DOCUMENT RESUME

ED 143 555 SE 023 039

AUTHOR	Blakers, A. L.
TITLE	Studies in Mathematics, Volume XVII. Mathematical Concepts of Elementary Measurement.
INSTITUTION	Stanford Univ., Calif. School Mathematics Study Group.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	67
NOTE	427p.; For related documents, see SE 023 028-041; Contains occasional light and broken type
EDRS PRICE	MF-$0.83 HC-$23.43 Plus Postage.
DESCRIPTORS	Arithmetic; Inservice Education; *Instructional Materials; Mathematical Applications; *Measurement; *Number Concepts; *Secondary School Mathematics; *Textbooks
IDENTIFIERS	*School Mathematics Study Group

ABSTRACT

The objective of this book is to identify those mathematical concepts which are relevant to elementary measurement, and to exhibit their logical interrelationships. The book was written with high school mathematics teachers in mind, but it is hoped that it will be useful also to elementary school teachers, science teachers, and college teachers. The book could be useful as a text for an advanced undergraduate course, or as an inservice course for teachers. The mathematical background required is approximately that which is included in a good high school education. Sections in the book include: (1) Measurement and Measure Functions; (2) The Measurement of Numerosity and Length; (3) The Measurement of Angles, Area, and Volume; and (4) Measurement and Dimension. A list of references conclude the publication. Each section includes background information, discussion of concepts, some suggestions for instruction, and exercises. (RH)

SCHOOL
MATHEMATICS
STUDY GROUP

STUDIES IN MATHEMATICS

VOLUME XVII

Mathematical Concepts of Elementary Measurement

By A. L. Blakers

SMSG

STUDIES IN MATHEMATICS

Volume XVII

MATHEMATICAL CONCEPTS OF ELEMENTARY MEASUREMENT

By A. L. Blakers

Permission to make verbatim use of material in this book must be secured from the Director of SMSG. Such permission will be granted except in unusual circumstances. Publications incorporating SMSG materials must include both an acknowledgment of the SMSG copyright (Yale University or Stanford University, as the case may be) and a disclaimer of SMSG endorsement. Exclusive license will not be granted save in exceptional circumstances, and then only by specific action of the Advisory Board of SMSG.

Financial support for the School Mathematics Study Group has been provided by the National Science Foundation.

TABLE OF CONTENTS

Preface 1

Chapter 1. MEASUREMENT AND MEASURE FUNCTIONS 5

1-1. Introduction 5

1-2. Measurement 6

1-3. Relations 9

1-4. Functions 12

1-5. The Algebra of Real Valued Functions 22

1-6. Some Special Sets of Functions 32

1-7. A Classification of Measure Functions 42

Chapter 2. THE MEASUREMENT OF NUMEROSITY AND LENGTH 51

2-1. Introduction 51

2-2. The Real Number System 52

2-3. The Measurement of Numerosity 77

2-4. The Physical Measurement of Length 91

2-5. Length in Formal Mathematical Systems 116

2-6. Coordinate Systems 155

2-7. Directed Segments, Orientation, and Vector Measures 162

2-8. Extension of the Domain for Length Functions: Curve Length. . 165

2-9. Extension of the Domain for Length Functions: Measure Theory. 179

Chapter 3. THE MEASUREMENT OF ANGLES, AREA, AND VOLUME 195

3-1. Introduction 195

3-2. Angle Measurement From An Empirical Standpoint 199

3-3. Angular Measure In Mathematics 207

3-4. Extension of the Domain for Angular Measure:
Directed Angles and Rotations 232

3-5. The Elementary Theory of Area 243

3-6. Extension of The Domain For Area Functions 295

3-7. The Measurement of Volume 308

Chapter 4. MEASUREMENT AND DIMENSION 315

4-1. Introduction 315

4-2. Ratio Scales 317

4-3. Primary and Secondary Quantities 322

4-4. A Mathematical Theory of Ratio Scales 354

4-5. Formulas, Dimensions, and Related Topics 399

References 419

PREFACE

This book began as an attempt to "explain what measurement is about, from the point of view of a mathematician". It was intended that there would be a companion book, written from a scientific viewpoint. It soon became apparent that a book about measurement, written from either point of view, would have to say something about the other. It also became apparent that the original objective was inappropriate: the question of "what measurement is about" is not mathematical, but philosophical. Much has been written (and much more will undoubtedly be written) by scientists and philosophers, on the subject of measurement. Most of these writers find it necessary to use mathematical ideas in their "explanations"; but, from a mathematical point of view, their treatment of the relevant mathematical ideas, and of the connections between these ideas, frequently leaves much to be desired. It is an objective of the present book to help to fill this gap: to identify those mathematical concepts which are relevant to elementary measurement, and to exhibit their logical interrelationships.

In comparatively recent times, it has been discovered that mathematics has no necessary logical connection with the real world. This discovery has been accompanied by an ever growing expansion of the activity, often called "model-building", which seeks to link the empirical structures of the sciences with the formal structures of mathematics. This link is established by means of functions, which map (or "model") empirical systems into mathematical systems, in such a way that structures arrived at empirically and inductively are carried over into corresponding mathematical structures. Sometimes this process makes use of existing mathematical systems, and sometimes new mathematical systems are created to provide appropriate model spaces. Among the mathematical systems in most frequent use as model spaces are the various number systems (the whole numbers, the rational numbers, the real numbers, the complex numbers), various geometric spaces, vector spaces, and so on. Many model functions are, in fact, complex and inter-related collections of simpler functions, and it is these simple functions (which are associated directly with the processes of measurement) and their relationships with one another, which are one of our main concerns in this book.

Measurement also has a place within mathematics: there are many situations in mathematics in which one mathematical system is mapped into another in such a way that we feel that a "measurement" process is involved (e.g., length and area concepts in geometry). When these mathematical systems are used in model-making, mathematical "measurement" frequently becomes a component in empirical measurement. We examine some of the simplest examples of this situation.

The book has been written with high school mathematics teachers in mind, but it is hoped that some of it will be within the grasp of elementary school teachers, and that it might be usefully read by teachers of science, and by college teachers. The principal concern of the book is to exploit the idea that "measurement involves structure-preserving functions" in order to provide a conceptual framework in which the elementary ideas of measurement can be understood. It is not necessary to follow all of the details (some of which are rather involved) in order to get a picture of this framework; and it is certainly not implied that this is the only framework which is suitable for the study of a theory of elementary measurement.

As you will see, the mathematical concepts which are relevant to elementary measurement come from a variety of branches of mathematics (classical real analysis, linear algebra, linear analysis, geometry, elementary topology, and so on). These concepts are usually encountered separately in more or less distinct mathematics courses, and the reader who has so encountered them should find it interesting to see how a study of the mathematical background of measurement ideas brings them together. For this reason the book should be a useful text for an advanced undergraduate course, or for an inservice course for high school mathematics teachers.

The mathematical background which is required of the reader is approximately that which is included in a good high school education, but it is expected that most readers will have gone further than this. We assume a general familiarity with the real number system, some knowledge of geometry, and some idea of what is involved in the concept of "function". In the functional approach to measurement, we are typically concerned with two systems, and we are looking for structure-preserving functions from one to the other. The structures commonly encountered may be described in terms of such notions as equivalence relations, order relations, and binary operations; and the systems themselves are often semi-groups, or groups. We introduce these terms, because it would be clumsy, and unnatural, to do without them; but you do not need any prior knowledge of them in order to understand their limited use in this book.

In most of the measurement situations which we treat, the values of the measure functions are positive real numbers. From the empirical point of view it is frequently sufficient to use positive rational numbers, but from the theoretical point of view we cannot answer many of the most interesting questions (e.g.: Why must two length functions be similar? Why must the area functions for rectangular regions be related as they are to length functions? Why are power functions and homogeneous functions so intimately connected with dimension questions?) unless we use the real numbers, and those properties of real numbers (i.e., topological completeness) which distinguish the real number system from the rational numbers. For many readers this distinction might not be clear, so we have devoted a section to an outline of the development of the structure of the real number system. In this section we prove a number of results which depend on the deeper properties of the real numbers, and which we need to use later in the book. The pace is fairly rapid, and the reader to whom the ideas of this section are totally unfamiliar will profit from a more detailed study of one of the expanded treatments to which we refer. However, in this and other places it is probably best to skip over some of the more complex details, and return to them if and when they become necessary for an understanding of what comes later. The pace of the book is necessarily uneven, and you should not attempt to master every topic before proceeding to the next.

A number of exercises are included, particularly in the earlier review sections. Some of these are really extensions of the text, so you should (at least) read them to see what they are about. In the later sections there are many unproved statements which can be treated as exercises.

Concerning measurement itself, nothing is assumed which is not part of the general knowledge of most citizens. However, before reading this book you might find it useful to review the elementary ideas on measurement which are contained in the School Mathematics Study Group publications (in the "Studies in Mathematics" Series):

Vol. V:	Concepts of Informal Geometry (Chapters 6, 7, 10)
Vol. VII:	Intuitive Geometry (Chapters 2, 7)
Vol. IX:	A Brief Course in Mathematics For Elementary School Teachers (Chapters 27, 28)

Measurement is a very big topic, and anything approaching a complete treatment would occupy a small library. With so much material available, it is inevitable that the choice of content has been somewhat arbitrary. Thus there is an emphasis on mathematical, rather than on empirical, ideas; there is an emphasis on those measurement concepts which have been motivated by the physical sciences, rather than on those which are more relevant to the biological and social sciences; and such important topics as the establishment and maintenance of standards, and the statistical analysis of data, are almost completely ignored.

To a considerable extent we have concentrated on giving a fairly complete treatment of elementary concepts, rather than a superficial picture of the whole subject. Thus much of our discussion falls in that no-man's-land of ideas which are usually considered too sophisticated for an elementary treatment, but which are later assumed to be "known" or "understood", in more advanced courses and texts.

In a few places (e.g., under the heading "Links With Other Parts of Mathematics") we have pointed out that some of the mathematical ideas which arise in connection with a theory of measurement, are directly related to some of the basic ideas of more advanced mathematics. (E.g., the notion of "dual space", in linear algebra; and the notion of the "tensor product" of modules.) These connections are pointed out for the benefit of any reader who happens to be familiar with these ideas, and to show that many so-called "advanced" notions are already present in the context of elementary measurement; but it is not assumed that the majority of mathematics teachers either are, or should be, familiar with these ideas at the present time. We expect to explore these connections more thoroughly in a later book.

The present book will have served its purpose if it gives you some feeling for the variety of mathematical ideas which are relevant to an elementary theory of measure, and if it encourages you not only to pursue these ideas, but also to read more widely (and critically) in the extensive literature which is devoted to the subject of measurement.

Chapter 1

MEASUREMENT AND MEASURE FUNCTIONS

1-1 Introduction

In the year 1900, the great philosopher-mathematician, Bertrand Russell, wrote:

> "Measurement of magnitudes is, in its most general sense, any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational or real, as the case may be. - - - - - It will be desirable that the order of the magnitudes measured should correspond to that of the numbers, i.e., that all relations of between should be the same for magnitudes and their measures."

While this statement might not give us a very complete picture of all the complexity that is involved in the notion of measurement, it does contain the germ of the idea that the present book attempts to convey: that a "measure" is a function, defined on some specified set of objects, and designed to reflect certain properties of those objects. In order to elaborate this idea, we need to use such mathematical notions as set, relation, function, group, semigroup, and so on. As these are not generally treated together in the way in which we need them, most of this first chapter is devoted to a basic review of ideas and terminology. If these ideas are familiar to you, you should be able to go through the chapter fairly quickly. If they are not, you are urged to stop and work the exercises. (For a few of the exercises you will have to draw on your knowledge of mathematics outside of this book.) References are given to more detailed treatments of many of the ideas introduced.

The chapter concludes with the outline of a scheme for classifying measure functions. This involves the informal use of some common ideas like length, time, area, etc., some of which are introduced more precisely in later chapters. This was done deliberately in order to give you a general framework in which to fit all, or at least most of the common measure functions, before getting involved in so much detail that it might obscure the overall picture.

1-2 Measurement

You might expect us to begin with a definition of "measurement", or by explaining the nature of measurement, so that you could determine, in a particular situation, whether the idea of measurement was involved. This, unfortunately, we are unable to do. Some attempts at definition tie the notion of measurement to the notion of number. Our point of view is certainly broader than this: we consider that each of the examples below involves the idea of measurement. These examples have been chosen to indicate the wide variety of contexts in which the idea of measurement is discernible.

1. Our identity is measured at birth, partially by the assignment to each of us a name, partially in terms of the identity of our parents and the date of our birth. Later our identity might be more accurately measured by a serial number if we are in the armed forces, or by a social security number, a fingerprint, or a passport.

2. The location of our place of residence is usually measured by a set of four items: a state name, a city name, a street name, and a number.

3. Our growth may be measured by a set of triples, each of which consists of a date, a number representing our weight and a number representing our height.

4. The size of our family is measured by a number, the cardinal number of the set of members of our family.

5. Our shoe size is measured by a pair of items consisting of a number and a letter (or combination of letters). These, in turn, are measures of the length and the width of our shoes.

6. The size of our house is measured by a collection of numbers representing such things as the floor area and the numbers of rooms of various types.

7. Our intelligence is measured by a number, our so-called IQ.

8. Our school report cards are a measure of our educational progress.

9. We can use their annual dividend rates as a measure of the success of our investments.

10. If we are farmers, our annual production of wheat is measured by a number which gives the size of our wheat crop in bushels; our annual production of eggs is measured by a number which gives the size of our egg "crop" in dozens.

11. A baseball player's batting average is a measure of his success at bat.

12. The half-life of a radioactive element is a measure of the stability of its atomic nucleus.

13. The symmetry of a crystal is measured by means of a certain group of transformations,

14. The extent to which a group fails to be abelian (commutative) is measured by its commutator subgroup.

15. The connectivity of a topological space is measured by its homotopy groups.

[Don't worry if the last few examples contain unfamiliar ideas.]

One could add to this list indefinitely, but this should be enough to convince you that the idea of measurement is found in a wide variety of contexts and forms. You might well ask whether there are any discernible common features, and if so, what are they.

Firstly, notice that in each case there are "objects" to be measured. In examples 1 , 3 , 7 , 8 , and 11 these objects are people. In 6 the objects are houses. In 4 the objects are families. In 13 the objects are crystals. In 15 the objects are topological spaces. In 14 the objects are groups.

Secondly, each measurement involves some attribute of the object, and some process by means of which this attribute is to be measured. We don't simply measure people: we measure them for weight, we measure them for height, we measure them for intelligence, and so on. It is convenient in many cases to use the common names of the measured attributes -- names such as length, speed, area, intelligence -- but we make no attempt to define these words, or to consider them independently of the processes by means of which they are measured. In most, if not all, cases, it is doubtful if the attribute has any objective meaning except in relation to the measurement process.*

Finally, in each case there is a "quantity" which results from applying a measurement process to one of the objects for which this particular process is applicable. In example 4 , this quantity is a whole number. In 9 , 10 , and 11 , the quantity is a real number. In 8 , it is a report card (i.e., the set of information contained on the card). In example 1 , it might be a fingerprint, a social security number, or a passport, depending on the particular process used.

*This question is of considerable philosophical interest. You will find reference to it in [1] , [2] , [3] , [4] , [5] and in various Encylopaedia Britannica articles under the headings "Dimensional Analysis", "Meaning", "Knowledge", "Logical Positivism".

To sum up: in each case there is a collection A of objects to be measured, a set B of "quantities" or "measurements", and a procedure for associating with each object in A an element of B. We immediately recognize this situation: what we are dealing with is a function. We shall have a lot to say about functions in later sections, but it might be useful to recall here the basic idea of a function.

Suppose that A and B are any two sets, and that we have a rule which assigns exactly one member of B to each member of A. Then the rule, together with the set A, is said to be a function (or a mapping), and the set A is called its domain. The elements of A are called arguments of the function. The set of those members of B which are actually assigned to members of A is called the range of the function. The particular member of B which is assigned to a particular argument a is called the value of the function at a, or the image of a under the function. If f is a name for the function, then the value of f at a is usually denoted by f(a). We say that "f is a function on A with values in B"; or "f is a function from A to B", and indicate this symbolically by such notations as:

$$f : A \rightarrow B$$

and

$$A \xrightarrow{f} B .$$

The set B is called the value space, or image space, of f. Strictly speaking, we should not use the definite article, as the same function can have different image spaces. (Some writers use "range", where we use "image space"; and "set of values", where we use "range". Other writers use "domain of definition" where we use "domain", and "domain of values", where we use "value space".)

In every situation in which the idea of measurement is involved, there is a related function: its domain is the set of objects to which the particular measurement process applies, and the process itself provides the rule by means of which a value, or measure, is assigned to each member of the set of objects.

It is convenient to refer to those functions which arise in measurement situations as measure functions. Some of the measure functions arising in our earlier examples can be roughly described by such expressions as "length in inches", "area in square feet", "numerosity in dozens".

It has sometimes been suggested that for any two sets A and B, any function $f : A \to B$ should be regarded as a measure function on A. We do not need to accept or reject this point of view. However, if adopted, it would mean that there are an awful lot of measure functions, for most of which we have, at present, no conceivable use. You will find a lot of interesting discussion and diversity of viewpoint on this, and on other ideas related to measurement, in [1].

1-3 Relations

In this section and the next, we review some of the basic ideas and properties of relations and functions. Much of this will probably be familiar to you, so we go fairly quickly with very few examples. Parts of this material are treated in much more detail, and with many examples, in the SMSG publications, "Intermediate Mathematics", "Elementary Functions", and "Calculus". A more advanced treatment can be found in [6].

The notion of function is one of the fundamental ideas of mathematics. It is also central to our treatment of measurement. There are different ways of approaching the function concept. One of these ways has been indicated in the last section, where a function from a set A to a set B was regarded as a rule of correspondence, or association, which pairs each element of A with exactly one element of B. An equivalent procedure is to regard a function as a special kind of relation. As we shall need the general concept of relation later, we devote this section to a review of some of the main ideas concerning relations, before continuing the discussion of functions.

Let A and B be two sets, not necessarily different. The cartesian product of A and B, denoted by $A \times B$, is the set of all ordered pairs (a,b), where $a \in A$, $b \in B$. (Equality of ordered pairs is defined by: $(a_1,b_1) = (a_2,b_2)$ if and only if $a_1 = a_2$ and $b_1 = b_2$.) A particular case, with which you are undoubtedly familiar, is the cartesian product $R \times R$ of the real number system with itself: the elements of $R \times R$ are ordered pairs of real numbers. Given any plane, we can set up a 1-1 correspondence* (i.e., a coordinate system) between points of the plane and elements of $R \times R$. Thus if we picture the real number system as the number line, we can similarly picture $R \times R$ as the (cartesian) plane.

*It is assumed that you are familiar with the notion of 1-1 correspondence. It is introduced formally in Section 1-4.

We are frequently interested in pairs of sets whose elements have some sort of relationship between them. For example, if A is the set of male residents of a certain town, and B the set of all residents, we can consider the relationship "is the son of" between members of A and B. That is, for any $a \in A$ and $b \in B$, either a is the son of b or a is not the son of b. We can abbreviate "a is the son of b" by $a \, S \, b$. Thus the relationship "is the son of" determines a set of ordered pairs (a,b) for which $a \, S \, b$. This set is a subset of $A \times B$. It is sometimes convenient to denote this subset by the same symbol S, so that

$$S = \{(a,b) : a \, S \, b\} .$$

We generalize this situation by defining a <u>binary relation from A to B</u>, to be any subset of the cartesian product $A \times B$; if $A = B$, we call this a <u>binary relation on A</u>.

Whenever a subset K of $A \times B$ is specified, we can use it to define a "relationship" between certain elements of A and B. For example, A might be the set of males at a dance, B the set of females, and K the subset of $A \times B$ consisting of those ordered pairs (a,b) such that a danced with b; or $A(=B)$ might be the set of all people living in the United States, and K the set of those ordered pairs (a,b) such that a and b live in the same state; or $A(=B)$ might be the set of real numbers, and K the set of those ordered pairs (a,b) such that $a < b$. (Can you picture this set of points as a subset of the cartesian plane?) The number of examples could be extended indefinitely; the concept of relation is clearly very general.

The set of those a in A which appear as a first member in at least one ordered pair $(a,b) \in K$, is called the <u>domain</u> of K; it is a subset of A. The set of those b in B which appear as a second member in at least one ordered pair $(a,b) \in K$, is called the <u>range</u> of K.

Let us temporarily confine our attention to situations where $A = B$. That is, we consider relations on a set A. Such relations can be classified in terms of their properties. Let K be a relation on A; (i.e., $K \subset A \times A$). Then we say that

(i) K is <u>reflexive</u> if $(a,a) \in K$ for every $a \in A$; (i.e., if $a \, K \, a$ for every $a \in A$);

(ii) K is <u>symmetric</u> if $(a,b) \in K$ whenever $(b,a) \in K$; (i.e., $b \, K \, a$ implies $a \, K \, b$);

(iii) K is <u>transitive</u> if $(a,c) \in K$ whenever $(a,b) \in K$ and $(b,c) \in K$; (i.e., $a \, K \, b$ and $b \, K \, c$ imply $a \, K \, c$).

To fix these ideas in mind you should examine the relevant examples above to see which of the relations in them have some or all of these properties, and you should construct other examples for yourself. You should also look for examples of relations which have none of these properties; (e.g., the relation F defined by: $a \; F \; b$ provided that $a = b + 1$, on the set of positive integers, $\{1,2,3,\ldots\}$).

Relations which have all three properties -- i.e., which are symmetric, reflexive, and transitive -- are particularly important; they are called <u>equivalence relations</u>. These have the important property of separating (or <u>partitioning</u>) the sets to which they apply into disjoint subsets. (A <u>partition</u> of a set is a collection of non-empty pairwise disjoint subsets, whose union is the whole set.) For an example of an equivalence relation, see the relationship above of residing in the same state. The disjoint subsets resulting from an equivalence relation are called <u>equivalence classes</u>: any two elements in the same equivalence class stand in the given relation to each other, and no two elements from different equivalence classes stand in the given relation to each other. Every equivalence relation on a set determines a partition of the set, and every partition determines an equivalence relation. Equivalence relations abound in mathematics: congruence and similarity of geometric figures are equivalence relations; congruence modulo a non-zero integer is an equivalence relation on the integers; the relation $(a,b) \; K \; (c,d)$ if and only if $a + d = b + c$, is an equivalence relation on the set of ordered pairs of natural (positive whole) numbers; the relation "is as tall as" is an equivalence relation on a set of people; the relation of 1-1 correspondence is an equivalence relation on a collection of sets. You should try to think of other examples.

Another kind of relation which is particularly important in the consideration of measurement is an <u>order relation</u>. There are a number of different types of order relation, all of which are transitive. A partial order relation, or <u>partial ordering</u> is transitive, reflexive and <u>antisymmetric</u>. (That is, $K \subset A \times A$ is a partial ordering on A if, in addition to being transitive and reflexive, $(a,b) \in K$ and $(b,a) \in K$ imply $a = b$.) Examples of partial orderings are:

(i) $a \; K \; b$ if and only if $a \leq b$; $A = R$ = the set of real numbers;

(ii) $a \; K \; b$ if and only if $a \subseteq b$; A is the set of all subsets of a fixed set S.

Note that for a partial ordering K on a set A , it is not necessary that a K b or b K a for each two elements a , b of A . (Look at example (ii) above from this point of view.) However, if a partial order relation K on a set A satisfies the additional condition that for every two elements a , b of A , either a K b or b K a , then the relation is called a weak total order relation. For an example see (i) above.

Another important class of order relations are those which, in addition to being transitive, are irreflexive. A relation K on a set A is irreflexive if, for every $a \in A$, $(a,a) \notin K$. (Note that reflexive and irreflexive are not complementary properties: there are some relations which are neither reflexive nor irreflexive. See if you can think of one.)

A strict total order relation K on a set A is a relation which is transitive, and which has the additional property (often called the "law of trichotomy") that for every two elements a , b of A , exactly one of the following three statements is true: a = b ; a K b ; b K a . The relations < and > on the real numbers are well known examples of strict total order relations. When there is no danger of confusion we abbreviate "strict total order relation" to "order relation". It is easy to prove that a strict total order relation is irreflexive.

1-4 Functions

In Section 1-2 we have described a function f from A to B as an association of exactly one element of B with each element of A . Thus for each $a \in A$, f determines an ordered pair (a,b) with b = f(a) , and hence f determines uniquely a set F of such ordered pairs, with the properties:

(i) each element of A occurs as a first member of some ordered pair (a,b) from F ;

(ii) each element of A occurs only once as a first member.

As we have seen in the previous section, the set F is a relation from A to B . Thus a function determines a particular kind of relation: one whose domain is the whole of A. (property (i)), and which also satisfies a condition of "single-valuedness" (property (ii)). On the other hand, it is clear that if we have a relation F satisfying (i) and (ii) , then we can use F to define a function $f : A \to B$, such that f(a) = b if and only if $(a,b) \in F$. Thus we have a natural 1-1 correspondence between functions and those relations which satisfy (i) and (ii) . This suggests that we could equally well define a function as a special kind of relation -- a definition which you will find in many books.

The function concept is so important that it is useful to use both approaches, and to realize their equivalence. Because of this equivalence we shall move freely from one approach to the other, and it will be convenient to economize on notation by using the same symbol for the set of ordered pairs and for the rule of association. Thus, referring back to our earlier notation, we use $f = F = \{(a, f(a)) : \text{for all } a \in A\}$.

The notion of a function as a set of ordered pairs has an obvious connection with the notion of the graph of a function. The graph of a function $f : A \to B$ can be defined as the subset of $A \times B$ which the function determines. This makes the ideas of function, and graph of a function, virtually the same. You will recall that in more elementary work, especially where $A = B = R$ = the set of real numbers, it is customary to refer to a diagrammatic representation of f as its "graph".

A function $f : A \to B$ is said to be one-one (1-1) if $f(a_1) = f(a_2)$ implies that $a_1 = a_2$. It is said to be onto, if every element of B appears at least once as a value; i.e., if the range of the function is B. Every function is onto its own range. A function which is both 1-1 and onto is the well-known 1-1 correspondence or isomorphism of sets. Such functions are particularly important, because they have inverses: the inverse of a 1-1 correspondence $f : A \to B$ is the function (also a 1-1 correspondence) f' defined by $f'(b) = a$, where a is the unique element of A satisfying $f(a) = b$. Clearly, if f' is the inverse of f, then f is the inverse of f'.

A 1-1 correspondence $f : A \to B$ is often denoted by $f : A \longleftrightarrow B$. In order to distinguish (in diagrams) between a 1-1 correspondence $f : A \longleftrightarrow B$ and its inverse, we introduce the notation

$$A \overset{f}{\leftarrow\!\!\twoheadrightarrow} B .$$

This should be read "f is a function from A to B, and f is a 1-1 correspondence". The inverse function (f') can be indicated by

$$A \overset{f'}{\twoheadleftarrow\!\!\rightarrow} B .$$

In other words, the double arrowhead indicates the "direction" which corresponds to the named function. We shall look at inverses again after we have considered the notion of composition of functions.

A particularly important type of function is one which maps each ordered pair of elements of a set into an element of the set. Such a function is known as a binary operation on the set. Thus a binary operation on a set A is a function

$$f : A \times A \rightarrow A .$$

If f is only defined on a subset of $A \times A$, then we call it a binary operation in A. The best known examples of binary operations on a set are the familiar operations of addition, subtraction, and multiplication, defined on the real numbers. Examples of binary operations in sets are: subtraction for the positive integers, division for the positive integers, division for the real numbers.

If a binary operation

$$f : A \times A \rightarrow A$$

satisfies

$$f(f(a,b), c) = f(a, f(b,c))$$

for all $a, b, c, \in A$ for which each side of the equation is defined, then f is said to be an associative operation. This property is more familiar in the form which uses a notation such as

$$f(a,b) = a \circ b .$$

With this notation the associative condition becomes

$$(a \circ b) \circ c = a \circ (b \circ c) .$$

Using the same notation, the operation is said to be commutative if $a \circ b = b \circ a$, for all a, b, for which each side of the equation is defined. You should remind yourself, by consideration of the familiar operations of subtraction and division in the real numbers, that not all binary operations are associate and/or commutative.

More generally, a function

$$f : A \times A \rightarrow B$$

is called a binary operation on A with values in B. For example, if A denotes the set of all cities in the United States, the function which assigns to each ordered pair of cities (a_1, a_2) the minimum highway distance between them (in miles) is a binary operation on A with values in the set of real numbers. The concept of commutativity is defined as before. If $B \neq A$, then the question of associativity does not arise.

If A, B, C are three sets, and if $f : A \to B$, $g : B \to C$ are functions, then we can compose f and g by considering their effect, in that order, on elements of A. This leads to the following definition of the composite function gf:

$$gf = \{(a,c) : a \in A, g(f(a)) = c\} .$$

In other words, $c = (gf)(a)$ is the value of g on $f(a)$. No ambiguity can arise if we omit the parentheses on gf. Note that we have denoted by gf the composite function "f first, then g" because of the way in which we write the value of a function on a particular element of its domain; some books use fg to denote "f first, then g". This is sometimes referred to as the "product" of f and g, but we avoid this term for reasons which will become clear later. (In considering functions whose values are real numbers we shall wish to consider the product of two functions, obtained by multiplying values. The notation for this, $f \cdot g = g \cdot f$, should be carefully distinguished from the notation gf, which we have introduced for the composite of g and f.) Observe that if $A = B = C$, then composition is a binary operation on the set of all functions from A to A.

The following diagram is useful in picturing composition of functions:

$$\begin{tikzcd}
A \arrow[rr, "gf"] \arrow[dr, "f"'] & & C \\
& B \arrow[ur, "g"'] &
\end{tikzcd}$$

The composite gf is so defined that, starting from an element a of A, each of the possible "function paths" from A to C leads to the same element $(gf)(a) = g(f(a))$ of C. I.e., we have

$$\begin{tikzcd}
a \arrow[rr, "gf"] \arrow[dr, "f"'] & & (gf)(a) = g(f(a)) \\
& f(a) \arrow[ur, "g"'] &
\end{tikzcd}$$

The idea of composition can be extended to an ordered set of three or more suitable functions; i.e., functions having the property that the range of any one is contained in the domain of the next, if there is a next. Thus if $f : A \to B$, $g : B \to C$, $h : C \to D$, we can form by composition $h(gf)$

and $(hg)f$. We then have, for any $a \in A$,

$$(h(gf))(a) = h(g(f(a))) = ((hg)f)(a)$$

so that composition is associative, and we can drop parentheses and write simply hgf for the composite function. The following diagram may be used in picturing this result:

$$\begin{tikzcd}
A \arrow[rr, "gf"] \arrow[dr, "f"'] \arrow[drrr, bend left=60, "hgf"] & & C \arrow[dr, "h"] & \\
& B \arrow[ur, "g"] \arrow[rr, "hg"'] & & D
\end{tikzcd}$$

We can interpret the result as indicating that, if we start from any element a of A, and proceed from it to D by any of the four possible "function-paths", the same element of D is reached in each case. Diagrams which picture sets and functions which are related in this way, are called commutative diagrams.

As we shall use commutative diagrams a great deal, we slow down here and say a little more about them. First of all, a diagram such as the one used above (whether or not it is commutative) should be considered as a natural extension of the commonly used simple diagram for a function:

$$A \xrightarrow{f} B$$

The term "commutative diagram" probably derives from one of the simplest examples of the use of such a diagram: the commutative diagram which corresponds to a commutative binary operation. Let $f : A \times A \to A$ be a binary operation on A, and denote $f(a_1,a_2)$ by $a_1 \circ a_2$. Then there is a natural function (ρ, say) on the elements of $A \times A$, which simply reverses the order of the terms in each ordered pair: that is, $\rho : (a_1,a_2) \to (a_2,a_1)$. Clearly ρ is a 1-1 correspondence. If we consider now the following function diagram

$$\begin{tikzcd}
A \times A \arrow[rr, leftrightarrow, "\rho"] \arrow[dr, "f"'] & & A \times A \arrow[dl, "f"] \\
& A &
\end{tikzcd}$$

then the diagram is commutative if and only if, for every $(a_1, a_2) \in A \times A$ we have

$$\begin{tikzcd}
(a_1, a_2) \arrow[rr, leftrightarrow, "\rho"] \arrow[dr, "f"'] & & (a_2, a_1) \arrow[dl, "f"] \\
& a_1 \circ a_2 = a_2 \circ a_1 . &
\end{tikzcd}$$

In other words, the diagram is commutative if and only if the binary operation is commutative.

In order to be commutative, a diagram (of functions) must have the property that every pair of composite functions represented (by directed "paths") in the diagram, which have the same domain and the same image space, must "agree", i.e., they must be the same function. If this condition fails for any element in the common domain of any pair of suitable composite functions, then the diagram is not commutative.

The earlier diagram which represented the associativity of functional composition, was commutative because of the way in which composition was defined. This is a fairly common situation. But we shall also encounter other instances of commutative diagrams in which the commutativity is a theorem, and not quite so obvious.

Commutative diagrams have been considerably used in more advanced parts of mathematics (especially in algebra and algebraic topology). They are particularly useful whenever we have a number of suitably interrelated sets and functions. Like any good diagram, their main purpose is as an aid to the imagination: they frequently help us to picture and summarize functional relationships which can be quite complicated when written out in "algebraic" form.

It is possible to prove theorems about commutative diagrams, but, as our use of them is quite elementary, we shall not take the time to do this. An example of such a theorem (which you can easily prove) is the following:

Theorem. If each of the triangular "subdiagrams" in the following diagram is commutative, then the whole diagram is commutative:

$$\begin{tikzcd}
A \arrow[r] \arrow[d] \arrow[dr] & B \arrow[d] \\
C \arrow[r] & D
\end{tikzcd}$$

The ideas of composition and inverse can be brought together through the notion of identity function. For any set A, the identity function I_A is the 1-1 correspondence $I_A : A \rightarrow A$ for which $I_A(a) = a$ for every $a \in A$. If we have a 1-1 correspondence $f : A \rightarrow B$ whose inverse is denoted by f', then we can compose f and f' in two ways, and we get

$$f'f = I_A ; \quad ff' = I_B .$$

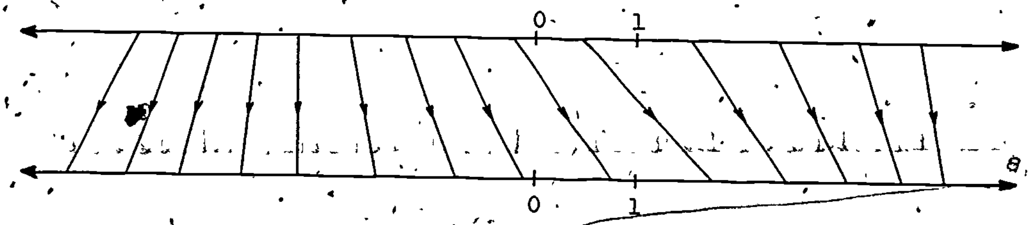
You should check this, and observe that there is some similarity between the composition of functions and the multiplication of real numbers, with the identity function playing the role of the number 1. Because of this similarity the notation f^{-1} is often used where we have used f', to denote the inverse of a 1-1 correspondence f. We do not use the notation f^{-1} here, because we need it later to denote something different, but is not always feasible to avoid it. In most cases the sense will be clear from the context.

Frequently we have to deal with a function $f : A \rightarrow B$., in a situation where there are relations K_A, K_B on A and B respectively. In this case we say that f is compatible with (or preserves) the relations K_A, K_B, if $(a_1, a_2) \in K_A \Longrightarrow (f(a_1), f(a_2)) \in K_B$. We also say that such an f is a homomorphism from (A, K_A) to (B, K_B). If $A = B$, and $K_A = K_B$, f is called an endomorphism of (A, K_A).

If, in addition to being compatible with the given relations, f is a 1-1 correspondence whose inverse f' is also compatible with K_A, K_B, then we say that f is an isomorphism of A onto B with respect to the given relations. We denote this by $(A, K_A) \approx (B, K_B)$. If f maps A isomorphically onto a proper subset of B, then we say that f is an isomorphism of A into B. The notions of homomorphism and isomorphism will keep recurring in different contexts, but they will generally have the same sort of meaning: we have two sets with some sort of "structure" (usually given by relations and operations -- see later) and a function which "preserves the structure". It is a simple matter to show that the composite of two homomorphisms (isomorphisms) is a homomorphism (isomorphism).

Important cases of relation-preserving functions, which we shall encounter in this book, occur when $A = B = R$ = the set of real numbers, and when $K_A = K_B$ is one of the well-known order relations $>$, $<$, $\geq$, $\leq$. A function $f : R \to R$ which is compatible with $\geq$ or with $\leq$ is said to be weakly increasing, or order-preserving; a function which preserves $>$ or $<$ is said to be increasing. A function which reverses order (e.g., for which $a < b \Longrightarrow f(a) \geq$ (or $>$) $f(b)$) is said to be weakly decreasing (or decreasing). A function which is either weakly increasing or weakly decreasing is called monotone. A function which is either increasing or decreasing is called strongly monotone. A strongly monotone function is 1-1. A function which is strongly monotone and onto (and hence a 1-1 correspondence) is called isotone (or isotonic). Thus an isotonic transformation is an isomorphism with respect to the order structure of R. More precisely, an isotone increasing function is an isomorphism of $(R,<)$ and $(R,<)$; an isotone decreasing function is an isomorphism of $(R,<)$ and $(R,>)$.

When dealing with a function f from R to R, it is sometimes useful to "picture" the function by means of a diagram in which the domain and the range are separately represented by parallel copies of the number line, and arguments, a, are joined by directed line segments to their values, $f(a)$. (Of course, not all can be drawn!) For example, such a diagram for a monotone increasing function might look like:



The monotone increasing property is reflected in the fact that no two segments (drawn or not) cross each other. For monotone decreasing functions, every two segments cross. You might find these ideas useful in thinking about some of the exercises below.

Exercises 1-4

1. Prove the assertion made above, that a strongly monotone function from R to R is 1-1.

2. If you are familiar with the notion of continuity, show that a 1-1 correspondence R ↔ R is isotonic if and only if it is continuous.

3. Prove that the inverse of an isotonic function is also isotonic, and has the same sense (i.e., increasing or decreasing).

4. If f and g are monotone (or strongly monotone) functions in the same sense (i.e., both increasing, or both decreasing) show that fg and gf are monotone increasing (or strongly monotone increasing); if f and g have opposite senses, then both composites are decreasing.

5. Prove that if f and g are isotone, then fg and gf are isotone.

6. Prove that the identity function I_R is isotonic and increasing.

We conclude this section with some further consideration of the various notations used in the description of functions:

If we are dealing with a finite set (not too large!) and a function for which there is no particular pattern in the assignment of a value to each argument, we usually list the set of ordered pairs which describe the function. For example, the state-of-residence function for a specified set of people (the domain) could be described by a set of ordered pairs:

{(Smith, New York) , (Jones, California) , - - - - - }

Another function normally described in this manner is that which associates a telephone number with each person in a specified domain: you may regard a telephone directory as giving an organized listing of the ordered pairs corresponding to this function. As far as this function is concerned, the alphabetical order of the listing is irrelevant.

When we are dealing with a function whose domain and/or image space has some "structure", it is often (but not always) possible to describe the function by means of an equation, or in some other way. For example, if domain = image space = R = set of real numbers, the function f which maps every number into its square may be described by such notations as $y = f(x) = x^2$; $f : x \to x^2$ Such notations are incomplete -- the domain must be specified separately -- but they have advantages in other respects. For example, we cannot specify a function whose domain is an infinite set, by listing separately all of the ordered pairs which go to make up the function.

When using the notation $y = x^2$, with domain = image space = R, we understand that the function described is the set of those ordered pairs of real numbers (x,y) which make up the truth set of the equation $y = x^2$. This idea is also used in situations involving a verbal statement. For example, we may define the <u>integer part function</u> f, whose domain is R, by

$$f = \{(x,y) : x, y \in R ; y \text{ is the largest integer for which } x - y \geq 0\} .$$

This integer part of x is often denoted by $[x]$, so that the integer part function on R is also described by

$$f : x \to [x] .$$

<u>Exercises 1-4</u> (continued)

7. Show that the function $f : R \to R$ defined by

$$f : \begin{cases} x \to x^2 & \text{for } x \geq 0 \\ x \to -x^2 & \text{for } x \leq 0 \end{cases}$$

is isotone.

8. If $f : x \to f(x)$ is monotone on R (strongly monotone; isotone) and a, b, c are real numbers, $a \neq 0$, show that

(a) $g : x \to af(x)$
(b) $h : x \to f(x + b)$
(c) $k : x \to f(x) + c$

and hence

(d) $j : x \to af(x + b) + c$

are monotone (strongly monotone; isotone). Sketch suitable graphs to help you to picture these results.

9. If a, b, c, d are real numbers, and f, g, are the functions on R given by

$$f : x \to ax + b , \quad g : x \to cx + d,$$

find expressions giving the values of fg and gf, at x.

1-5 The Algebra of Real Valued Functions

We introduce first the notion of equality of functions. Two functions f and g are said to be equal provided that they have the same domain, and provided that for every element a of the domain, $f(a) = g(a)$. Clearly, equal functions determine the same set of ordered pairs; and equality of functions is an equivalence relation.

Let A be any set, and let R be the set of real numbers. We use R^A to denote the set of all functions from A to R. The motivation for this notation lies in the exercises below.

Exercises 1-5

1. Prove the assertion made above, that equality of functions is an equivalence relation.

2. Let A and B be finite sets containing a and b elements respectively. Prove that B^A (the set of functions from A to B) contains b^a elements.

3. Let A be a finite set with a elements, and let S denote the set of all subsets of A. Let B be the 2-element set $\{0,1\}$. If $T \in S$ (i.e., T is any subset of A) define $f_T : A \to B$, (i.e., $f_T \in B^A$)

by
$$f_T(x) = \begin{cases} 0 & \text{if } x \notin T \\ 1 & \text{if } x \in T . \end{cases}$$

Now define:

$$F : S \to B^A$$

by $F(T) = f_T$, and show that

(a) F is 1-1 and onto;
(b) S has 2^a elements.

(For this reason the symbol 2^A is sometimes used to denote the set of all subsets of a given set A, whether or not A is finite.)

The elements of the set R^A are functions. Particular elements of R^A which can be singled out are the so-called constant functions: corresponding to each real number r, we define the constant function $\underline{r} : A \to R$, by $\underline{r}(a) = r$ for every $a \in A$. Important constant functions are $\underline{0}$ and $\underline{1}$.

Operations of addition and multiplication can be defined on the set of functions R^A, by using the corresponding operations in R. It is customary to use the usual symbols, "+" for addition of functions, and "$\cdot$" for multiplication of functions. Thus if

$$f_1, f_2 \in R^A ,$$

we define

$$f_1 + f_2 = \{(a, f_1(a) + f_2(a)) : a \in A\} ;$$

$$f_1 \cdot f_2 = \{(a, (f_1(a)) \cdot (f_2(a))) : a \in A\} .$$

Clearly $f_1 + f_2$ and $f_1 \cdot f_2$ belong to R^A, which is therefore closed under addition and multiplication. You should verify for yourself that this addition and multiplication are both associative and commutative; that the multiplication is distributive over the addition; and that $\underline{1} \cdot f = f \cdot \underline{1} = f$, and $\underline{0} \cdot f = f \cdot \underline{0} = \underline{0}$ for all $f \in R^A$.

For each function $f \in R^A$ we can define a unique negative, or additive inverse, $-f \in R^A$ by

$$-f = \{(a, -f(a)) : a \in A\}$$

and we readily verify that

$$f + (-f) = \underline{0} .$$

Subtraction of functions can now be introduced in the usual way.

You might be tempted to think that R^A has all of the algebraic structure of the real numbers, but this is not generally the case:

Firstly, while all non-zero real numbers have multiplicative inverses, we can define

$$f^{-1} = \{(a, \frac{1}{f(a)}) : a \in A\}$$

only for those functions f whose range does not include the number zero. (In general, the set of those functions whose range includes zero contains much more than the constant function $\underline{0}$.) We call f^{-1} the <u>multiplicative inverse</u>, or <u>reciprocal</u> of f. If f^{-1} exists, we have $f \cdot f^{-1} = \underline{1}$. We can define division by $\frac{f}{g} = f \cdot g^{-1}$ only when $0 \notin$ range of g. One consequence of this restriction on the existence of multiplicative inverses is that the set R^A can have what are called <u>divisors of zero</u>. These are elements f , g , such that $f \neq \underline{0}$, $g \neq \underline{0}$ but $f \cdot g = \underline{0}$. You should convince yourself that such functions exist, by constructing examples.

Secondly, we can set up an order relation in R^A by the definition: $f \leq g$ provided that $f(a) \leq g(a)$ for all $a \in A$; but in general, this relation is only a partial ordering, whereas the corresponding relation for the real numbers is a total ordering.

Exercises 1-5 (continued)

4. Let A be a 2-element set. Find functions f , $g \in R^A$ such that $f \nleq g$ and $g \nleq f$.

5. If A has exactly one element, show that the sets R^A and R are isomorphic with respect to their algebraic operations and order structures.

Another operation which can be introduced on the set of functions from A to R , is the so-called "scalar multiplication" of elements of R^A by real numbers. Let $r \in R$ and $f \in R^A$. Then we define $rf : A \to R$ by

$$(rf)(a) = r \cdot (f(a)) \quad \text{for all } a \in A .$$

This "multiplication" is related to the notions of constant function and the multiplication of functions: see Exercise 6 below.

Exercises 1-5 (continued)

6. With the notation above, show that

$$rf = \underline{r} \cdot f = f \cdot \underline{r}$$

where $\underline{r} : a \to r$, for every $a \in A$.

7. If n is a positive integer, and $f : A \to R$, show that $nf = f + f + \ldots + f$ (n terms).

8. (a) Prove that addition in R^A is commutative and associative.

(b) If p , q are real numbers, and f , $g \in R^A$, prove that

(i) $(pq)f = p(qf) = q(pf)$;

(ii) $(p + q)f = pf + qf$;

(iii) $p(f + g) = pf + pg$;

(iv) $1f = f$.

9. If f, g are monotone (strongly monotone; isotone) functions in the same sense, from R to R, show that

   (a) $f + g$ is monotone (strongly monotone; isotone) in the same sense as f and g;

   (b) if $r > 0$, rf is monotone (strongly monotone; isotone) in the same sense as f;

   (c) if $r < 0$, rf is monotone (strongly monotone; isotone) in the opposite sense to f.

A particularly important function space is the set R^R, the well known set of "real functions of a real variable". In this case, in addition to the algebraic structure for the set R^R as described above, we can compose functions: the composite of two functions in R^R is also a function in R^R. Thus we have the additional binary operation of composition in R^R. This operation is associative, but not commutative. Composition is related in a number of interesting ways to the other operations in R^R, but it would take us too far afield to investigate these relationships fully.

Similarity Transformations and Similar Functions. A subset of R^R which is important in questions of measurement, is the set of those functions which arise from the multiplication of every number in R by a fixed number. If $k \in R$, we denote the corresponding function by the symbol $\overline{k}$. Thus $\overline{k} : R \to R$ is defined by

$$\overline{k} : r \to kr, \text{ for all } r \in R .$$

It is frequently convenient to denote the function $\overline{k}$ simply by the number k, but we must be careful not to confuse it with the constant function $\underline{k}$. These functions are related to the identify function I_R in a simple way: $\overline{k} = kI_R = \underline{k} \cdot I_R$. If this is combined with the results of the exercises above, we obtain the following properties for $\overline{k}$:

(i) if $k \neq 0$, $\overline{k}$ is a 1-1 correspondence from R to itself;

(ii) if $k > 0$, $\overline{k}$ is isotonic (i.e., 1-1 and strictly monotonic) and order preserving;

(iii) if $k < 0$, $\overline{k}$ is isotonic but reverses order;

(iv) $\overline{1} = I_R$; $\overline{pq} = \overline{p}\,\overline{q} = \overline{q}\,\overline{p}$;

(v) if $f : A \rightarrow R$ then

$$kf = \overline{kf} = (\underline{k} \cdot I_R)f = \underline{k} \cdot f = f \cdot \underline{k} .$$

This result can be pictured using the commutative diagram:

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[dr, "kf"'] & R \arrow[d, "\overline{k} = \underline{k} \cdot I_R"] \\ & R \end{tikzcd}$$

and noting that if we start from any element a of A, move by f to its value $f(a)$ in R, then down by $\overline{k}$ to $\overline{k}(f(a))$, we reach the same element of R as if we had gone directly to R by kf. If $A = R$, observe that, in general, $\overline{k}f \neq f\overline{k}$. (Consider $f : x \rightarrow x^2$.)

If $k \neq 0$, $\overline{k}$ is called a <u>similarity transformation</u>, or <u>similitude</u>, of the real numbers. A similitude transforms any subset of the real line into a similar subset, in the geometrical sense. If $0 < k < 1$, $\overline{k}$ could be described as a "uniform contraction"; if $k > 1$, $\overline{k}$ could be described as a "uniform expansion".

If f and g are two functions from A to R, and if there exists $k \in R$ $(k \neq 0)$ such that $g = \overline{k}f$, then f and g are said to be <u>similar functions</u>. If $k > 0$, f and g are said to be <u>positively similar functions</u>. In connection with measurement, sets of positively similar functions (with values in the set R^+ of positive real numbers) are extremely common. (E.g., the set of all length functions, with a common domain.) When dealing with such functions we usually omit the word "positively", and refer to them simply as "similar functions". Similarity, and positive similarity, are equivalence relations.

A similarity transformation on R is a special case of a linear function. A <u>linear function</u> on R is a function f

$$f : x \rightarrow ax + b$$

where a, b are fixed real numbers. If $a \neq 0$, the linear function is <u>non-singular</u>; such a function is also known as a <u>polynomial function of the first degree</u>. If $a \neq 0$ and $b = 0$, the linear function is a similarity transformation. If $a = 0$, the linear function is <u>singular</u>: a singular linear function on R is, of course, a constant function.

Exercises 1-5 (continued)

10. Assume that $f : x \to ax + b$, $g : x \to cx + d$ are linear functions on R, and let $k \in R$ $(k \neq 0)$. Prove that

(a) kf is a linear function; kf is non-singular if and only if f is non-singular;

(b) if f and g are non-singular, then $f \cdot g$ is not a linear function;

(c) $f + g$ is a linear function; $f + g$ is non-singular if and only if $a \neq -c$;

(d) fg and gf are linear functions; fg and gf are non-singular if and only if f and g are non-singular;

(e) $fg = gf$ if and only if $ad + b = bc + d$;

(f) $fg = gf = I_R$ if and only if $ac = 1$ and $bc + d = ad + b = 0$;

(g) I_R is linear;

(h) every non-singular linear function f has an inverse f' (with respect to composition) such that f' is linear and non-singular, and $ff' = f'f = I_R$;

(i) if $a > 0$, f is isotonic increasing; if $a < 0$, f is isotonic decreasing.

11. If $A \neq \emptyset$, prove that

(a) similarity is an equivalence relation on R^A ;

(b) positive similarity is an equivalence relation on $(R^+)^A$ and on R^A;

(c) all constant functions in R^A are similar.

As you are undoubtedly aware, the graphical representation of a linear function in the cartesian plane is a straight line. You might find it helpful to use a graphical picture when working some of the above exercises.

If you have taken a course in calculus, you will have encountered many of the ideas mentioned above, but not in a context which emphasizes the algebraic structure of R^R . You will recall that in differential calculus we are interested in those functions from R^R (or from R^A , where A is a specified subset of R) which have derivatives. This is the subset of so-called differentiable functions. In the development of the properties of derivatives, you undoubtedly discussed rules for differentiating sums of functions, products of functions, quotients of functions, multiples of functions by real numbers, and functions of functions (i.e., composite functions).

These various operations were precisely those which we have discussed above, but restricted to the subset of differentiable functions. Thus differentia~~tion itself~~ is a function whose domain is the set of differentiable functions in R^R, and whose values (not necessarily differentiable) lie in R^R ; and the various "rules of differentiation" express the relationship of the derivative function to the algebraic structure of R^R .

Another situation which arises in connection with measure functions, (especially in relation to "derived" measures, and "dimension") concerns the subject of induced functions on cartesian products: Suppose that we are given two functions

$$f : A_1 \to B_1 \; ; \; g : A_2 \to B_2$$

Then f and g induce, in a natural way, the function $f \times g : A_1 \times A_2 \to B_1 \times B_2$, defined by

$$f \times g : (a_1,a_2) \to (f(a_1) , g(a_2)) \; ; (a_1 \in A_1 , a_2 \in A_2) \; .$$

If $B_1 = B_2 = B$, and there is a binary operation on B (such as addition, multiplication, or division) then $f \times g$ can be composed with the binary operation (which, you will recall, can be considered as a function from $B \times B \to B$) to give a mapping from $A_1 \times A_2$ into B . We shall see an example of this in connection with the relationship of length and area functions, where $B = R^+$, and the relevant binary operation on R^+ is multiplication. Other examples concern angular measures and velocity measures, where the relevant operation is division. These and related questions of "dimension" will be considered in a later chapter.

Linear and Homogeneous Functions. Let $a , x , b \in R$. A polynomial $ax + b \; (a \neq 0)$ of degree 1 , determines a non-singular linear function $x \to ax + b$. This polynomial is homogeneous if $b = 0$, and the corresponding function $f : x \to ax$, satisfies the condition $f(kx) = kf(x)$ for every positive k . Such a function is said to be a homogeneous function of degree 1 in a single variable, or argument.

These ideas can be generalized to define concepts of linearity and homogeneity for functions of several variables (i.e., on a finite cartesian product ΠR , or on some specified subspace of this product.) Another generalization leads to the more restricted notions of multilinear and multihomogeneous functions, which we shall encounter in connection with the treatment of derived measures and dimensions. We introduce these ideas here in their simplest form (i.e., with domains and image spaces derived from the real numbers) and consider further generalizations as the need arises.

A function $f : R \times R \to R$ is <u>linear</u> if and only if it has the following properties:

(i) for every (x_1,y_1) and $(x_2,y_2) \in R \times R$,
$f(x_1 + x_2, y_1 + y_2) = f(x_1,y_1) + f(x_2,y_2)$;

(ii) for every $(x,y) \in R \times R$, and every $k \in R$, $f(kx, ky) = kf(x,y)$.

<u>Remark</u>: Perhaps you are surprised that the second property should be required, as it is a "homogeneity" condition; that is, our definition is really a generalization to two variables of the notion of a homogeneous linear function of one variable. It so happens that the notion of homogeneous linear function is the important one in generalizations, and the homogeneity property is therefore included in the definition. Thus our (generalized) linear functions are all homogeneous, and our multilinear functions (see below) will be multihomogeneous. Many writers use the term "linear transformation" (especially in generalizations to so-called "linear spaces") where we have used "linear function"; a linear transformation is thus a homogeneous linear function.

It follows readily, from the definition, that $(x,y) \to 2x + 3y$ is a linear function on $R \times R$, and that the function $(x,y) \to 2x + 3y + 1$ is not. The definition is easily extended to the case of a finite number of real variables.

A function $f : R \to R$ is defined to be <u>homogeneous</u> if and only if it has the property that there exists a fixed $\alpha \in R$, such that for every $k > 0$, and every $x \in R$,

$$f(kx) = k^{\alpha}f(x) .$$

The number α is called the <u>degree of f</u>. An example of such a homogeneous function is the function $x \to 2x^3$. The definition can be suitably modified to apply to a function whose domain is a subset of R. (E.g., the function $x \to 3x^{1/2}$.)

The concept of homogeneity can be simply extended to functions of several real variables; we give the definition for the case of two variables only. A function $f : R \times R \to R$ is <u>homogeneous (of degree α)</u> if and only if there exists a fixed $\alpha \in R$, such that for every $k > 0$, and every $(x,y) \in R \times R$,

$$f(kx\,,ky) = k^{\alpha}f(x,y)$$

This is the natural generalization of the idea of a homogeneous polynomial function; e.g., the function $(x,y) \to x^2 + 3xy + 3y^2$ is easily shown to be homogeneous of degree 2. A nonpolynomial example is the function $(x,y) \to \frac{x}{\sqrt{x^2 + y^2}}$, which is homogeneous of degree 0. You should observe that, with the definition which we have given for a linear function of two or more variables, every such linear function is homogeneous of degree 1, but the converse is definitely false. (See exercises below.)

Multilinear and Multihomogeneous Functions. The concepts of multilinear and multihomogeneous function are more restrictive generalizations of the notions of linearity and homogeneity. We give the definitions for the case of two variables only: these definitions are easily extended.

A function $f : R \times R \to R$ is bilinear if and only if it has the following properties:

(i) for every real x_1, x_2, y_1, and y_2,

$$f(x_1 + x_2\,,y_1) = f(x_1,y_1) + f(x_2,y_1) \text{ , and}$$

$$f(x_1\,,y_1 + y_2) = f(x_1,y_1) + f(x_1,y_2) \text{ ;}$$

(ii) for every real x, y, and k, $f(kx\,,y) = f(x\,,ky) = kf(x,y)$.

Observe that the second property is again a homogeneity requirement: it implies that every bilinear function is homogeneous of degree 2. Every bilinear function is also bihomogeneous (see below) of degree $(1,1)$. An example of a bilinear function is the function $(x,y) \to 3xy$.

A function $f : R \times R \to R$ is bihomogeneous if and only if there exist real α_1, α_2, such that, for every $(x,y) \in R \times R$, and every positive real k_1, k_2,

$$f(k_1x\,,k_2y) = k_1^{\alpha_1}k_2^{\alpha_2}f(x,y) \text{ .}$$

The ordered pair of real numbers (α_1,α_2) is called the degree of the bihomogeneous function. (We also say that such a function has degree α_1 in x and α_2 in y.) An example is the function $(x,y) \to 3x^2y^3$, which is bihomogeneous of degree $(2,3)$. The function $(x,y) \to 2x^{-3}y^{1/2}$, on the domain $R^+ \times R^+$, is bihomogeneous of degree $(-3\,,\frac{1}{2})$.

Exercises 1-5 (continued)

12. If $f : R^+ \to R^+$ is a homogeneous function of degree α, prove that there is a $c \in R^+$ such that for all $x \in R^+$, $f : x \to cx^\alpha$; hence show that f' is homogeneous of degree zero if and only if it is a constant function.

13. If $f : R \times R \to R$ is a polynomial function, prove that f is homogeneous of degree n (n a positive integer) if and only if all of its terms have the same degree, n. (This latter property is, of course, the usual definition of "homogeneous polynomial of degree n".)

14. Prove that for each positive integer n, the function $(x,y) \to (x^n + y^n)^{1/n}$, defined on $R^+ \times R^+$, is homogeneous of degree 1, but it is not linear unless $n = 1$.

15. Prove that every bilinear function is homogeneous of degree 2, and bihomogeneous of degree $(1,1)$.

16. $f : R^+ \times R^+ \to R^+$ is bihomogeneous of degree (α_1,α_2). Prove that

    (a) there is a $c \in R^+$ such that $f : (x,y) \to cx^{\alpha_1}y^{\alpha_2}$;

    (b) f is homogeneous of degree $\alpha_1 + \alpha_2$;

    (c) f is linear if and only if $\alpha_1 = 0$, $\alpha_2 = 1$; or $\alpha_1 = 1$, $\alpha_2 = 0$.

17. Find examples of homogeneous functions (of 2 variables) which are not bihomogeneous.

18. If you are familiar with the notion of "Venn diagram", draw a Venn diagram which illustrates the relationship of the sets of linear, bilinear, homogeneous, and bihomogeneous functions from $R^+ \times R^+$ to R^+.

19. f and g are homogeneous functions with the same domain, and with degrees α_1, α_2, respectively.

    (a) Prove that $f + g$ is homogeneous if and only if $\alpha_1 = \alpha_2$, and that, in this case, the degree of $f_1 + f_2$ is α_1.

    (b) Prove that $f_1 \cdot f_2$ is homogeneous of degree $\alpha_1 + \alpha_2$.

20. If $f : R^+ \to R^+$ and $g : R^+ \to R^+$ are homogeneous functions of degrees α_1 and α_2, respectively, prove that the composite functions fg and gf are each homogeneous of degree $\alpha_1\alpha_2$.

1-6 Some Special Sets of Functions

In this section we introduce some special sets of functions, most of which have the structure of a group (or a semigroup) with respect to the operation of composition. These groups are used in the next section in connection with the classification of measure functions. [For a more detailed introduction to group theory, including most of the groups discussed here, see [7].]

Before discussing these particular groups, we introduce the concept of group formally, and summarize some of the main group ideas which we shall need. A group consists of a non-empty set G, together with a binary operation on G (the value of this operation on (x,y) is indicated in this definition by the juxtaposition xy) such that

(i) G is closed under the operation. (Actually this is implicit in the requirement of a binary operation on G);

(ii) the operation is associative; i.e., if $x, y, z \in G$, then $(xy)z = x(yz)$;

(iii) G contains a special element e, called an identity element (or null element), such that

$$ex = xe = x \quad \text{for all } x \in G ;$$

(iv) corresponding to each $x \in G$, there is a unique element $x^{-1} \in G$, such that $xx^{-1} = x^{-1}x = e$.

The element x^{-1} is called the inverse of x with respect to the given operation.

A group G is called abelian, or commutative if for each pair of elements x, y, of G,

$$xy = yx .$$

A non-empty subset of a group, which is itself a group with respect to the given group operation, is called a subgroup of the original group.

Exercises 1-6

1. Verify that the following sets and operations are groups:

   (a) the integers (or the rational numbers, or the real numbers), under addition;

   (b) the non-zero rational numbers (or the non-zero real numbers), under multiplication;

   (c) the positive rational numbers (or the positive real numbers), under multiplication;

   (d) the equivalence classes of integers, mod 12, under addition;

   (e) the equivalence classes of integers mod 7, with the zero class excluded, under multiplication;

   (f) the set of all 1-1 functions of a 3-element set onto itself, under function composition;

   (g) the set of linear functions on R, under composition;

   (h) the set R^A of all functions from a set A to the real numbers R, under addition of functions;

   (i) the set R_+^A of all functions from a set A to the positive real numbers R^+, under multiplication of functions.

2. Which of the groups in the previous exercise are abelian?

3. If G is a group, $H \subset G$, show that H is a subgroup of G provided that

   (a) $h_1h_2 \in H$, for all h_1, h_2 in H (i.e., the subset H is closed under the group operation);

   (b) $e \in H$;

   (c) $h \in H \Longrightarrow h^{-1} \in H$.

4. Show that the set of even integers is a subgroup of the group of integers under addition.

5. Show that ~~the~~ set of non-zero rational numbers is a subgroup of the group of non-zero real numbers under multiplication.

6. Show that the set of positive reals is a subgroup of the group of non-zero reals under multiplication.

7. Show that the 2-element set {1,-1} is a subgroup of the group of non-zero reals under multiplication.

8. Show that the relation ρ defined on the set $\{H_i\}$ of all subgroups of a group G by:

$$H_1 \ \rho \ H_2 \text{ if and only if } H_1 \text{ is a subgroup of } H_2 ,$$

is a reflexive, antisymmetric, and transitive relation; (i.e., a partial order relation).

Permutation Groups. A *permutation* of a finite set of objects is a 1-1 mapping of the set onto itself. In other words, a permutation is just a 1-1 correspondence. Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of n objects. The permutations of A can be composed by functional composition, and the composite of any two permutations is again a permutation; the composition operation is associative; the permutation I_A is an identity element for the set of permutations; and each permutation is 1-1 onto, and hence has an inverse with respect to composition. It follows that the set of all permutations of A is a group under composition. This very important group is known as the *permutation group* (or *symmetric group*) on n objects, and denoted by $\underline{P}_n$. The definite article is used because, for fixed n, the nature of the n objects does not affect the structure of the permutation group: all permutation groups on n objects are similarly structured or *isomorphic*. This is another example of the idea of isomorphism: Two groups are *isomorphic* if there is a 1-1 correspondence between their elements which preserves the group structure. (Actually it is sufficient to require the existence of a 1-1 correspondence which is compatible with the group operations: see Exercise 9 below). A function which is compatible with the group operations, and which is onto, but not necessarily 1-1, is called a *homomorphism*. Thus an isomorphism is a 1-1 homomorphism. (See exercises below.) An isomorphism of a group onto itself is called an *automorphism*.

A function which maps a group G isomorphically onto a proper subgroup of a group H, is referred to as an *isomorphism* of ~~G into~~ H. A *homomorphism into* is similarly defined. A homomorphism of a group to itself (onto or into) is called an *endomorphism*.

Homomorphisms are conveniently pictured by means of commutative diagrams. If G and H are groups, a function $f : G \to H$ induces (in a natural way) a function $f \times f : G \times G \to H \times H$ as described in Section 1-5. If the vertical arrows in the following diagram indicate the group operations for G and H, then f is a homomorphism of G into H if and only if the following diagram is commutative:

$$\begin{tikzcd}
G \times G \arrow[r, "f \times f"] \arrow[d] & H \times H \arrow[d] \\
G \arrow[r, "f"'] & H
\end{tikzcd}$$

The isomorphism of the permutation groups on equivalent sets, is a direct consequence of the 1-1 correspondence of the sets of permuted objects. You might be tempted to think that isomorphism of groups is always this simple, but this is not the case. For example, if G denotes the group of positive real numbers under multiplication, and H the group of all real numbers under addition, the function

$$f : x \to \log_{10} x$$

is an isomorphism from G to H.

Exercises 1-6 (continued)

9. G and H are groups and $f : G \to H$ is a 1-1 correspondence and a homomorphism (i.e., $f(g_1 g_2) = f(g_1)f(g_2)$ for all $g_1 , g_2 \in G$). Show that f satisfies ;

   (a) $f(e_G) = e_H$; (e_G , e_H are the respective identity elements),

   (b) $f(g^{-1}) = (f(g))^{-1}$ for every $g \in G$.

   Thus f preserves products, the identity element, and inverses (i.e., the whole group structure) and hence f is an isomorphism.

10. Prove the assertion made above, that $f : x \to \log_{10} x$ is an isomorphism from the positive reals under multiplication to the reals under addition.

11. What is the inverse function to the f of Exercise 10? Is this also an isomorphism?

12. Show that the composite of homomorphisms (of groups) is also a homomorphism, and that the composite of isomorphisms is an isomorphism.

13. If $k \neq 0$, show that the function

$$f : x \to kx$$

   is an automorphism on the group of real numbers under addition.

14. Show that $\underline{P}_n$ contains $n!$ elements.

15. Show that each of the groups $\underline{P}_1$, $\underline{P}_2$, is abelian, but that $\underline{P}_3$ is not abelian.

The set of all 1-1 correspondences of a non-finite set (e.g., the real numbers) with itself, is also a group under functional composition, as you may readily check. We use the symbol $\underline{P}_R$ for the group of 1-1 correspondences of R. Although the word "permutation" is usually used only in the finite case, it is sometimes convenient to refer to $\underline{P}_R$ as the permutation group of the real numbers. The function I_R is, of course, the identity element of $\underline{P}_R$.

The Isotonic Group. You will recall that in Section 1-4 we defined an isotonic transformation of the real numbers as one which was strongly monotone and onto (and hence 1-1 and continuous). In the exercises of Section 1-4 we asked you to prove that

(i) the composite of two isotonic functions is isotonic;

(ii) the identity function I_R is isotonic;

(iii) the inverse of an isotonic function is isotonic.

If you did not prove these before, you should do so now. These properties, and the fact that composition of functions is always associative, show that the set of isotonic functions is a group under composition. Observe that this group, which we call the isotonic group, is a subgroup of the group $\underline{P}_R$. We denote the isotonic group by the symbol $\underline{I}$.

In another of the exercises of Section 1-4 you were asked to show that

(i) the composite of two strongly monotone increasing functions is strongly monotone increasing;

(ii) the identity function is isotonic increasing;

(iii) the inverse of an isotonic increasing function is isotonic increasing.

These results show that the set of isotonic increasing functions forms a subgroup of the isotonic group. By analogy with the multiplicative properties of the real numbers, we call this subgroup the positive isotonic group, and denote it by the suggestive notation $\underline{I}^+$.

Exercises 1-6 (continued)

16. By means of examples, show that neither $\underline{I}$ nor $\underline{I}^+$ is commutative.

The Affine Group. A linear function $f : x \to ax + b$ on R is also called an affine transformation. A non-singular affine transformation is a function

$$f : x \to ax + b , \qquad \text{with } a \neq 0 .$$

In Exercise 1-5.10 we asked you to prove that if f and g are non-singular affine transformations, then

(i) fg and gf are non-singular affine transformations;

(ii) I_R is a non-singular affine transformation;

(iii) every such transformation has an inverse (with respect to composition) which is also a non-singular affine transformation.

It follows that the set of all such functions on R is a group under composition. This group is called the affine group on R, or the 1-dimensional affine group. We denote this group by the symbol $\underline{A}$.

A non-singular affine transformation $x \to ax + b$ can be regarded as the composite of a homogeneous non-singular linear transformation, $x \to ax$, and a translation, $x \to x + b$, in the given order. It is also the composite of the translation $x \to x + \frac{b}{a}$ and the non-singular linear transformation $x \to ax$, in the given order.

Exercises 1-6 (continued)

17. Show by means of examples that $\underline{A}$ is not commutative.

18. Show that the homogeneous non-singular linear transformations $x \to ax$, $a \neq 0$ form a subgroup of $\underline{A}$.

19. Show that the translations $x \to x + b$ form a subgroup of $\underline{A}$.

We consider next those affine transformations $f : x \to ax + b$ for which $a > 0$. For convenience we call these functions positive affine transformations. It is a simple matter to verify that

(i) the composite of positive affine transformations is positive;

(ii) I_R is a positive affine transformation;

(iii) the inverse of a positive affine transformation is affine and positive.

Thus the positive affine transformations form a subgroup of the affine group. We denote this by $\underline{A}^+$, and refer to it as the positive affine group (of dimension one).

We examine next the relationship of the affine groups to the isotonic groups discussed earlier. As we saw in Section 1-5, a non-singular linear function is isotonic, and a positive non-singular linear function is isotonic increasing. Hence the affine groups $\underline{A}$, $\underline{A}^+$, are subgroups of the isotonic groups $\underline{I}$, $\underline{I}^+$; respectively.

The Similarity Group. In the discussion of the affine group you were asked to show that the set of transformations

$$f : x \to ax , \qquad a \neq 0$$

formed a group under composition, and that this group is a subgroup of the affine group. We call this group the similarity group on R, because of the connection of the transformations in this group with the notion of similarity in geometry. (A similarity transformation on R is the same as a non-singular homogeneous linear transformation on R, but this is not true for the corresponding transformations of the plane and higher dimensional spaces.) As we saw in Section 1-5, individual functions in the similarity group $\underline{S}$ are called similarity transformations (abbreviated to "similarities"), or similitudes. We call those similarities for which $a > 0$, positive similarities. You can easily verify that the positive similarities form a subgroup of $\underline{S}$. We denote this by $\underline{S}^+$.

The relationship between the various groups of functions introduced in this section is exhibited in the following diagram, in which each arrow indicates that the group at the tail of the arrow is a subgroup of the group at the head. It follows from the transivity of the subgroup relation that each group is a subgroup of any group reached from it along a sequence of arrows. The arrows may also be thought of as representing the natural inclusion functions, which map each element of a subset of a set, into itself.

$$\begin{tikzcd}
\underline{S}^+ \arrow[r] \arrow[d] & \underline{A}^+ \arrow[r] \arrow[d] & \underline{I}^+ \arrow[d] & \\
\underline{S} \arrow[r] & \underline{A} \arrow[r] & \underline{I} \arrow[r] & \underline{P}_R
\end{tikzcd}$$

Exercises 1-6 (continued)

20. Show that $\underline{S}$ and $\underline{S}^+$ are commutative groups.

21. Show that $\underline{S}$ is isomorphic to the multiplicative group of non-zero real numbers, and that $\underline{S}^+$ is isomorphic to the multiplicative group of positive real numbers.

Semigroups. We shall also need the more general ideas of semigroup and ordered semigroup, so we explain these briefly: A semigroup is a set of elements, together with an associative binary operation on the set. (This implies that the set is closed under the operation, but that is all; it is not necessary that there be an identity element, or inverses.) It follows that every group is a semigroup, but you can easily find examples of semigroups which are not groups. (See exercises.) If the operation is also commutative, the semigroup is said to be abelian. If a right-cancellation property holds (i.e., $ab = cb$ implies that $a = c$) the set is called a right-cancellation semigroup; if both right and left cancellation hold, the set is called a cancellation semigroup.

A semigroup H which has an order relation ρ, linked with the semigroup operation by the property: $a \rho b$ implies that $ac \rho bc$ and $ca \rho cb$ for all $c \in H$, is called an ordered semigroup. A group which is an ordered semigroup is called an ordered group.

The concepts of homomorphism and isomorphism are defined for the various types of semigroup in the natural way, and it is easily shown that the composite of two homomorphisms (isomorphisms) is a homomorphism (isomorphism).

Exercises 1-6 (continued)

22. (a) Show that the set of positive integers (positive rationals; positive reals) under addition, with the usual ordering "<", is an ordered abelian semigroup with cancellation.

    (b) Similarly for the set of all integers (rationals, reals) greater than some fixed positive integer (rational number, real number).

23. Similar to Exercise 22(a), but with respect to the operation of multiplication. In this case each of the semigroups has the additional property of possessing an identity element; which of these semigroups are groups?

24. Show that every group is a cancellation semigroup.

25. Show that the set of all real numbers is a semigroup with identity, with respect to multiplication. Is it an ordered semigroup? Is it a cancellation semigroup? Is it a group?

26. (a) Show that the set R^R of all functions from R to R is a semigroup with respect to composition; that it has a left and a right identity; that it is not commutative; and that it is neither a right nor a left cancellation semigroup.

    (b) Show that if $f, g, h \in R^R$, and if

    (i) $gf = gh$ and g is 1-1, then $f = h$;

    (ii) $fg = hg$ and g is onto, then $f = h$.

27. Show that if m, n are any positive integers, then the set of all transformations $\bar{n}$ on the positive integers, defined by

$$\bar{n} : m \to nm$$

is an abelian cancellation semigroup with identity, under composition; and that it is isomorphic to the semigroup of the positive integers under multiplication.

28. Similarly to Exercise 27, but for the set of all transformations $\overline{\frac{1}{n}}$, defined by

$$\overline{\frac{1}{n}} : x \to \frac{x}{n} ,$$

on the set R of all real numbers.

29. Prove that an ordered abelian semigroup is a cancellation semigroup.

30. $(H,+)$ is an ordered abelian semigroup. Prove that the relation: $(a,b) \sim (c,d)$ if, and only if, $a + d = b + c$, $(a,b,c,d \in H)$ is an equivalence relation on $H \times H$.

31. $(H,+)$ is an ordered abelian semigroup, n is a positive integer, and na denotes the n-fold iterated sum. Prove that the relation $\sim$ defined by:

$(a,b) \sim (c,d)$ if, and only if, for all positive integers m and n, $ma < nb$ if, and only if, $mc < nd$,

is an equivalence relation on $H \times H$.

[The results of the next two exercises are important in the discussion of measure functions. We give proofs in Chapter 2, after reviewing the properties of the real numbers.]

32. If $f : R \to R$ belongs to the affine group, prove that f preserves ratios (in particular, equality) of differences. I.e., prove that if $x_1, x_2, x_3, x_4 \in R$, and $x_3 \neq x_4$, then $f(x_3) \neq f(x_4)$, and

$$\frac{x_1 - x_2}{x_3 - x_4} = \frac{f(x_1) - f(x_2)}{f(x_3) - f(x_4)} .$$

Conversely, if f is an isotone transformation which preserves ratios of differences, prove that f belongs to the affine group.

33. Prove that any positive similarity transformation on R^+, is an automorphism of the ordered semigroup $(R^+,+,<)$. Conversely, if f is any automorphism of $(R^+,+,<)$ prove that f is a positive similarity transformation.

34. (a) Prove that every similarity transformation on R preserves ratios; and, conversely, that every ratio-preserving transformation is a similarity.

(b) Similarly show that the ratio-preserving transformations of R^+ are the positive similarities.

35. Let J^+ denote the set of positive integers, and Q^+ the set of positive rational numbers. Let $H = \{kx : k, x \in J^+, k \text{ fixed}\}$. Prove that $(H,+,<)$ is an ordered semigroup, and that if $k > 1$ this is a proper sub-semigroup of $(J^+,+,<)$. What is the corresponding situation if we replace J^+ by Q^+ or R^+?

36. Prove that the set of monotone increasing functions on R is a non-abelian semigroup (under composition) with a two-sided identity, and with left-cancellation.

1-7 A Classification of Measure Functions

In a subject as large as mathematics there is a continuous effort to find ways of giving conceptual order to the growing diversity of ideas and theories. One method which has proved fruitful and which was formally presented (in relation to the classification of geometries) by Felix Klein in a famous address (the Erlanger program) given in 1872, is to study the relationship between certain sets with mathematical structures (e.g., operations, relations) and certain sets of transformations which leave invariant the essential features of these structures. Thus in euclidean geometry, we might be concerned with the study of those properties of subsets of euclidean spaces which are unaffected by rigid motions (congruences), or by similarity transformations; in projective geometry the concern might be with the invariants of projective transformations; in affine geometry, with affine transformations; in topology, with topological transformations (homeomorphisms); in group theory, with isomorphisms; and so on. This is the spirit in which the present section is written. But before going into detail, it must be emphasized that we are not going to describe a nice tidy finished theory with a complete classification of all possible measure functions. There are many loose ends, and it is not clear that these could all be tidied up. Nevertheless we believe that you will find this partial classification of considerable value and interest.

We are all familiar with the fact that, in our everyday experience, we encounter a variety of length functions. In the next chapter we discuss the construction and properties of these functions in some detail, but for the purpose of this section we assume that you are familiar with the general properties and relationships of these, and other common measure functions. For length functions we assume that there is a domain D of objects (a term you should interpret very broadly) which possess the attribute of "length", and that there is a length-in-feet function, λ_f, from D to the real numbers R. We also have a length-in-inches function, λ_i; a length-in-miles function, λ_m; a length-in-centimeters function, λ_c; and so on. All of these functions have the same domain D, and the same value-space R, and they all purport to measure the same attribute, length. It is reasonable to ask whether there are any relationships between them. Of course the answer is "yes" : for example, the value of λ_i on any element $d \in D$ is 12 times the value of λ_f on d ; the value of λ_f on each element of D is 5280 times the corresponding value of λ_m ; and so on. In terms of the notation developed earlier, $\lambda_i = \overline{12}\,\lambda_f$; $\lambda_f = \overline{5280}\,\lambda_m$; etc. We can indicate these relations in the following commutative diagram: (Where an automorphism and its inverse are indicated by a double-headed arrow, the appropriate name is placed near the head of the arrow).

$$\begin{tikzcd}
 & & R \arrow[dd, leftrightarrow, "\overline{63360}"' near start, "\overline{1/63360}" near end] \\
D \arrow[urr, "\lambda_i"] \arrow[r, "\lambda_f"] \arrow[drr, "\lambda_m"'] & R \arrow[ur, leftrightarrow, "\overline{12}"' near end, "\overline{1/12}" near start] \arrow[dr, leftrightarrow, "\overline{5280}" near start, "\overline{1/5280}"' near end] & \\
 & & R
\end{tikzcd}$$

Any one of these length functions is similar to any other: i.e., any one can be obtained from any other by composition with an appropriate positive similarity of R ; and, more generally, any function obtained by composition of a length function with a positive similarity is also a perfectly suitable length function. We can summarize this situation by asserting that the essential properties of a length function are unchanged by composition with any element of the positive similarity group $\underline{S}^+$. We anticipate the next chapter by saying that the basic properties of a length function on a set D of real objects, are that it should assume only positive values, and that it should preserve an empirically determined "length structure" of D. This empirically determined structure will include an equivalence relation on D ; (a length-function must assign equal values to length-equivalent objects); an order relation on the set $\tilde{D}$ of equivalence classes; and an equivalence relation on $\tilde{D} \times \tilde{D}$, which determines equivalence of "ratios". As we will see, the group of transformations of R which (by composition) takes one length-function into another, is precisely the positive similarity group $\underline{S}^+$ introduced in the last section. Each element of $\underline{S}^+$

determines a transformation (actually a 1-1 correspondence) of the set Λ of all length functions onto itself. This is sometimes expressed by saying that $\underline{S}^+$ acts as a *group of operators* on Λ. Actually Λ has the structure of an ordered semigroup under addition, and these operators determine automorphisms of Λ as an ordered semigroup. In this sense, the ordered semigroup Λ of length functions is "invariant" under the positive similarity group of transformations.

As the lengths of all real objects are positive, we could consider $\underline{S}^+$ as a group of transformations on R^+: $\underline{S}^+$ is the largest group of such transformations which (by composition) leaves Λ invariant. Choosing a unit or a "scale" for length corresponds to selecting a particular function of Λ. As we shall see later, if d is any object in the domain D, and p any positive real number, then (with suitable assumptions) there is exactly one length function $\lambda \in \Lambda$ such that $\lambda(d) = p$. Moreover, for fixed d, the set of all such functions is the set of all length functions. Thus the set Λ has as many elements as there are positive real numbers.

The situation which we have described for length functions is common to many of the so-called "scalar measures" of the physical sciences (e.g., mass, area, volume, work, density, time intervals). The fact that a "scale" for measuring each of these is only unique up to a similarity transformation, is well known; it plays an important role in the method of "dimensional analysis".

Measure functions of the type which we have considered above, might well be referred to as *similarity-invariant measures*. Another name sometimes used for them is *ratio scale*, a name which is related to the fact that the preservation of ratios is the distinguishing feature of a similarity transformation. (See Exercise 1-6.34.)

Some of the measure functions used in the physical sciences, and many of those used in the social sciences, are determined only up to a transformation by composition with a larger group than the group of positive similarities. For example, in the measurement of temperature (not absolute temperature) you are undoubtedly familiar with the transformations between the centigrade function T_c and the fahrenheit function T_f: if d belongs to the domain of these functions, these transformations are

$$T_f(d) \to T_c(d) = \frac{5}{9}\,(T_f(d) - 32)$$

and

$$T_c(d) \to T_f(d) = \frac{9}{5}\,T_c(d) + 32\ .$$

That is, T_f and T_c differ by composition with the positive affine functions

$$x \rightarrow \frac{5}{9}x - \frac{160}{9} = \frac{5}{9}(x - 32)$$

and its inverse

$$x \rightarrow \frac{9}{5}x + 32 .$$

The important feature of these transformations (as far as temperature is concerned) is that they preserve equal differences. A little thought should convince you that any positive affine transformation on T_c (or T_f) will yield a suitable temperature function, and that the essential features of temperature functions are unaffected by composition with a positive affine transformation. In fact, there is no reason why we could not reverse the roles of hotter and colder, and construct a temperature scale (function) on which the values for hotter objects were smaller real numbers; this corresponds to permitting the variation of temperature functions by composition with any element of the full affine group.

A quite similar situation holds with respect to the measurement of position on a line. The measurement of position on a line, by an appropriate assignment of real numbers to points of the line, is the process of giving the line a coordinate system. A "coordinate function" on the line is a measure of location. It is well known that if any coordinate function is composed with any non-singular affine transformation of R, then another coordinate function is obtained. (You will find more detail on this question in the SMSG books "Geometry", "Geometry with Coordinates", "Analytic Geometry", and "Geometry Based on Ruler and Protractor Axioms".) Other measure functions whose properties are affine-invariant are the measure of location in time (e.g., calendar time), and potential energy.

When one looks beyond the physical sciences, one finds examples of measurement situations in which the image space is R, but the domain of the appropriate measure function has less structure with respect to the attribute being measured than in the cases of such attributes as length and temperature. This is reflected in the fact that a larger group of transformations of R leaves intact the essential features of the relevant measure functions. Most measurement procedures for ranking sets of objects in a transitive order, permit composition with elements of the isotonic (or positive isotonic) group, or even with elements of the corresponding semigroups of strongly monotone functions. An example is the ranking of a class of students by means of the scores on a test: These scores might range from, say, 0 to

200, but we would not generally infer that a student with a score of 80 was twice as good as one with a score of 40; or that the difference in ability between students with scores of 180 and 190 was the same as the ability difference of students with scores of 30 and 40. The scores merely yield an order relation on the domain, and composition with any strongly monotonic transformation of R does not disturb this feature. Of course, in practice, we often use (generally implicitly) a monotonic increasing transformation which transforms the raw score function into a function whose range is a segment (1,2,3,...,n) of the positive integers, and we regard this function as a sort of "canonical function" for the measurement of this particular attribute. Other examples of measure functions of this type are: hardness measurement for minerals; grading measures for the quality of materials; location of houses on a street, by numbers (where east and west, or north and south, are introduced, these can be regarded as positive and negative, in either order); many kinds of preference measurement in psychology; and so on. An early stage in the development of such measures as loudness and temperature (where we might only have the means for deciding for each pair of objects an order; such as "warmer than", which yields an empirically transitive relation) would put them into this category. The simplest form of the notion of utility (in economic theory) might be considered to belong to the <u>isotonic-invariant</u> category of measure functions; a more advanced viewpoint of this notion, which would put the measurement of individual utilities into the category of affine-invariant measure functions, is contained in Chapter 1 of the modern classic, <u>Theory of Games and Economic Behavior</u>, by J. Von Neumann and O. Morgenstern [8]. In this chapter one finds a thorough discussion of the sort of empirical "structure" on the domain of the utility-measure function, which would enable it to be considered as affine-invariant. It is interesting to note that, in the same chapter, there is brief mention of the basic idea of this section: that real-valued measure functions might be classified in terms of the sets (often groups) of transformations on R which lead to equivalent functions. This idea, which seems to have occurred independently to the psychologist S. S. Stevens, is also discussed (in more detail) in [1] and [9]. Stevens uses the term <u>interval scale</u> to describe a type of measure function which is affine-invariant, because affine transformations (and hence, of course, similarities) on R preserve equality of intervals (see Exercise 1-6.32), but strongly monotonic functions generally do not; he uses the term <u>ordinal scale</u> to describe a type of measure function which is isotone-invariant, and <u>ratio scale</u> to describe a type of measure function which is similarity-invariant. He also introduces the term <u>nominal scale</u> to describe those real-valued measure functions whose essential character is unchanged by the permutation group $\underline{P}_R$. In this case the only structure on the domain is an equivalence relation (which might be trivial), and the only requirement is that the measure function assign the same value to equivalent elements of the domain. In other words, the measure function simply uses numbers to name, or identify, equivalence classes. Examples of this type of measure are the identification of team members by numbers, the assignment of telephone numbers to individuals (not generally 1-1: as a rule members of the same family have the same number) and the assignment of social security numbers. (In the last example, if the assignment of a social security number were required to indicate order of entry to the scheme as well as to provide identification, then the measure would be ordinal rather than nominal.) In some countries, as many bewildered tourists have discovered, there are towns where house numbers in a street are assigned serially in the order of construction! Such an assignment is nominal, as far as the measurement of location is concerned, but ordinal when regarded as a measure of the time of construction, or of age. As an age measure, this would not be an interval scale, because it would not be generally true that pairs of houses with the same difference in their assigned numbers would have the same difference in their ages.

You will have noticed that the classification is rather "forced" or oversimplified, in several places. For example, although there would be nothing wrong, in principle, in using arbitrary real numbers, such as π , $\sqrt{2}$, or even negative numbers, to indicate social security numbers, in practice we prefer to stick to positive integers. Thus the "invariance set" might, in practice, be restricted to the set of permutations of the positive integers. Similar remarks apply to the numerical measurement of house position on a street, where we usually use integers. But these minor exceptions do not detract from the value of this transformation-set/invariance idea in giving a general classification of real-valued measure functions. Moreover, the idea can be extended beyond those measure functions whose values are real numbers.

The following table summarizes some of the above ideas:

CLASSIFICATION OF MEASURE FUNCTIONS

Type of Function	Empirical Structure in Domain	Invariance Group	Examples
Permutation-invariant or Nominal	Equivalence relation	$\underline{P}_R$	Social security numbers; identification numbers assigned to members of a team.
Isotone-invariant or Ordinal	Above, and also an order relation	$\underline{I}$ or $\underline{I}^+$	Street numbering; hardness of minerals; ranking of students.
Affine-invariant or Interval	All of above, and also an "equal-interval" relation on ordered pairs	$\underline{A}$ or $\underline{A}^+$	Location of position in space or time; temperature (not absolute); utility.
Similarity-invariant or Ratio	All of above, and also an "equal-ratio" relation on ordered pairs	$\underline{S}$ or $\underline{S}^+$	Length, absolute temperature, mass, density, work, area, volume, elapsed time, numerosity.

We make several comments on this table:

1. Roughly speaking, the domain of each type of measure function has a structure which includes that of the types listed above it in the table; i.e., we have an increasing complexity of domain structure (with respect to the particular attribute under consideration) as we read down the table.

2. The invariance groups become "smaller" as we read down the table. (Roughly speaking, each is a subgroup of the one above.) This is a natural consequence of the fact that there is more structure to be preserved.

3. We often wish to put additional restrictions on various measure functions. These include such restrictions as: positive values only; integer values only; rational values only; values on a certain segment of the integers only; and so on. These restrictions can be reflected in corresponding restrictions on the admissible transformations, and the resulting measure functions can be further classified according to the appropriate sub-groups or semigroups which result. (For example, in using numbers to identity a finite set of objects (e.g., members of a team), we often use the segment

of the integers from 1 to n, for suitable n. The appropriate transformation group $\underset{\sim}{P}_n$ can be considered as the subgroup of those elements of $\underset{\sim}{P}_R$ which are otherwise constant.) Proceeding in this way we observe that the categories of real-valued measure functions (classified by the appropriate groups of transformations) form a partially ordered set under the relation of "subgroup", and not a totally ordered set, as you might conclude from the over-simplified table.

4. You will have noticed that we have included the measurement of numerosity as a similarity-invariant measure. Of course the simplest measure of numerosity is the ordinary cardinal number measure. In a certain sense this has a "natural" unit, and there are no "different but equivalent" measures. In this case the set of those transformations of $\underset{\sim}{R}$ which yield (by composition) equivalent measure functions is the single-element group consisting of the identity element only. Thus cardinal number measure could be put in a class by itself, and referred to as "identity-invariant". In practice we do accept other measure functions for numerosity-measurement: in dozens, by the score, by thousands, and so on: These differ from the cardinal number measure by positive similarity transformations, so it is appropriate to include numerosity measures in the similarity-invariant category. If we wished to restrict such measures to those which correspond to integer "units", then the appropriate subset of $\underset{\sim}{S}^+$ would be the semigroup of similarity transformations $\overline{\frac{1}{n}}$, where n is restricted to positive integral values.

One final comment: As you are undoubtedly aware, measurement is not generally an end in itself. For example, in many situations the numbers resulting from measurements are subjected to statistical analyses, leading to the calculation of such statistics as means, modes, standard deviations, and so on. The question of what statistical procedures are appropriate for what types of measurements is strongly related to the classification of measure functions by invariance-groups. You can find this question treated in the articles of Stevens [1] and [9]. It should be pointed out that these articles have stimulated a considerable amount of current controversy, largely revolving about the meaning to be given to "appropriate" in the consideration of the relationship of measure function classification, and "appropriate" statistical procedures.

Chapter 2

THE MEASUREMENT OF NUMEROSITY AND LENGTH

2-1 Introduction

In Chapter 1 we tried to convince you that a functional viewpoint of measurement was both natural and useful. In this chapter we take a much more detailed look at some simple measure functions, especially those for the measurement of numerosity and length. These are "simple" in a sense that will become clearer when we discuss "non-simple", or derived measures, such as area, volume, and velocity. (We shall see that simple and derived are relative, and not absolute terms.) They are also simple in the sense that they represent the outcome of some of man's earliest attempts to come to grips with the idea of measurement.

The history of the development of measurement ideas parallels the history of the development of number ideas, and the inter-relationship of the two is a fascinating study: It is hardly an exaggeration to say that the familiar arithmetic operations of addition and multiplication (for the positive whole numbers and the positive rational numbers) were "invented" in order to satisfy the needs of measurement, especially numerosity measurement, length measurement, and area measurement. But our concern is not so much with the history of the subject of measurement, as it is to give you a conceptual viewpoint which is appropriate to our current level of mathematical development, and which exploits the precision of mathematical ideas to make clear what is involved in the setting up of measure functions, and in such related ideas as units and dimensions.

Roughly speaking, our viewpoint is that we know all about the real number system and its various important sub-systems (natural numbers, integers, rationals, etc.) and their inter-relationships, and that we are interested in describing certain functions, whose domains are sets of real or mathematical objects, and which will, in a sense to be made clear, preserve an empirically suggested or mathematically determined structure of relations and operations. In view of our earlier remarks concerning the way in which our ideas about numbers have been influenced by our ideas about measurement, it is necessary to recall that we now know that the real number system can be logically developed from certain axiomatic assumptions, without the use (except as motivation) of any of the results of measurement processes. This is important, because it would not make much sense to use number properties which ~~depended~~ on measurement, in an attempt to explain a theory of measurement.

Many of you will be familiar with an axiomatic development of the real number system from some appropriate set of axioms. (See, for example, [6], [10].) If you have worked through such a development, you will have learned a great deal not only about the reals, but also about the integers and the rational numbers, and the way in which these various systems are inter-related. As we shall need some of these ideas in fairly precise form, we devote the next section to a brief review of the real number system, with particular emphasis on those ideas needed later on. You will find these ideas treated much more systematically, with most of the necessary proofs, in [6] and [10].

2-2 The Real Number System

We might attempt to explain to you the structure of the real number system by defining it to be a complete ordered field, and then explaining what these terms mean. Unfortunately such a postulational approach tells us nothing about the natural numbers, the integers, and the rational numbers, or how these are related to one another and to the real numbers. If we need to know about these things (as we do for the purposes of this book), we must "work backwards" from the postulated real number system in order to obtain them. In many ways it is simpler, and more instructive, to start with a much more primitive number system, the natural numbers, and show how its properties can be developed from a simple axiomatic description. Then we can define successively the integers, the rationals and the reals, without the need to introduce any new undefined terms, or any additional axioms. This program is carried through in [6] and [10], and in many of the similar books now available. It occupies far too much space to be included here: all we can do is indicate some of the more important steps in the development.

The Natural Numbers. As a starting point (in addition to fundamental ideas from logic and set theory) we take the so-called Peano axioms for the natural numbers. Several different sets of axioms (variations of the set given by the Italian mathematician G. Peano in 1889) go by this name. A suitable set is:

(i) There exists a set N of objects which we call natural numbers. ("Natural number" is an undefined term.)

(ii) There exists a function $\varphi : N \to N$ with the properties

(a) φ is 1-1;

(b) every element of N with one exception, occurs as an image under φ. (We designate this exceptional element by the symbol "1".)

(iii) (Axiom of Induction.) If M is a subset of N, such that

(a) $1 \in M$;

(b) $\varphi(n) \in M$ whenever $n \in M$;

then $M = N$.

If φ is interpreted as the function corresponding to the intuitive idea "addition of 1", these are well known properties of the set of positive integers. What is not so obvious, unless you have gone through it, is that these few properties, taken as axioms, enable us to develop logically a system which has all of the properties which we have learned to associate with the positive whole numbers. In this book, whenever we refer to the natural numbers, or the positive integers, it is this formally-developed system which we have in mind.

The basic properties developed for the natural numbers are:

(i) There exists a binary operation on N, called addition (+) which is associative and commutative, and which has the properties

(a) $m + 1 = \varphi(m)$, for each $m \in N$;

(b) $m + \varphi(n) = \varphi(m + n)$, for each $m, n \in N$.

(ii) There exists a binary operation on N, called multiplication, (denoted by $\cdot$ or by juxtaposition) which is associative and commutative, and which distributes over addition; and which has the properties

(a) $m \cdot 1 = m$ for each $m \in N$;

(b) $m \cdot \varphi(n) = m \cdot n + m$ for each $m, n \in N$.

(iii) There exists an order relation ($<$) on N, defined by: $m < n$, if and only if there exists r, such that $m + r = n$. This relation is connected with the operations of addition and multiplication in such a way that the set $(N,+,<)$ is an ordered semigroup, and the set $(N,\cdot,<)$ is an ordered semigroup with identity. I.e., for $m, n, p \in N$,

(a) $m < n$, $n < p \implies m < p$ (transitivity);

(b) exactly one of the statements $m = n$, $m < n$, $n < m$ is true (trichotomy);

(c) $m < n$ if and only if $m + p < n + p$;

(d) $m < n$ if and only if $m \cdot p < n \cdot p$.

(iv) If an initial segment, I_m, of the natural numbers, is defined to be the set of those natural numbers less than or equal to m, then

(a) for $m \neq n$, there is no 1-1 mapping of I_m onto I_n;

(b) if A and B are disjoint sets, and there exist 1-1 correspondences

$$A \longleftrightarrow I_m, \; B \longleftrightarrow I_n,$$

then there exist 1-1 correspondences

$$A \cup B \longleftrightarrow I_{m+n}$$

and

$$A \times B \longleftrightarrow I_{m \cdot n} .$$

These are, of course, only a few of the properties of the natural numbers. As you are no doubt aware, there is a whole branch of mathematics, called number theory, which is mainly concerned with the natural numbers.

The Integers. From the natural numbers we can proceed (without additional assumption) in either of two directions: we can either define the integers, or we can define the positive rational numbers. In elementary work, in order to obtain the integers we usually postulate zero and the negatives, but we construct the positive rationals as equivalence classes of ordered pairs (fractions) of natural numbers, under the equivalence relation

$$\frac{m}{n} \sim \frac{p}{q} \text{ if and only if } mq = np .$$

Actually this procedure for constructing the positive rationals has its exact counterpart in a construction for the integers, which may be defined as equivalence classes of ordered pairs (m,n) of natural numbers (we can think of these pairs as "formal differences") under the equivalence relation

$$(m,n) \sim (p,q) \text{ if and only if } m + q = n + p .$$

The operations of addition and multiplication, and an order relation, can be introduced into this set of equivalence classes in a natural way, and we find that this new system, J, has all of the properties which we normally associate with the integers, and that J contains a subset, J^+, which is isomorphic to the natural numbers under the correspondence

$$[(n+1,1)] \longleftrightarrow n ,$$

where the left side denotes the equivalence class of $(n+1,1)$. In this sense we may regard J as an "extension" of the natural number system. The main difference between J and N is that J contains the negatives (additive inverses) of the elements which correspond to elements of N, and zero, and that J is a commutative group under addition.

The Rational Numbers. These may be constructed as equivalence classes of ordered pairs of integers (written as $\frac{p}{q}$, $q \neq 0$) under the relation

$$\frac{p}{q} \sim \frac{r}{s} \text{ if and only if } ps = qr .$$

Operations of addition and multiplication, and an order relation, are introduced in a natural way, to yield a system Q, which we call the rational numbers. This system has similar structural properties to J, and in addition the elements of Q, with zero omitted, form a commutative group under multiplication. Q is an example of an ordered field. Q contains a subset which is isomorphic to J under the correspondence

$$[\tfrac{p}{1}] \longleftrightarrow p$$

where p is an integer, and the left side denotes the equivalence class of $\frac{p}{1}$. In this sense we may regard Q as an extension of J.

The order relation in Q has many important properties:

(i) It is dense in the sense that, given any a and $b \in Q$ with $a < b$, there exists at least one (and hence infinitely many) c such that $a < c < b$. (Observe that this is not a property of the order relation for the integers.)

(ii) It is archimedean, in the sense that given any positive rational numbers a, b, there exists at least one (and hence infinitely many) positive integer m such that $ma > b$. (The relation $>$ is, of course, defined in the normal way: $a > b$ if and only if $b < a$.)

The Real Numbers. In spite of the denseness of the ordering of Q, there are many ways in which Q is "incomplete". For example, it is well-known (and easily proved) that there is no rational number q with the property that $q^2 = 2$, and that this lack is highly significant in relation to questions of segment-length in geometry. Q is also incomplete in other ways, which relate to its so-called topological (or continuity) structure. Both of these deficiencies can be overcome at the same time, by using the rationals to construct a new number system, the real numbers. This can be done in a variety of ways (e.g., Dedekind cuts, Cauchy sequences, infinite decimal or binary expansions) which lead to isomorphic systems. For our purposes, the (Dedekind) cut procedure is the most useful, a fact which is not surprising if we remember that Dedekind's idea was directly derived from the method invented by Eudoxus (about 370 B.C.) for the development of a satisfactory theory of proportionality for segments. Eudoxus' procedure may be considered to be a substitute for the fact that no suitable system of numbers (i.e., the real numbers) was then available for the measurement of length. As we shall have to imitate this procedure in our discussion of the measurement of length, you will be able to judge for yourself the greatness of Eudoxus' achievement.

Because of the importance of the idea of a (Dedekind) cut in connection with questions of measurement, we shall describe the idea briefly, and indicate how it leads to a new system of numbers. For the sake of simplicity, let us confine our attention to the development of the positive real numbers from the positive rational numbers. (We designate the set of positive rationals by Q^+; this set is an ordered semigroup under addition, and an ordered group under multiplication.)

We have already observed that there is no rational number whose square is 2. It is easy to prove that there are positive rational numbers whose squares are less than 2 (e.g., 1) and that there are positive rationals whose squares are greater than 2 (e.g., 2); that every positive rational belongs to exactly one of these categories; that every positive rational in the first category is less than every positive rational in the second; that the first category contains no greatest positive rational, and the second category contains no least positive rational; and that if any positive rational belongs to the first (second) category, then every smaller (greater) positive rational also belongs to that category.

These properties of the above partition of the positive rationals are not all independent: some are consequences of the others. After we have constructed the real numbers, we shall find that such a partition is characteristic of a positive irrational number. (I.e., the sets of positive rationals respectively less and greater than a specified positive irrational, constitute such a partition.) What Dedekind did was, in effect, to reverse this idea to obtain the positive irrationals. In addition, in order that the new system of numbers should contain a subset isomorphic to the positive rationals, he introduced a minor modification by permitting the second category to contain a least positive rational. (Each positive rational p determines a partition of the positive rationals into the set of those positive rationals $< p$, and the complementary set of those positive rationals $\geqq p$.) We could now develop the positive reals by defining "ordered partitions" as certain ordered pairs of sets of rationals, and introducing operations and relations into the set of such partitions. Because of the complementary character of the pair of sets in an ordered partition, it is sufficient (and simpler to manage) if we concentrate our attention on the "lower" set in a partition. This we call a cut. More precisely, a cut, C, is a set of positive rational numbers, such that

(i) $C \neq \emptyset$ (the empty set), and $C \neq Q$;

(ii) if $r \in C$ and $q < r$, then $q \in C$;

(iii) if $r \in C$, then there exists $p \in C$, with $p > r$.

We denote the set of all cuts by R^+ and proceed to define operations of addition and multiplication, and an order relation, in R^+ as follows:

$$C_1 + C_2 = \{r_1 + r_2 : r_1 \in C_1, r_2 \in C_2\}$$

$$C_1 \cdot C_2 = \{r_1 \cdot r_2 : r_1 \in C_1, r_2 \in C_2\}$$

$C_1 < C_2$ if and only if there exists $r \in C_2$, such that $r \notin C_1$.

If you think of these cuts as candidates for the role of positive real numbers, and keep in mind that R^+ should contain a subset of rational cuts (corresponding to the rational partitions) which is isomorphic to Q^+, then you will see that these definitions are the natural ones.

Without too much difficulty, it can be shown that the defined operations are associative and commutative, that multiplication distributes over addition, and that order is preserved under each operation in the sense that, for cuts C_1, C_2, C_3,

(i) $C_1 < C_2$ if and only if $C_1 + C_3 < C_2 + C_3$

(ii) $C_1 < C_2$ if and only if $C_1 \cdot C_3 < C_2 \cdot C_3$:

We can also show that R^+ has a multiplicative identity element (the rational cut determined by 1) and that each cut has a multiplicative inverse. (I.e., R^+ is an ordered abelian group under multiplication and an ordered abelian semigroup under addition.)

We can now introduce a zero and negatives (additive inverses), by a procedure which is entirely analogous to the construction of the integers from the natural numbers, to yield the system R of real numbers. Addition, multiplication, and an order relation are defined in a natural way. R is an abelian group under addition; and, with the zero element omitted, it is an abelian group under multiplication. The distributive property holds, so that R is a field. It has an order relation, which is preserved under addition and under multiplication by positive real numbers (i.e., those real numbers greater than the additive identity element) making R an ordered field. The ordering in R is dense and archimedean. (Observe that all of these properties were also properties of the rational numbers.) R contains a subset which is isomorphic to R^+, and in this sense R may be considered as an extension of R^+. Moreover R contains a subset isomorphic to Q, and hence R may be considered as an extension of Q.

A simple, but important, property of R, is that if r is any positive real number, and $C = \{k : k \in Q^+, k < r\}$, then r is the real number which corresponds to the cut C.

It is hardly surprising, because of the method of construction of R^+, that R^+ contains a number (cut) whose square is the cut "2". (I.e., the cut determined by the rational number 2); you can easily verify that if

$$C = \{r : r \in Q^+, r^2 < 2\},$$

then $C^2 = 2$. What is perhaps more surprising, is that R has all of the properties which are implied by the (topological) notion of completeness. We do not need to discuss this idea in detail, so we merely remind you that the idea of completeness is contained in each of the following properties of R, all of which can be proved from the definition of R which we have given:

(i) Every Cauchy sequence in R converges. (A Cauchy sequence (a_n) of real numbers is one which has the property that, given any real $\epsilon > 0$, there exists a positive integer n_ϵ such that for all positive integers $p, q, > n_\epsilon$, $|a_p - a_q| < \epsilon$.)

(ii) Every (non-empty) set of real numbers which is bounded from above has a supremum, or least upper bound (denoted by l.u.b., or sup); every non-empty set bounded from below has an infimum, or greatest lower bound (denoted by g.l.b., or inf).

*(iii) If $R = R_1 \cup R_2$ is a partition of the real numbers (i.e., $R_1 \neq \emptyset$, $R_2 \neq \emptyset$, and $R_1 \cap R_2 = \emptyset$) such that every number in R_1 is less than every number in R_2; then either R_1 contains a greatest real number, or R_2 contains a least real number.

(iv) If (a_n) and (b_n) are non-decreasing and non-increasing sequences of real numbers, with an $\leq$ bn for every n, then the intersection of all closed "intervals" $[a_n,b_n]$ ($[a_n,b_n] = \{x : x \in R, a_n \leq x \leq b_n\}$) is not empty. [This property can be thought of as a "geometric" expression of the notion of completeness. You can prove it by showing that the set A of those real numbers which are less than at least one a_n is non-empty and bounded above, and then proving that the least upper bound, a, of this set belongs to every $[a_n,b_n]$. To see that this property does not hold for the rational numbers, consider increasing and decreasing sequences (a_n), (b_n), of rational numbers, each of which converges to $\sqrt{2}$, and prove that the intersection of the rational "intervals" $[a_n,b_n]$ ($[a_n,b_n] = \{q : q \in Q, a_n \leq q \leq b_n\}$) is empty. (If you are not familiar with the notion of sequential convergence you can find it treated in any good calculus text.)]

Some additional properties of R, which can be proved quite easily, and which are useful in relation to the theory of measurement, are:

(a) Both the rational numbers and the irrational numbers are dense subsets of R in the topological sense; i.e., if r is any rational (irrational) number, then, for every real $\epsilon > 0$, there exists at least one (and hence infinitely many) irrational (rational) number x, such that $|r - x| < \epsilon$. (This implies that every interval $[a,b]$ of real numbers, with $a < b$, contains both rational and irrational numbers.)

(b) Let C be a cut. Then there are strictly monotone sequences (q_n), (q'_n) of rational numbers ((q_n) increasing, (q'_n) decreasing) such that for every n, $q_n \in C$, $q'_n \notin C$, and $q'_n - q_n < \frac{1}{n}$.

(c) If C is a cut, and $q \in C$, then there is a positive integer m, such that for all $n > m$, $q + \frac{1}{n} \in C$.

(d) If C is a cut, and n is a positive integer, then there is a $q \in C$ such that $q + \frac{1}{n} \notin C$, and such that $q + \frac{1}{n}$ is not the least rational which is not in C.

Exercises 2-2

1. Prove the completeness properties (i), (ii), (iii), and (iv) above.
2. Prove the properties (a) -- (d) above.
3. If C_1, C_2, C_3 are cuts, such that

   (a) if $q_1 \in C_1$, $q_2 \in C_2$, then $q_1 + q_2 \in C_3$,

   (b) if $q_1 \notin C_1$, $q_2 \notin C_2$, then $q_1 + q_2 \notin C_3$,

   show that $C_3 = C_1 + C_2$.

   (I.e., $\{q : q = q_1 + q_2; q_1 \notin C_1, q_2 \notin C_2\} = \{q : q \notin C_1 + C_2\}$).
4. Similar to 3., but with multiplication instead of addition.

The Use of Dedekind Cuts. To illustrate the use of cuts, we first prove a theorem which relates to the classification of measure functions as discussed in Section 1-7, and we prove Exercise 1-6.32 as a corollary. The way in which cuts enter into these proofs is typical of the use of real number properties in the theory of measurement. If you work through the details you will see how much simpler matters would be if we were able to restrict our attention to rational numbers only!

Theorem 2-2.1. If $f : R \to R$ is a strictly monotone function which preserves equality of differences, then f also preserves ratios of differences. [This result will be used in the discussion of coordinate systems in Section 2-6.]

Lemma 1. If $x \in R$, $f : R \to R$ is a function which preserves equality of differences, and n is a positive integer, then

$$f(nx) = n(f(x) - f(0)) + f(0)$$

and

$$f(nx_1) - f(nx_2) = n(f(x_1) - f(x_2)) .$$

Proof.

$$nx - (n - 1)x = (n - 1)x - (n - 2)x = \dots = x - 0 .$$

Hence, because f preserves equality of differences,

$$f(nx) - f((n - 1)x) = f((n - 1)x) - f((n - 2)x) = \dots = f(x) - f(0) .$$

Hence, by addition,

$$n(f(x) - f(0)) = f(nx) - f(0)$$

and therefore

$$f(nx) = n(f(x) - f(0)) + f(0) .$$

Hence

$$\begin{aligned} f(nx_1) - f(nx_2) &= n(f(x_1) - f(0)) + f(0) - [n(f(x_2) - f(0)) + f(0)] \\ &= n(f(x_1) - f(x_2)) . \end{aligned}$$

Lemma 2. If $f : R \to R$ is a strictly monotone function which preserves equality of differences, then f is monotone with respect to differences. (I.e., f either preserves or reverses the ordering of differences, according as f is monotone increasing or monotone decreasing.)

Proof. Let x_1, x_2, x_3, $x_4 \in R$, and let $x_1 - x_2 < x_3 - x_4$. Assume also that $x_2 < x_1$, $x_4 < x_3$. (The treatment when the first or both of these inequalities are reversed is entirely similar.) Then there exists x_5 such that

$$x_4 < x_5 < x_3 , \text{ and } x_1 - x_2 = x_5 - x_4 .$$

Hence, because f preserves equality of differences,

$$f(x_1) - f(x_2) = f(x_5) - f(x_4) .$$

If f is monotone increasing,

$$f(x_4) < f(x_5) < f(x_3) ,$$

and hence

$$f(x_1) - f(x_2) = f(x_5) - f(x_4) < f(x_3) - f(x_4) ,$$

and f preserves the order of differences. If f is monotone decreasing,

$$f(x_3) < f(x_5) < f(x_4) ,$$

and

$$f(x_1) - f(x_2) = f(x_5) - f(x_4) > f(x_3) - f(x_4) .$$

Hence f reverses the order of differences.

Proof of Theorem 2-2.1. Let $x_1 , x_2 , x_3 , x_4 \in R$, $x_3 \neq x_4$, and let

$$\frac{x_1 - x_2}{x_3 - x_4} = k .$$

If k is positive and rational, let $k = \frac{m}{n}$, where m and n are positive integers. We wish to show that

$$\frac{f(x_1) - f(x_2)}{f(x_3) - f(x_4)} = k = \frac{m}{n} .$$

We have

$$n(x_1 - x_2) = m(x_3 - x_4) ;$$

i.e., $nx_1 - nx_2 = mx_3 - mx_4$. Hence, since f preserves equality of differences,

$$f(nx_1) - f(nx_2) = f(mx_3) - f(mx_4) .$$

Hence, by Lemma 1,

$$n(f(x_1) - f(x_2)) = m(f(x_3) - f(x_4)) .$$

Since f is strictly monotone, $f(x_3) \neq f(x_4)$, and hence

$$\frac{f(x_1) - f(x_2)}{f(x_3) - f(x_4)} = \frac{m}{n} = k ,$$

as required. The corresponding result for k rational and negative follows immediately if we first reverse the order of one of the differences; for $k = 0$, the result is trivial. Hence f preserves ratios of differences when the ratio is rational.

In order to complete the proof of the theorem, we must show that the result still holds when the ratio of differences is not necessarily rational: this is where we make use of the definition of a real number in terms of cuts. We treat only the case of f monotone increasing: the treatment when f is decreasing is quite similar.

Suppose that $x_3 \neq x_4$, and that

$$\frac{x_1 - x_2}{x_3 - x_4} = r_1 , \quad \frac{f(x_1) - f(x_2)}{f(x_3) - f(x_4)} = r_2 ,$$

and assume that r_1 (and hence, from the strictly monotone property of f, r_2) is positive. (The case r_1, r_2 negative is easily handled, as before.) Let $\frac{m}{n}$ be any rational number in the cut corresponding to r_1, with m and n positive integers. Then

$$\frac{m}{n} < r_1$$

and hence
$$\frac{m}{n} < \frac{x_1 - x_2}{x_3 - x_4} ;$$

i.e.,
$$m(x_3 - x_4) < n(x_1 - x_2) ;$$

i.e.,
$$mx_3 - mx_4 < nx_1 - nx_2 .$$

Hence, from Lemma 2, and the assumption that f is monotone increasing,

$$f(mx_3) - f(mx_4) < f(nx_1) - f(nx_2) .$$

Hence, from Lemma 1,

$$m(f(x_3) - f(x_4)) < n(f(x_1) - f(x_2)) ;$$

i.e.,
$$\frac{m}{n} < \frac{f(x_1) - f(x_2)}{f(x_3) - f(x_4)} = r_2 .$$

Thus $\frac{m}{n}$ belongs to the cut which corresponds to r_2. By an argument which is completely similar, we can show that if $\frac{m}{n}$ does not belong to the cut which corresponds to r_1, then $\frac{m}{n}$ does not belong to the cut which corresponds to r_2. Thus the set of positive rationals less than r_1 is the same as the set of positive rationals less than r_2. Hence $r_1 = r_2$; and the theorem is proved.

Corollary. (cf. Exercise 1-6.32) If $f : R \rightarrow R$ is strictly monotone, and if f preserves equality of differences, then f is a non-singular affine transformation, and hence f is isotone.

Proof. From the theorem, f preserves ratios of differences. Hence

$$x = \frac{x - 0}{1 - 0} = \frac{f(x) - f(0)}{f(1) - f(0)} .$$

Let $f(0) = p$, and let $f(1) - f(0) = q$. Then, because f is strictly monotone, $q \neq 0$. Thus

$$x = \frac{f(x) - p}{q}$$

$$\therefore f(x) = qx + p ;$$

i.e., $$f : x \rightarrow qx + p , \qquad q \neq 0 ,$$

and therefore f is affine and non-singular. Clearly f is onto, hence f is isotone.

Comments:

1. Theorem 2-2.1 and its corollary show that the non-singular affine transformations of R are those isotone transformations which preserve equality of differences (and hence ratios of differences).
2. By comparison (cf. Exercise 1-6.34) the similarity transformations of R are those which preserve ratios.
3. A similarity transformation is, of course, a non-singular affine transformation, and preserves ratios of differences as well as ratios.

As a further example of the properties of the real numbers, we give the promised proof for Exercise 1-6.33. This result will be used many times in the subsequent discussion of measure functions, in relation to the matter of change of unit/change of scale, and in the determination of the structure (as ratio scales) of the sets of admissible length functions, area functions, volume functions, etc.

Theorem 2-2.2.

(a) A function, $f : (R^+,+) \to (R^+,+)$, is a homomorphism, if and only if it is a positive similarity.

(b) Every endomorphism of $(R^+,+)$ is an automorphism.

(c) The set of automorphisms of $(R^+,+)$ is a group (under composition) and this group is isomorphic to $(R^+,\cdot)$.

Proof.

(a) In one direction the proof is trivial: if f is a positive similarity, we leave to you the proof that f is a homomorphism. We shall prove that if f is a homomorphism of $(R^+,+)$, then f is order-preserving, and f is a positive similarity.

If $x, y \in R^+$, $x < y$, then there exists $z \in R^+$, such that $x + z = y$. Hence $f(x) + f(z) = f(y)$, and $f(z) > 0$, therefore $f(x) < f(y)$, and f is order preserving. (I.e., monotone increasing.)

For any $x \in R^+$, and m a positive integer,

$$\begin{aligned} f(mx) &= \underset{(m\ \text{terms})}{f(x + x + \dots + x)} \\ &= \underset{(m\ \text{terms})}{f(x) + f(x) + \dots + f(x)} \\ &= mf(x). \end{aligned}$$

Hence $$f(x) = f(m\tfrac{x}{m}) = mf(\tfrac{x}{m}),$$

so that $$f(\tfrac{x}{m}) = \tfrac{1}{m} f(x).$$

Combining these results, we get

$$f(\tfrac{m}{n} x) = \tfrac{m}{n} f(x)$$

for every positive rational number $\frac{m}{n}$. Thus, if q is rational, and if $f(1) = k > 0$, we have $f(q) = qf(1) = kq$.

Now suppose that r is any positive real number, and let $f(r) = t$. If r is rational, then we have shown above that $t = kr$. If r is not rational, we know (from trichotomy) that exactly one of $(t = kr, t < kr, t > kr)$ holds.

If $t < kr$, then $\frac{t}{k} < r$. Hence there exists a rational number q with $\frac{t}{k} < q < r$. Since f preserves order, we have $f(q) < f(r)$. But q is rational, hence $f(q) = kq < f(r) = t$. But $\frac{t}{k} < q$; i.e., $t < kq$. Hence we have a contradiction, and therefore $t \nless kr$. Similarly $t \ngtr kr$, so that $t = kr$, and the proof that f is a positive similarity, is complete. Parts (b) and (c) are left for you to prove: the proofs are quite straightforward.

Corollary. Every endomorphism of $(R^+,+)$ is a non-singular, homogeneous linear function.

Proof. Homogeneity is all that remains to be proved. We have shown that there is a $k \in R^+$ such that, for all $x \in R^+$, $f : x \to kx$. If now c is any positive real number, $f(cx) = k(cx) = c(kx) = cf(x)$, hence f is homogeneous of degree 1.

Exercises 2-2 (continued)

5. If a, b, are positive real numbers, denote by $A = (0,a)$, $B = (0,b)$, the "open" initial segments of positive reals less than a and b, respectively. Each of the sets A, B, has a structure with respect to addition, and it has the usual order, but neither is an additive semigroup, because neither set is closed under addition. A homomorphism $f : (A,+) \to (R^+,+)$ is a function which preserves the (incomplete) additive structure. Prove that

   (a) every such homomorphism is monotone increasing and 1-1;

   (b) there is a unique b such that f is an isomorphism from A to B, and $f(x) = \frac{b}{a} x$ for every $x \in A$.

   (This exercise is related to the measurement of angles.)

6. Prove that the set of automorphisms of the ordered group $(R,+,<)$ is the positive similarity group $\underline{S}^+$. (Remember to consider the behavior of the negative numbers.)

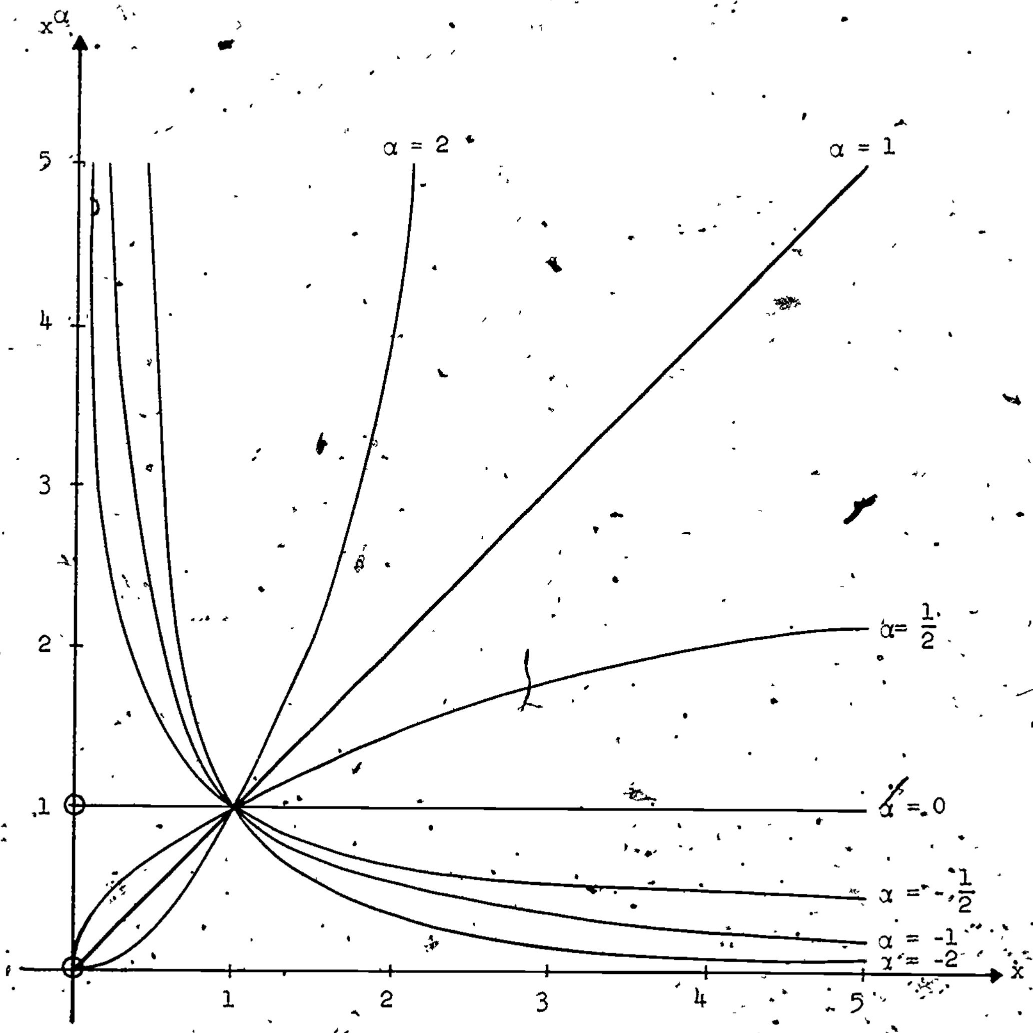
7. Similarly prove that the automorphism groups of $(Q,+,<)$ and $(Q^+,+,<)$, are the respective positive rational similarity groups. (i.e., the group of transformations $k : x \to kx$, for x rational (positive rational) and k a fixed positive rational.)

8. If f is an endomorphism of $(R^+,+)$, prove that f is fully determined by its value on a single element of R^+.

9. If $f : R \to R$ is a monotone function which is not 1-1 (i.e., for some a, b, with $a \neq b$, $f(a) = f(b)$) and which preserves equality of differences, prove that f is a constant function. [Hint: First prove that f is constant on the interval $[a,b]$, then prove that f has the same value at every point $a + n|b - a|$, for all integral n, then use the archimedean property.]

Our next theorem concerns the monotone endomorphisms of the multiplicative group $(R^+, \cdot)$ of positive real numbers. The result of the theorem (that all such endomorphisms are power functions) is needed for the discussion of the theory of "dimension", in relation to those categories of measure functions which are ratio scales. In the proof of the theorem we need to use basic properties of power functions, and also properties of logarithmic and exponential functions, so we review these briefly before stating and proving the theorem.

Power Functions. If α is any real number, the function $f : R^+ \to R^+$ defined by $f : x \to x^\alpha$ is called a power function. As you know, if x is a real number, and if α is a positive integer, x^α is defined as an iterated product; but when α is fractional or irrational, the definition depends on the deeper properties (i.e., completeness) of the real numbers, and x^α can only be defined for all α, if x is positive. To remind you of the way power functions "look" for various values of α, a number of the graphs of power functions are illustrated in the diagram below:



Partial Graphs of Functions $x \to x^\alpha$ on the Domain R^+

We summarize the important properties of the power functions $f_\alpha : x \to x^\alpha$, defined on R^+:

(i) If $\alpha = 0$, f_0 is the constant function $f_0 : x \to 1$ for every $x \in R^+$.

(ii) If $\alpha \neq 0$, f_α is a 1-1 correspondence of R^+ onto R^+; in this case the inverse of f_α is also a power function, and $(f_\alpha)^{-1} = f_{1/\alpha}$.

(iii) If $\alpha > 0$, f_α is monotone increasing.

(iv) If $\alpha < 0$, f_α is monotone decreasing.

(v) The composite of two power functions is a power function, and $f_\alpha f_\beta = f_{\alpha\beta}$.

(vi) For each value of α, f_α is continuous.

[We do not wish to get too deeply involved with the concept of continuity, which is probably familiar to you from your courses in calculus. You can find a formal definition in any good calculus text; informally, continuity simply means that, for all "sufficiently close" arguments, the values must be "arbitrarily close" to one another.]

Exponential and Logarithmic Functions. The exponential and logarithmic functions are closely related to the power functions. If, in the diagram above, you imagine the graphs drawn for all $\alpha \in R$, and imagine the vertical line drawn through any point $x = a$ ($a \in R^+$, $a \neq 1$) then this ordinate will intersect the graph of each power function at exactly one point. (This is not completely obvious from the partial graphs which we have illustrated, but you can easily verify that the assertion is true.) These points of intersection are the points (a, a^α), one for every real α. The set of pairs (α, a^α) is a function with domain R. In other words, for each $a \neq 1$, this process gives a 1-1 function $x \to a^x$, whose domain is the set of all real numbers and whose range is in R^+. It is not hard to show that this function is onto R^+, and hence it is a 1-1 correspondence of R and R^+. These functions $x \to a^x$, one for each positive number a except the number 1, are, of course, the exponential functions, and their inverses are the logarithmic functions. If you refer again to the power function graphs, and picture the ordinates $x = a$, then you will see that if $a > 1$, a^α increases as α increases; while if $a < 1$, a^α decreases as α increases. That is, the exponential function $x \to a^x$ is monotone increasing if $a > 1$, and monotone decreasing if $a < 1$. We summarize some of the properties of the exponential and logarithmic functions, and illustrate the graphs of the exponential functions for several values of a. (The graphs of their inverses, the logarithmic functions, may be easily obtained from these.)

(i) Each exponential function $x \to a^x$ is a 1-1 correspondence from R to R^+; the inverse, which is the logarithmic function $x \to \log_a x$, is a 1-1 correspondence from R^+ to R ...

(ii) For $a > 1$, the function $x \to a^x$ is monotone increasing, and the function $x \to \log_a x$ is monotone increasing.

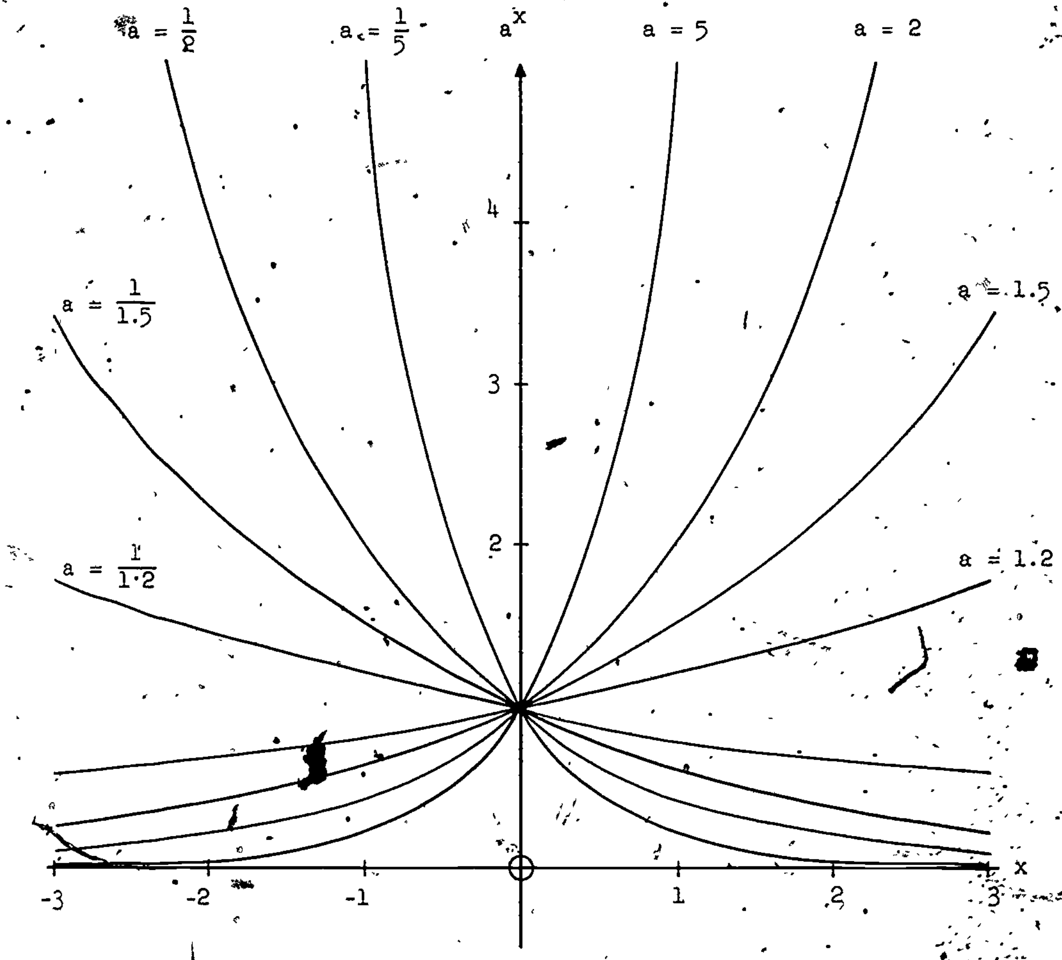
(iii) For $a < 1$, the functions $x \to a^x$ and $x \to \log_a x$ are each monotone decreasing.

(iv) Each exponential function is an isomorphism from $(R,+)$ to $(R^+, \cdot)$; i.e., $a^{x_1+x_2} = a^{x_1}a^{x_2}$.

(v) Each logarithmic function is an isomorphism from $(R^+, \cdot)$ to $(R,+)$; i.e., $\log_a(x_1x_2) = \log_a x_1 + \log_a x_2$.

(vi) $a^x = (\frac{1}{a})^{-x}$, and $\log_a x = -\log_{1/a} x$.

[The first of the properties (vi) is reflected in the diagram below, in the symmetric relationship of the corresponding graphs.]



Partial Graphs of Functions $x \to a^x$ $(a > 0, a \neq 1)$

Theorem 2-2.3.

(a) Every monotone endomorphism f of the group $(R^+, \cdot)$ is a power function, and every power function on R^+ is a monotone endomorphism.

(b) The set of monotone automorphisms of the group $(R^+, \cdot)$ is a group under composition, and this group is isomorphic to the multiplicative group $(R - 0, \cdot)$ of the non-zero real numbers, under the correspondence $f \longleftrightarrow \alpha$ (where f is, of course, the function $f : x \to x^\alpha$.)

Proof.

(a) Let $f : (R^+, \cdot) \to (R^+, \cdot)$ be a monotone endomorphism. Then $f(1) = 1$. Let $f(2) = b$. (We could use any $a \neq 1$ instead of the number 2.) Then there is an $\alpha \in R$ such that $b = 2^\alpha$. That is, $f(2) = 2^\alpha$. (If $\alpha > 0$, then $2^\alpha > 1$; hence f must be monotone non-decreasing; if $\alpha < 0$, f must be monotone non-increasing.) We shall show that $f(x) = x^\alpha$, for every $x \in R^+$. From the homomorphism property of f, it is easy to show that $f(2^n) = [f(2)]^n$ for every positive integer n, hence for every number 2^n, $f(2^n) = (2^\alpha)^n = 2^{n\alpha} = (2^n)^\alpha$. In the same way, $f(2) = f[(2^{1/n})^n] = [f(2^{1/n})]^n$, so that $f(2^{1/n}) = (2^\alpha)^{1/n} = (2^{1/n})^\alpha$. We can combine these results to obtain $f(2^{m/n}) = (2^{m/n})^\alpha$ for every positive rational number $\frac{m}{n}$. We wish to extend this result to show that $f(2^y) = (2^y)^\alpha$ for every y in R, because the set of all such numbers 2^y is the whole of R^+.

Firstly, suppose that α is 0. In this case, $f(2^q) = (2^q)^0 = 1$ for every positive rational q, and hence, from monotonicity, $f(2^y) = 1 = (2^y)^0$ for every positive number y. We can deal with y negative (whether or not α is zero) after dealing with the case α non-zero, y positive.

Suppose next that f is monotone and non-decreasing (i.e., $\alpha > 0$). We seek to prove that, for all $y > 0$, $f(2^y) = (2^y)^\alpha$. Suppose that this is not true for some $y = r > 0$, and that $f(2^r) < (2^r)^\alpha$. Then, $2^r > 1$ and hence $f(2^r) \geq 1$. That is

$$1 \leq f(2^r) < (2^r)^\alpha .$$

Hence, using the fact that the function $x \to x^{1/\alpha}$ is monotone increasing, we obtain

$$1 \leq [f(2^r)]^{1/\alpha} < 2^r .$$

Hence, from the monotone increasing property of the function $x \to \log_2 x$, and putting $z = \log_2([f(2^r)]^{1/\alpha})$, we get

$$0 \leq z < r .$$

Hence, there is a positive rational number q, such that

$$z < q < r ,$$

and therefore, using the monotone increasing property of the function $x \to 2^x$;

$$[f(2^r)]^{1/\alpha} = 2^z < 2^q < 2^r .$$

Hence, from the monotone increasing property of the function $x \to x^{\alpha}$, $(\alpha > 0)$,

$$(1) \qquad f(2^r) < (2^q)^{\alpha} < (2^r)^{\alpha} .$$

But f is monotone non-decreasing, therefore, since $2^q < 2^r$,

$$f(2^q) \leq f(2^r) .$$

Because q is rational, $f(2^q) = (2^q)^{\alpha}$, hence

$$(2^q)^{\alpha} \leq f(2^r) .$$

But this contradicts (1), and hence $f(2^r) \not< (2^r)^{\alpha}$. Similarly we can prove that $f(2^r) \not> (2^r)^{\alpha}$. Hence $f(2^y) = (2^y)^{\alpha}$ for all positive y, and $\alpha \geq 0$. If $y = 0$, this statement becomes $f(1) = 1$, which we have seen is true because f is a homomorphism. If $y < 0$, then we have

$$1 = f(1) = f(2^0) = f(2^y \cdot 2^{-y}) = f(2^y) \cdot f(2^{-y}) .$$

Hence $f(2^y) = [f(2^{-y})]^{-1}$. But $-y$ is positive, hence

$$\therefore f(2^y) = [(2^{-y})^{\alpha}]^{-1} = (2^y)^{\alpha} .$$

Hence, for all real y,

$$f(2^y) = (2y)^{\alpha} .$$

That is, for all $x \in R^+$,

$$f : x \to x^{\alpha} .$$

The treatment when $\alpha < 0$ is entirely similar, and we conclude that every monotone endomorphism of $(R^+, \cdot)$ is a power function. The converse (that every power function is a monotone endomorphism) follows directly from well-known properties of the power functions.

(b) Because each monotone endomorphism is a power function $x \to x^{\alpha}$; a monotone endomorphism is a monotone automorphism if and only if $\alpha \neq 0$, and there is thus a 1-1 correspondence $\varphi : f \to \alpha$ of monotone automorphisms and non-zero real numbers.

If $f_1 : x \to x^{\alpha_1}$, $f_2 : x \to x^{\alpha_2}$, with α_1 and α_2 non-zero, then the composite function $f_1 f_2$ takes $x \to (x^{\alpha_2})^{\alpha_1} = x^{\alpha_1 \alpha_2}$. Hence $\varphi(f_1 f_2) = \varphi(f_1)\varphi(f_2)$ and φ is an isomorphism.

Remarks:

1. The above proof is quite lengthy, but not inherently difficult. As is common in similar theorems (e.g., Theorem 2-2.2) it is quite easy to go from the homomorphism property to arguments involving the rational numbers (in this case the rational powers 2^q) and the difficult step is to move to an argument which applies for all real (or positive real) numbers. In the above theorems, we were able to achieve this extension by using monotone properties (proved in Theorem 2-2.2; part of the hypothesis of Theorem 2-2.3). We could equally well have assumed continuity (in Theorem 2-2.3) and achieved the extension from the rationals to the reals by a "continuity" argument. Carried through in detail, this is more complicated than the argument from monotonicity. This is not surprising, as continuity is a more complex notion.

2. The proof of Theorem 2-2.3 could have been deduced from that of Theorem 2-2.1 by using a logarithmic/exponential isomorphism of $(R^+, \cdot)$ and $(R,+)$, and the fact that such an isomorphism is monotone. This idea is conveyed in the exercises below, in which we have an "additive" equivalent of Theorem 2-2.3.

Exercises 2-2 (continued)

10. Let $f : (R,+) \to (R,+)$ be a monotone homomorphism. Prove that
    (a) if, for any $b \neq 0$, $f(b) = 0$, then f is the constant function $f : x \to 0$ for all $x \in R$;
    (b) if f is not the constant function, then f is 1-1, and f is strictly monotone and preserves the equality of differences.

11. Use Exercise 10, and Theorem 2-2.1, to prove that if $f : (R,+) \to (R,+)$ is a monotone homomorphism, then there is a $k \in R$, such that $f : x \to kx$ for every $x \in R$. (I.e., f is either the constant "zero" function, or f is a similarity transformation of R.)

12. Prove Exercise 11 (as a direct consequence of Theorem 2-3.3) by using the exponential/logarithmic relationship, $x \longleftrightarrow 2^x$, as illustrated in the commutative diagram below, in which the function g is defined (by composition) so as to make the diagram commutative:

$$\begin{tikzcd}
x \arrow[d, leftrightarrow] & (R,+) \arrow[r, "f"] \arrow[d, leftrightarrow] & (R,+) \arrow[d, leftrightarrow] & x \arrow[d, leftrightarrow] \\
2^x & (R^+,\cdot) \arrow[r, dashed, "g"'] & (R^+,\cdot) & 2^x
\end{tikzcd}$$

13. If $f : (R^+,\cdot) \to (R^+,\cdot)$ is a homomorphism, which is monotone on an open interval $(a;b)$ $(a \neq b)$, prove that f is monotone on R^+. [Hint: proceed as in the proof of Theorem 2-2.3, to prove that $f : x \to x^{\alpha}$; first, for all $x = 2^q$ (all rational q); and then for all positive real x in (a,b). Then show that any $z \in R^+$ can be written as $z = xy$, where $x \in (a,b)$ and $y = 2^q$ for some rational q, and use the homomorphism hypothesis to show that f is a power function, and hence monotone.]

14. If $f : (R,+) \to (R,+)$ is a homomorphism, which is monotone in any open interval (a,b) $(a \neq b)$ prove that f is a similarity transformation, and hence monotone on R.

15. Prove that $f : (R,+) \to (R,+)$ is an endomorphism of the group $(R,+)$ if and only if $f(0) = 0$ and f preserves the equality of differences.

16. Prove that $f : (R^+,\cdot) \to (R^+,\cdot)$ is an endomorphism of the group $(R^+,\cdot)$ if and only if $f(1) = 1$, and f preserves the equality of ratios. [You should prove this directly, and also indirectly by using Exercise 15 and the exponential/logarithmic relationship as in Exercise 12 above.]

17. If $f : R \to R$ preserves the equality of differences, prove that the function $g : R \to R$ defined by $g : x \to f(x) - f(0)$, is an endomorphism of the group $(R,+)$.

18. If $f : R^+ \to R^+$ preserves the equality of ratios, prove that the function $g : R^+ \to R^+$ defined by $g : x \to \frac{f(x)}{f(1)}$, is an endomorphism of the group $(R^+, \cdot)$.

Remark: The various results proved in this section and stated in the exercises, illustrate the close parallelism between the homogeneous affine functions on R, and the power functions on R^+; and between the affine functions on R, and the positive-constant multiples of the power functions on R^+. (I.e., the functions $x \to cx^\alpha$, $c > 0$, $\alpha \in R$.)

Exercises 2-2 (continued)

19. If $f : R^+ \to R^+$ is a monotone function, such that the restricted function $f|Q^+$ maps $(Q^+,+)$ homomorphically into $(R^+,+)$, prove that f is a positive similarity.

Remark: The result of Exercise 19 above can be further generalized. If you are familiar with the notions of topological density, and continuity, you can prove that if $f : R^+ \to R^+$ is a monotone (alternatively: continuous) function with the property that $f : y \to ky$ $(k \in R^+)$ for each y in a dense subset of R^+, (e.g., Q^+) then $f : x \to kx$ for every $x \in R^+$.

Powers of Functions. Let A be any set, and let $f : A \to R^+$. For every real number α, we can define a function $f^\alpha : A \to R^+$ by

$$f^\alpha(a) = [f(a)]^\alpha .$$

The function f^α is called "f to the power α", or "f to the α". Thus, for each α, the power operation is an operation in the set G of all functions from A to R^+. You should prove that this operation has the following properties, related to the algebraic structure of G, which was discussed in Section 1-5:

(i) $f^0 = \underline{1}$ (the constant function).

(ii) $(f^\alpha)^\beta = (f^\beta)^\alpha = f^{\alpha\beta}$.

(iii) $(f_1 \cdot f_2)^\alpha = f_1^\alpha \cdot f_2^\alpha$.

(iv) $(kf)^\alpha = k^\alpha f^\alpha$ $(k > 0)$.

If you remember that many of the common measure functions have values in R^+, you should not be surprised to learn that we shall need to use the "power" operation in our study of the relationships between ratio scales, when we consider the subject of "dimension".

Exercises 2-2 (continued)

20. Let $d : R \times R \to R$ be the difference operation (i.e., $d(x,y) = x - y$) and let $f : R \to R$ be a function which preserves the equality of differences. Prove that the function $g : R \to R$, defined in Exercise 17 above, makes the following diagram commutative:

$$\begin{tikzcd}
R \times R \arrow[r, "f \times f"] \arrow[d, "d"'] & R \times R \arrow[d, "d"] \\
R \arrow[r, "g"'] & R
\end{tikzcd}$$

Conversely, prove that if, given $f : R \to R$, there exists a function $g : R \to R$ which makes the above diagram commutative, then f preserves the equality of differences. (In other words, for any given $f : R \to R$, the existence of a function g which makes the diagram commutative, is equivalent to the condition that f should preserve the equality of differences.) Now use the result of Exercises 17 and 12 to give an alternate proof of Theorem 2-2.1 and its Corollary.

2-3 The Measurement of Numerosity

The ideas involved in the empirical measurement of numerosity (a clumsy word, but probably better than numerousness!) are so simple, and the structure of the most appropriate value set (the set of positive integers, whose prehistorical development was undoubtedly due to the empirical properties of the measurement operation known as "counting") is so closely related to the attribute that we wish to measure, that numerosity measurement is ignored in many treatments of the general subject of measurement. We shall not ignore it, because it is instructive to examine the measurement of numerosity in the same spirit that we shall use later in the discussion of length measurement, and in other more complicated measurement situations.

As we indicated earlier, we shall assume that we are fully familiar with the real number system and its various subsystems, and that (except for motivation) we have developed these systems without direct appeal to the real world and empirical ideas. We will have some difficulties, due to the fact that most of us work with an intuitive rather than an axiomatic set theory. It is hoped that this will not obscure the main features of the development.

In outline, the informal idea of what we are going to try to do when setting up measure functions, is as follows:

(i) There is an attribute which we set out to measure. We cannot define this attribute, but our feeling that certain objects (which make up the domain) possess it, is sufficiently clear that we are prepared to establish empirical procedures which, generally aided by induction (i.e., empirically-inspired guess work, generalizing from particular cases; not mathematical induction, which is really deduction) enable us to give a structure to the domain. In most cases this will involve the establishment of relations and operations on the set of objects in the domain. The domain will thus be given an empirical structure. In many cases the real numbers, or some appropriate subsystem, will have a comparable structure (and usually much additional structure).

(ii) We select an appropriate subset (call it V) of the reals, and we ask ourselves the following questions:

(a) Is it possible to map the domain into V in such a way that the structure is preserved? I.e., the mapping should be a homomorphism. [Sometimes the mapping is 1-1, (i.e., an isomorphism), and sometimes it is onto. Frequently we are able to establish an (empirical) equivalence relation on the domain (related to the selected attribute which we wish to measure), and we look for a suitable structure-preserving mapping on the resulting set of equivalence classes.]

(b) If such a mapping exists, is it unique?

(c) If more than one structure-preserving mapping (measure function) exists, is there any relationship between these functions? (E.g., can we get from one to another by composition with an appropriate automorphism of V ?)

Let us keep this plan in mind, and return to the question of numerosity measurement. The ideas of attribute, domain, and "equivalence with respect to the attribute" tend to blur together, but roughly speaking, the idea we have in mind for numerosity relates to the empirical process of pairing-off, or matching. As elements of the domain, D, we think of sets of objects such as bags of stones, or flocks of sheep, and we confine our attention to finite sets. We must say what we mean by "finite", so, to make matters as simple as possible, we draw on our assumed knowledge about the number systems and define a set to be <u>finite</u> if there exists a 1-1 correspondence between the set and a segment $(1,n)$ of the positive integers. (We could, if necessary, avoid the use of J^+ at this stage of the development.)

In the domain D, we can compare elements (i.e., finite sets) with respect to numerosity by the empirical process of pairing-off members of the sets: if we can complete this pairing-off of the members of two finite sets S, T without having any members of either set left over, we say that S has the same numerosity as T, or that S and T are equally numerous. (We write this $S \sim T$; we may read it "S is equally numerous to T".) As far as we can tell empirically, $\sim$ is an equivalence relation on our domain. (Due to the difficulty mentioned earlier, you probably find it hard to picture $\sim$ as an experimentally-defined relation, whose transitivity, for example, must be verified experimentally; when we study length measurement, the corresponding situation should be more clear.) If our domain is infinite, (or very large) we cannot test all possibilities, so we use induction and simply guess that, if we had unlimited time and patience, we could verify that $\sim$ is an equivalence relation. (Scientists make inductive guesses like this all the time, generally in much less "obvious" situations, with the expectation that, should they have guessed wrongly, this will eventually become empirically evident.) We denote the equivalence class of a set S by $\tilde{S}$, and the set of all equivalence classes by $\tilde{D}$.

The very same empirical process (pairing) which enabled us to organize our domain into classes of "equally numerous" sets, enables us to define another relation, that of "less numerous than" ($<$) : $S < T$ if, in the pairing process, the members of S are exhausted first; i.e., each member of S is paired with one of T, and there are elements of T remaining. We assume that, as a result of many trials, we have discovered that this relationship does not appear to depend on the way in which we go about the matching process, and that as far as we have tested it, the relation $<$ is transitive. We also discover that, for any two finite sets S and T, exactly one of the following holds: $S \sim T$, $S < T$, $T < S$; and that if $S_1 \sim S_2$, $T_1 \sim T_2$, and $S_1 < T_1$, then $S_2 < T_2$. From all of this we feel justified in assuming that the relation $<$ is transitive on D; that it yields a corresponding relation (for which we use the same symbol) on the set $\tilde{D}$ of equivalence classes of D; and that this is a (strict total) order relation on $\tilde{D}$. In other words, the relation $<$ on $\tilde{D}$, defined by

$$\tilde{S} < \tilde{T} \text{ if } S_1 < T_1 \text{ for any } S_1 \in \tilde{S}, T_1 \in \tilde{T},$$

is an order relation on $\tilde{D}$.

We consider next whether any other empirical structure can be established in D, bearing in mind the attribute (numerosity) in which we are interested. Several possibilities need to be considered: for example, the operations of union and intersection can be carried out as physical operations, and experience would suggest the assumption that each is associative and commutative, and that each distributes over the other. However, not all of this structure seems to be immediately relevant to our intuitive idea of numerosity. After some consideration of possibilities, let us imagine that we have concluded that the operation of union for disjoint sets is going to be relevant, and that (bearing in mind our eventual objective of setting up a function to the positive integers) unions of disjoint sets should map into sums, under a numerosity-measure function. At this point a difficulty appears: not all pairs of sets in our domain are disjoint, and even if (for $S_1, S_2, S_3 \in D$), $S_1 \cap S_2 = S_2 \cap S_3 = \emptyset$, it is not necessarily true that $S_1 \cap S_3 = \emptyset$. On the other hand, we discover that if $S_1 \cap S_2 = \emptyset$, $S_1' \cap S_2' = \emptyset$ and $S_1 \sim S_1'$, $S_2 \sim S_2'$, then $S_1 \cup S_2 \sim S_1' \cup S_2'$. This suggests that we might shift our attention from D to $\tilde{D}$, and consider what further empirical numerosity structure we can establish on $\tilde{D}$. We could then look for a structure-preserving map from $\tilde{D}$ to the set J^+ of positive integers. While thinking along these lines, we recall that the natural mapping $D \to \tilde{D}$ (defined by

$S \to \tilde{S}$, for each finite set S) preserves order, and that if this were to be composed with a suitable mapping from $\tilde{D}$ to J^+, the composite map would have the desired property of mapping equally numerous sets into the same positive integer.

We now proceed to establish an operation $\underline{\cup}$ in $\tilde{D}$ as follows: if $\tilde{S}$, $\tilde{T} \in \tilde{D}$, select $S_1 \in \tilde{S}$, $T_1 \in \tilde{T}$, such that $S_1 \cap T_1 = \emptyset$. (This involves an assumption for which we expect to have empirical support.) Now define $\tilde{S} \underline{\cup} \tilde{T} = \widetilde{S_1 \cup T_1}$, (The notation on the right denotes, of course, the equivalence class of $S_1 \cup T_1$; it is sometimes convenient to indicate this "disjoint union" by the notation $S_1 \underline{\cup} T_1$.) Empirical examination of the properties of the operation $\underline{\cup}$ on $\tilde{D}$ suggests that we are justified in making the following assumptions:

(i) $\underline{\cup}$ is associative and commutative, and $\tilde{D}$ is closed under $\underline{\cup}$;

(ii) $\underline{\cup}$ preserves order, in the sense that for $\tilde{S}_1$, $\tilde{S}_2$, $\tilde{T} \in \tilde{D}$,
$$\tilde{S}_1 < \tilde{S}_2 \Longrightarrow \tilde{S}_1 \underline{\cup} \tilde{T} < \tilde{S}_2 \underline{\cup} \tilde{T}.$$

In other words, the system $(\tilde{D}, \underline{\cup}, <)$ is an ordered abelian semigroup.

We now turn our attention to the question of whether there is a mapping of $\tilde{D}$ into J^+ (the set of positive integers) which preserves this "numerosity structure". Actually there are two possibilities to consider: J^+ is an ordered abelian semigroup with respect to addition, and it is also an ordered abelian semigroup with respect to multiplication, so we should keep both possibilities in mind. The procedure by which we attempt to set up such a suitable function is, of course, the well known process of counting: given any $\tilde{S} \in \tilde{D}$, we take any $S \in \tilde{S}$, and pair off the elements of S with the ordered set of positive integers 1, 2, 3, ... Our assumption of finiteness ensures that this process will terminate at some positive integer s, and it seems natural to test the relation $\phi = \{(\tilde{S}, s)\}$ to see whether it yields a suitable structure-preserving function. We must first make certain that, for given $\tilde{S}$, s is uniquely determined; i.e., that the relation is single-valued. That this is so, follows from the fact that the counting procedure establishes a 1-1 correspondence between S and the segment $(1, s)$, so that if, using a different set from $\tilde{S}$, or by a different pairing in the counting procedure, we were to arrive at a different positive integer s', the transitivity of the 1-1 correspondence relation would imply that there existed a 1-1 correspondence between the segments $(1, s)$ and $(1, s')$. But (as mentioned in the previous section) the impossibility of this can be proved within the formal system J^+. Thus the counting process has established a function $\varphi : D \to J^+$, and a related function $\underline{\varphi} : \tilde{D} \to J^+$. An argument similar to those used above shows that $\underline{\varphi}$ is 1-1.

We consider next the way in which the counting process is related to the empirical order in D , and the resulting order in $\tilde{D}$. For $\underline{S} < \underline{T}$, the empirical procedure for establishing this relation yields a 1-1 correspondence of S with a proper subset of T . It follows that if $\varphi(S) = s$, $\varphi(T) = t$, then there is a 1-1 correspondence of the segment $(1,s)$ with a proper subset of the segment $(1,t)$. From this it can be proved, purely within the formal system J^+ , that $s < t$. Hence φ and $\underline{\varphi}$ preserve order.

We look next to see whether $\underline{\varphi}$ carries "unions" in $\tilde{D}$ into either sums or products in J^+ . We soon discover that "unions" are not carried into products. As far as sums are concerned, it can be proved, within J^+ , that the segment $(1,t)$ corresponds 1-1 with the set $(s + 1 , s + t) = \{n \in J^+ : s + 1 \leq n \leq s + t\}$; that $(1,s) \cap (s + 1 , s + t) = \emptyset$; and that $(1,s) \cup (s + 1 , s + t) = (1 , s + t)$. Hence $\underline{\varphi}$ is an isomorphism of the ordered semigroup $(\tilde{D} , \underline{U} , <)$ into the ordered semigroup $(J^+, +, <)$.

It is natural to ask whether $\underline{\varphi}$ (and hence φ) is onto. Before we can consider this question seriously, we have to be much clearer about the domain D than we have been. If we restrict D to sets of material objects, which we might think of as objects composed of fundamental physical particles, our question would be equivalent to asking whether the number of fundamental particles in our universe is finite. This is not a question that can be easily answered! In this connection we point out that the question of the ontoness of $\underline{\varphi}$ is closely related to the empirical properties which we have assumed (inductively) for $\tilde{D}$, and to our description of the operation $\underline{U}$. If S is any 1-element set in D , our assumption that the operation $\underline{U}$ can always be performed, implies that we can iterate the operation to obtain $\tilde{S}_n = S \,\underline{U}\, S \,\underline{U}\, S \ldots \underline{U}\, S$ (n terms) for every $n \in J^+$, and hence we conclude that $\underline{\varphi}$ is onto. If this operation becomes impossible for some $m + 1 \in J^+$, it means that we have a set S_m with m objects, and that there is no single-element set which is disjoint with $\underline{S}_m$. That is, all elements of our "universe" belong to S_m , and our "universe" of objects is finite. This means that our assumption that $\underline{U}$ can always be carried out as an empirical operation (for any pair of sets not necessarily disjoint) is invalid, and hence the system $(\tilde{D},\underline{U})$ is not a semigroup. On the other hand, if we allow D to include sets of mathematical objects, such as natural numbers, then it is easy to show that $(\tilde{D} , \underline{U})$ is a semigroup, and that $\underline{\varphi}$ is onto J^+ .

The measure function which we have set up for the measurement of our intuitive notion of numerosity, is known as the *cardinal measure* of a finite set, and its value on a set is the *cardinal number* of that set. At this stage there are several questions which we might reasonably ask ourselves:

(i) Is φ the only numerosity measure on D, such that $\underline{\varphi}$ is an isomorphism of the ordered semigroup $(\tilde{D}, \underline{\cup}, <)$ into $(J^+, +, <)$? And what if we consider other value spaces such as Q^+ or R^+?

(ii) Is it possible to extend the domain of φ, so that φ still corresponds to our intuitive idea of a numerosity measure, by dropping the restriction that sets in D be finite?

(iii) What about the other operation (multiplication) in J^+? Does it fit into the picture anywhere?

The second question is of considerable mathematical interest, and its investigation led Cantor to the theory of transfinite cardinals. You will not be surprised to learn that these are important in the mathematical theory of measure.

The first question has two parts. In order to attack the first part, let us make the assumption that the cardinal number measure function φ is onto J^+. This implies that $\underline{\varphi}$ is an isomorphism. If there exists a second isomorphism

$$\underline{\varphi}' : (\tilde{D}, \underline{\cup}, <) \to (J^+, +, <)$$

(not necessarily onto), then we can combine the (compositional) inverse of $\underline{\varphi}$ with $\underline{\varphi}'$ and obtain an isomorphism $\underline{\varphi}' \, \underline{\varphi}^{-1} = \rho$ (say) of the ordered semigroup $(J^+, +, <)$ into itself. Let us see what this implies: if $\rho(1) = k$, then

$$\rho(n) = \rho(\underbrace{1 + 1 + \ldots + 1}_{(n \text{ terms})}) = \underbrace{\rho(1) + \rho(1) + \ldots + \rho(1)}_{(n \text{ terms})} = n\rho(1) = nk .$$

Let ρ_k denote the function $n \to nk$ for each $n \in J^+$. We check readily that ρ_k is, in fact, an isomorphism of J^+ into a proper subset of itself, the ordered semigroup of all positive multiples of k. (See Exercise 1-6.35.)

Thus, for each positive integer k, the composite map $\rho_k \underline{\varphi}$ seems to have all of the structure-preserving properties which we demanded for a numerosity measure. It is easy to show that the set of all maps ρ_k has the structure of a semigroup under composition. (This semigroup is isomorphic to the semigroup $(J^+, \cdot)$.)

If we consider the next question, that of the existence of numerosity measures whose values are not necessarily integers, we find that the situation is quite similar to that just discussed. If the value space is taken to be the set Q^+ of positive rationals, then the appropriate functions for composition with the cardinal measure function are the functions

$$\rho_k : x \to kx$$

for each positive rational k. In passing, we observe that the numerosity measure functions corresponding to measurement in dozens, or by the score, by the hundred, by the million, etc., belong to this category. If we go further, and permit values in R^+, we reach a similar conclusion, except that k may then take on any positive real value.

At this point it is convenient to give a formal definition of "numerosity function", a definition which is in keeping with definitions which we shall give later for length functions, area functions, and so on. We define a <u>numerosity function</u> (with integer values) for D, to be a function $\varphi : D \to J^+$ such that

(i) if S and T belong to T, with $S \sim T$, then $\varphi(S) = \varphi(T)$;

(ii) if S and T belong to D, and $S \cap T = \emptyset$, then $\varphi(S \cup T) = \varphi(S) + \varphi(T)$.

In other words, all we require is that φ should give the same value to equivalent sets, and that "disjoint unions" should map into sums. The modification needed to permit values in Q^+ and R^+ are obvious. We point out that if $\tilde{D}$ is a semigroup, then it follows that a numerosity function φ will induce an isomorphism $\underline{\varphi} : (\tilde{D}, \underline{\cup}) \to (J^+, +)$ and that, as you should verify, the definition implies that a numerosity function preserves order. (This is a result of the close connection between the order properties of $\tilde{D}$ and J^+ and the operations $\underline{\cup}$ and $+$; e.g., if $m, n \in J^+$, then $m < n$ if and only if there is a $p \in J^+$ such that $m + p = n$; and similarly for $\tilde{D}$.)

The operation $\underline{\cup}$ is a binary operation on $\tilde{D}$, and hence determines a function $\underline{\cup} : \tilde{D} \times \tilde{D} \to \tilde{D}$. The relationship of this to addition in J^+ is conveniently indicated by the following diagram, whose commutativity is equivalent to the definition of φ. (The mapping $\underline{\varphi} \times \underline{\varphi}$ is the cartesian product mapping defined at the end of Section 1-5.)

$$\begin{tikzcd}
\tilde{D} \times \tilde{D} \arrow[r, "\underline{\varphi} \times \underline{\varphi}"] \arrow[d, "\underline{U}"'] & J^+ \times J^+ \arrow[d, "\text{addition}"] \\
\tilde{D} \arrow[r, "\underline{\varphi}"'] & J^+
\end{tikzcd}$$

We have still to consider question (iii), which asked whether there was any place for the second operation in J^+ (multiplication) in questions of numerosity.

If we apply the procedure for forming the cartesian product to the cardinal measure function $\varphi: D \to J^+$, we get a natural mapping $\varphi \times \varphi : D \times D \to J^+ \times J^+$. As mentioned in Section 1-5, we can compose this with the "multiplication" mapping of $J^+ \times J^+$ into J^+. We also recall that the elements of D are themselves finite sets, and that the cartesian product of two finite sets is a finite set; i.e., D is closed under the cartesian product operation. This binary operation corresponds to a mapping from $D \times D$ to D. (Don't confuse the cartesian product of two elements (finite sets) of D with the cartesian product $D \times D$.) Let us put all of these mappings together in a single diagram:

$$\begin{tikzcd}
D \times D \arrow[r, "\varphi \times \varphi"] \arrow[d, "\begin{array}{l}\text{cartesian}\\\text{product of}\\\text{elements}\\\text{(finite}\\\text{sets)}\end{array}"'] & J^+ \times J^+ \arrow[d, "\text{multiplication}"] \\
D \arrow[r, "\varphi"'] & J^+
\end{tikzcd}$$

A corresponding diagram for a "typical" element of $D \times D$ is useful in keeping track of the various functions:

$$\begin{tikzcd}
(S,T) \arrow[r] \arrow[dd] & (\varphi(S), \varphi(T)) \arrow[d] \\
 & \varphi(S) \cdot \varphi(T) \\
S \times T \arrow[r] & \varphi(S \times T)
\end{tikzcd}$$

It is a perfectly reasonable question to ask: "Does $\varphi(S) \cdot \varphi(T) = \varphi(S \times T)$?" In other words, "is the first diagram a commutative diagram?" That the answer is "yes" follows from our definition of the "counting function". φ, from the definition of cartesian product, and from properties of J^+ and of multiplication in J^+, which we stated in Section 2-2.

We can give an even clearer diagram which represents this situation, if we assume that S and T each contain a reasonably small number of elements. For example, if $\varphi(S) = 5$, and $\varphi(T) = 3$, the last commutative diagram can be replaced by

$$\begin{tikzcd}
(\{a_1,a_2,a_3,a_4,a_5\}, \{b_1,b_2,b_3\}) \arrow[r, "\varphi \times \varphi"] \arrow[d, "\begin{matrix}\text{cartesian}\\ \text{product}\\ \text{operation}\end{matrix}"'] & (5,3) \arrow[d, "\begin{matrix}\text{multiplication}\\ \text{operation}\end{matrix}"] \\
\left\{\begin{matrix} (a_1,b_1), (a_1,b_2), (a_1,b_3) \\ (a_2,b_1), (a_2,b_2), (a_2,b_3) \\ (a_3,b_1), (a_3,b_2), (a_3,b_3) \\ (a_4,b_1), (a_4,b_2), (a_4,b_3) \\ (a_5,b_1), (a_5,b_2), (a_5,b_3) \end{matrix}\right\} \arrow[r, "\varphi"'] & 15 = 5 \cdot 3
\end{tikzcd}$$

What about the other numerosity functions, $\rho_k \varphi$, where $\rho_k : x \to kx$, $k \neq 1$? It is trivial to check that, whether we use J^+, Q^+, or R^+ as the image space, $(\rho_k \varphi)(S \times T) = k[\varphi(S) \cdot \varphi(T)]$, while $[(\rho_k \varphi)(S)][(\rho_k \varphi)(T)] = k^2[\varphi(S) \cdot \varphi(T)]$. Hence $(\rho_k \varphi)$ does not carry cartesian products in D into products in the image space, unless $k = 1$. In other words, the cardinal number numerosity measure function is distinguished from all other numerosity functions by this property of mapping cartesian products of elements of D into corresponding number products in the image space.

The situation pictured in the diagrams above contains the germ of an idea that will recur over and over again in discussing the relationship of measure functions. The fact that we had a second operation (cartesian product) mapping $D \times D$ into D is quite special to the measurement of numerosity. In other situations (e.g., angular measures, area, velocity), we will have measure functions μ_1, μ_2, with domains D_1, D_2, (not necessarily different), a relation between D_1 and D_2 (i.e., a subset of $D_1 \times D_2$), and a mapping from the domain D_3 of another measure function, μ_3, onto this relation. Commutativity will then hold in a diagram

$$\begin{tikzcd}
D_1 \times D_2 \arrow[r, "\mu_1 \times \mu_2"] & R^+ \times R^+ \arrow[d, "\nu"] \\
D_3 \arrow[u] \arrow[r, "\mu_3"'] & R^+
\end{tikzcd}$$

When μ_3 is angle measurement, μ_1 and μ_2 are length measures, ν is division, and the commutative diagram expresses the relationship between length and angle measures which is involved in the idea of radian measure. If μ_3 is an area measure function, μ_1 and μ_2 are length measures, and ν is multiplication. When μ_3 is "average velocity", μ_1 is a length function, μ_2 is a time-interval function, and ν is division. We shall have more to say later about these ideas.

By now you might have reached the conclusion that we are spending a great deal of time stating the obvious. In a sense this is true, because the formal system J^+, the operations of addition and multiplication, and the inequality relation, were all constructed precisely for the purpose of providing a "model" for the empirical ideas of numerosity measurement. To some extent this will be true again in the next section, in which we discuss the empirical measurement of length, but we feel that it is worthwhile going through the development, in an attempt to clarify what is empirical, and what is the formal (mathematical) model, and what are the relations between the two. Mathematical systems devised for modeling the real world have a habit of developing a life of their own, and this does not necessarily have any empirical counterpart.

Exercises 2-3

1. Prove the assertion made above, that the set of functions which map the ordered semigroup $(J^+,+,<)$ isomorphically into itself, is an ordered abelian semigroup (with composition as the semigroup operation, and an appropriate inequality relation). Prove also that this ordered semigroup of transformations is isomorphic to the ordered semigroup $(J^+,\cdot,<)$.
2. What is the structure of the group of automorphisms of $(J^+,+,<)$?
3. Similar to previous exercise, but with Q^+ instead of J^+.

We conclude this section with some remarks on a number of matters which arise in connection with most measure functions:

Domain. From a mathematician's viewpoint, a function is not properly described unless its domain is specified. In most measurement situations one is not only concerned with the domain, but also with efforts to extend the domain in some meaningful way. In the case of empirical measures, this usually involves important practical questions of procedure -- accuracy, improved instrumentation and so on -- and in many cases it gives rise to philosophical problems. In fact, when analyzed carefully, many of the philosophical arguments concerning measurement turn out to involve questions of domain.

In most measurement situations, the initial methods used for defining a suitable measure function have to be modified when the domain is extended. For example, in cardinal number measurement (by counting), the pairing-off process "by hand" is not appropriate for counting large numbers of physical particles, such as are involved in radioactive decay, and complicated electronic devices are used. In other situations we resort to mathematical methods in order to "count without counting". (See [12].) As a general rule, the domain of a useful (empirical) measure function is extended to the point where the process of measurement involves a complex mixture of empirical evidence, induction, and mathematical theory.

Accuracy. As mentioned above, the subjects of accuracy and domain are inter-related. We shall not concern ourselves very much with the practical side of this important question, but, even in connection with numerosity measurement, we point out that our implied assumption of absolute accuracy (in counting) is not generally valid unless we restrict the domain severely.

Accuracy involves both physiological and psychological considerations, as well as instrumentation. On the physiological side, we might instance the limited ability to discriminate (visually or otherwise) between the elements of a set in the domain. (The elements of the domain D are finite sets.) We recall that, in mathematics, we do not accept that a set is well-defined unless there is a precise method for determining whether or not a given object belongs to the set, and we assume, in effect, perfect discrimination. In physical situations this is often impossible to achieve, even when aided by refined instruments. On the psychological side, we have all the problems of human error, memory and so on.

From a mathematical point of view, this aspect of measurement (accuracy) raises many interesting statistical questions, including those involved in the process of "rounding off", but consideration of these is not part of this book.

<u>Units</u>. This topic will assume greater importance when we consider a measure function, such as length, whose image space is the reals (or the positive reals). For the present we merely remark that, for a given measure function, the corresponding unit is the set (i.e., equivalence class) of those elements of the domain whose image is the number 1. (Sometimes a particular object in this set is referred to as the unit, much as we refer to a fraction as a rational number.) Where there is more than one function which is suitable for the measurement of a given attribute, the choice of function and the choice of unit are often equivalent. Once a function is chosen, the unit is fixed; on the other hand, if a unit is specified, we shall see that the function is uniquely determined. There are important questions concerning so-called "change of units"; we discuss these in later sections.

For cardinal number measure, the unit is the equivalence class of those sets which each contain a single element. For numerosity measurement in dozens, the unit ("dozen") is the equivalence class of sets with 12 elements; and so on. It is sometimes convenient to use the same name for the unit and for the corresponding function.

<u>Language</u>. Much of the language which we have used above in connection with numerosity measurement, is not the language which we normally use in our daily lives. It is useful, therefore, to look at a few equivalent statements using both forms of expression. We will come back to this question of language again, after the introduction of empirical length measures.

	Common Usage	Usage As In This Section
(i)	There are 28 students in this class.	The cardinal number measure of the set of students in this class is 28.
(ii)	This box contains $12\frac{1}{3}$ dozen oranges.	The numerosity in dozens of the set of oranges in this box is $12\frac{1}{3}$.
(iii)	Company X made a profit of \$14.7 million in 1964.	The numerosity in millions of the profit in dollars made by Company X in 1964 is 14.7 .

In these examples the everyday language is clearly more compact. Observe also that there are various conventions regarding units and objects. In example (i), the unit is unspecified, but it is clearly understood; the set of objects (the students in this class) is identified after the value is given. In example (ii), the unit (dozen) is named after the number; and the words describing the set of objects ("This box - - - oranges") are separated. In example (iii), the unit is "million", the "objects" are dollars, and the words describing the set of objects are split up. ("Company X - - - profit - - - 1964").

Of course the functional description becomes much more concise if we introduce appropriate symbolism. E.g., if φ, φ_{12}, φ_m, denote the measure functions, and S_1, S_2, S_3 the elements of the respective domains, for examples (i), (ii), (iii) above, we may write

(i) $\varphi(S_1) = 28$;

(ii) $\varphi_{12}(S_2) = 12\frac{1}{3}$;

(iii) $\varphi_m(S_3) = 14.7$.

2-4 The Physical Measurement of Length

The word "physical" in the title of this section is used by way of contrast with the word "mathematical" in the title of the next section, in which we discuss the question of length measurement in formal mathematical systems. It does not indicate that we shall be concerned with physical equipment, or with most of the practical problems that attend the physical process of length measurement in varied situations.

As in the preceding section, we are concerned with an undefined attribute (in this case, length) of certain "objects". These objects may be such entities as an edge of our desk, rigid "rods", pairs of fixed "points" which represent space locations, and so on. We may think of these objects, intuitively, as the physical counterparts of pairs of points (or, equivalently, the line segments which these pairs of points determine) in a geometric space, but this is in no sense a definition. In order to have a simple "neutral" word for the elements of our domain, let us call these elements rods. Thus a rod is an object which we wish to include in the domain of any length-measurement function.

Our strategy will be similar to that used in the last section. Corresponding to our intuitive idea of length, we establish procedures for comparing lengths, and these lead to a certain structure on the domain. (Unless we restrict the domain severely, many of the structural properties of the domain will be only inductive hypotheses.) We then ask ourselves whether there are any functions on our domain, with values in a suitable number system (e.g., the positive reals R^+) , which preserve the structure of the domain; and should there be more than one, how these different length-measurement functions are related.

Let us assume that we have decided on a physical procedure for comparing "lengths" (i.e., comparing rods with respect to this intuitively felt attribute, length). We may think of this as placing two rods side-by-side, with one end of each matched, and then deciding (visually, by touch, or some other way) whether or not the other ends match. If two rods, d_1 , d_2 (from the domain D) , match exactly, we say that they are "equivalent with respect to length" (or have the same length). We denote this relation by $d_1 \sim d_2$ and investigate its properties. We assume that our comparison procedure makes this relation symmetric. Moreover our intuitive idea of length suggests that we should assume that the relation is reflexive. Transitivity is a more interesting matter. In practice, if we take three rods d_1 , d_2 , d_3 such that $d_1 \sim d_2$, and $d_2 \sim d_3$, then it is likely that we shall find that $d_1 \sim d_3$. But if we were to take a considerable number (say 1000) rods, with the empirical properties

$$d_1 \sim d_2 , d_2 \sim d_3 , \dots , d_{999} \sim d_{1000}$$

it is most unlikely (whatever physical procedure we have adopted) that d_1 will appear equivalent to d_{1000}. (This is well known to carpenters: a carpenter who wants to cut a large number of "copies" of a length of timber, does not follow a procedure of cutting copy 2 equivalent to copy 1 , copy 3 equivalent to copy 2 , and so on!)

This immediately raises the question of accuracy, and what we meant above by the expression "match exactly". The actual matching procedure is likely to involve us in the use of vision, or touch, and hence might involve properties of light, and physiological and psychological properties of the observer. It might also involve us in moving rods about, and in making comparisons at different times and places, under different conditions of temperature and atmospheric pressure, in different gravitational and magnetic fields, and in various states of motion. We know, of course, that some of these things are important in relation to what we are considering (length), and we must decide what to do about them. Problems of this sort are discussed in some detail in Chapter 1 of [2] . The conclusion that is forced upon us is that no physical procedure for length comparison can be exact, in the sense that it will lead, with no ambiguity, to the structure discussed below. In fact, if we probe a little more deeply, we will see that the impossibility of exactness is not just a property of our procedure, but, since there are no physical counterparts of the points and line segments of geometry, there is no physical way of even specifying "exactly" the elements of our domain.

In order to make progress in our attempt to establish a structure in our domain, we shall largely ignore the practically-important question of precision. In particular, we assume, for the moment, that our comparison procedure for the rods in our domain is exact, in the sense that it enables us to make a definite decision with respect to each pair of objects; and that the resulting relation on the domain is an equivalence relation. We denote the equivalence class of an element d , by $\tilde{d}$, and we denote the set of all equivalence classes, by $\tilde{D}$.

The comparison procedure used for determining length-equivalence, leads (with similar assumptions of exactness) to a relation of "less than with respect to length" ($<$) between rods. We assume that our empirical evidence justifies the assumption that this is a transitive relation on D, and that it yields a corresponding order relation ($<$) on $\tilde{D}$.

We now examine our domain in relation to the density of the order relation; i.e., given $d_1, d_2 \in D$, with $d_1 < d_2$; we search for a d_3 such that $d_1 < d_3 < d_2$. Let us assume that we can always find such a d_3, so that our order relation is dense. (Notice that this is really a very big assumption: not only does it involve the question of the absolute accuracy of the comparison procedure, but it implies that, if our domain has at least two length-different objects, then it has infinitely many.) It follows that the order relation in $\tilde{D}$ is also dense.

The next step is to introduce an operation in D (which, hopefully, will yield a derived operation in $\tilde{D}$), which we will expect to map into addition under an appropriate measure function. We establish this operation by a physical procedure of "joining", which we can think of as placing two rods in line, "end-to-end". We assume that this composite entity is an object of our domain D. Denote the join operation by $*$. Then for $d_1, d_2 \in D$, $d_1 * d_2 \in D$; i.e., D is closed under the join operation. The empirical procedure suggests that we should assume that this operation is associative, and that it has the property that $d_1 \sim d_1', d_2 \sim d_2' \Longrightarrow d_1 * d_2 \sim d_1' * d_2'$. In other words, the join operation leads to a corresponding associative operation (for which we use the same symbol) on equivalence classes. We examine next the commutativity of the join operation. Clearly we would not regard $d_1 * d_2$ and $d_2 * d_1$ as the same rod, so we do not find that $*$ is commutative on D. If you think of the possible physical processes involved, you will see that there is a minor difficulty in examining commutativity in D: for $d_1, d_2 \in D$, it is unlikely that we can directly compare $d_1 * d_2$ with $d_2 * d_1$. This seems to violate our assumptions that (a) both of these elements belong to D; and (b) we have a length comparison procedure for each pair of elements in D. This difficulty is not serious: we can overcome it by assuming that we can always find a third element $d \sim (d_1 * d_2)$, to compare with $d_2 * d_1$, or by assuming that we have more than one "copy" of each rod. Either way, let us assume that our empirical evidence is consistent with the assumption that $*$ is commutative on $\tilde{D}$.

We look next at the relationship of the order relation, and the join operation in $\tilde{D}$, and assume that empirical evidence justifies the assumption that, for $\tilde{a}, \tilde{b}, \tilde{c} \in \tilde{D}$,

(i) $\tilde{a} < \tilde{b}$ if and only if there exists $\tilde{c}$ with $\tilde{a} * \tilde{c} = \tilde{b}$;

(ii) $\tilde{a} < \tilde{b} \Longrightarrow \tilde{a} * \tilde{c} < \tilde{b} * \tilde{c}$ for all $\tilde{c} \in \tilde{D}$.

In other words, we feel justified, as a result of our investigation, in assuming that the system $(\tilde{D}, *, <)$ which we have obtained by physical operations, induction, and abstraction, is a densely-ordered abelian semigroup.

At this stage we might stop our empirical investigation, and consider our objective: to establish a length measure function on D . Before doing this we have to say a little more clearly what we require of such a function. We define a <u>length function</u> to be a function $\lambda : D \to R^+$ which has the properties

(i) $\lambda(d_1) = \lambda(d_2)$ if and only if $d_1 \sim d_2$;

(ii) if $d_3 = d_1 * d_2$, then $\lambda(d_3) = \lambda(d_1) + \lambda(d_2)$.

The first requirement ensures that λ will induce a corresponding length function (for which we use the same symbol, λ) on $\tilde{D}$. The second "additivity" condition implies that λ must be "finitely additive". And the earlier assumptions which we have made ensure that λ must preserve order, both as a mapping on D and as a mapping on $\tilde{D}$. The systems $(\tilde{D}, *, <)$ and $(R^+, +, <)$ are both ordered, abelian semigroups, and our definition of length function implies that $\lambda : (\tilde{D}, *, <) \to (R^+, +, <)$ must be an isomorphism, but not necessarily onto.

<u>Value Space and Range for Length Functions</u>. We recall that there are subsets of the positive reals which have the structure of an ordered abelian semigroup; in particular, the positive integers, J^+ , the positive rationals, Q^+ , and the positive reals, R^+ are all ordered abelian semigroups under addition. We therefore ask ourselves the following questions:

1. Is there any structure-preserving 1-1 mapping (isomorphism) of $\tilde{D}$ into J^+ , Q^+ , or R^+ ?

2. If there is such a mapping, is there more than one? And if so, how are they related?

(We can quickly dispose of question 1 as far as J^+ is concerned. The order relation in $\tilde{D}$ is dense, while that in J^+ is not. It follows that there cannot be a 1-1 order-preserving map from $\tilde{D}$ to J^+. (Intuitively, the reason for this is that, for any two elements of $\tilde{D}$, there are infinitely many elements between them; but their images in J^+ would be positive integers, and there are only finitely many integers between each pair of integers.)

We recall that both Q^+ and R^+ have dense order relations, so neither is ruled out as a possible image space on account of the denseness of the order in $\tilde{D}$. But we shall have to do quite a bit more work before we can decide whether or not there is any structure-preserving map from $\tilde{D}$ to either Q^+ or R^+. We set aside this question for the time being, merely noting that we will, in fact, be able to show the existence of such a function as far as R^+ is concerned, but that we shall not (because of questions of accuracy, both of procedure and definition) be able to decide empirically whether R^+ is really needed, or whether Q^+, or some other subset of R^+, will suffice. When we come to look at the question of length-measurement in mathematical systems, we shall see that Q^+ will not suffice for the measurement of length in the euclidean plane; but we shall also see that, if we start from the axioms of classical synthetic geometry, then we cannot prove that R^+ is needed. In fact, there are geometries which satisfy the classical axioms, and which do not require R^+ for length measurement: but more about that in the next section.

Relationships Between Length Functions. We turn now to our second question: If there is a length function, λ, from $(\tilde{D},*,<)$ to $(R^+,+,<)$, are there any others? And if there are, how are they related? The first question is easily answered: we have seen (Theorem 2-2.2) that any positive similarity transformation $\bar{k}$, on R^+ is an automorphism of the ordered semigroup $(R^+,+,<)$, and hence (as you may easily show) the composite function, $\bar{k}\lambda$, is also an automorphism of $(\tilde{D},*,<)$ into $(R^+,+,<)$. That is, $\bar{k}\lambda$ is also a suitable length function. If we denote $\bar{k}\lambda$ by λ_1, then we see immediately that $\lambda = \overline{\frac{1}{k}}\lambda_1$; i.e., λ is the composite of λ_1 with a positive similarity of R^+. In other words, there is a symmetry between λ and λ_1, in the sense that each may be obtained from the other by composition with a suitable positive similarity, and the two similarities are inverse elements in the positive similarity group. (As we shall see in the next section, if λ is not onto, then some of these composite functions might exhibit such unexpected behavior as failing to possess a unit, but we leave discussion of this interesting question until later.)

It is now natural to ask whether all suitable length functions (if there are any) are related in this way, by positive similarities. We cannot yet answer this question absolutely, but we can give a conditional answer: if there exists at least one such function, λ, which is onto, (i.e., λ is an isomorphism of $\tilde{D}$ onto R^+: note that this implies, in particular, that λ takes on arbitrarily large values -- which is actually implied by our assumption that $\tilde{D}$ is closed under the join operation -- and arbitrarily small values) and if λ_1 is another suitable function (not assumed onto) then it is easy to show that

$$\lambda_1\lambda^{-1} : R^+ \to R^+$$

is an isomorphism of $(R^+,+,<)$ into itself. But it follows from Theorem 2-2.2, that the only such isomorphisms are the positive similarity automorphisms. Hence λ_1 (and hence every other admissible length function) is also onto, and any two such functions differ by a positive similarity: i.e., they are similar functions. In other words, there is nothing which obviously distinguishes any one function from the other; and, in the sense of Section 1-7, "length" is a positive-similarity-invariant measure, or a ratio scale.

We now look at the length functions themselves, with the assumption that they are onto, and hence similar. Denote the set of all such functions by Λ. Then, as in Section 1-5, we can define an operation of addition $(+)$ on functions from $\tilde{D}$ to R^+, and we find that if $\lambda_1, \lambda_2 \in \Lambda$, then $\lambda_1 + \lambda_2 \in \Lambda$ (i.e., Λ is closed under addition). Again, as in Section 1-5, Λ has an order relation $(<)$ and $(\Lambda, +, <)$ is an ordered abelian semigroup. The order relation in Λ is dense. Moreover it follows from the assumption that each λ is onto R^+, that if we take any $\lambda_0 \in \Lambda$ then for each $\lambda \in \Lambda$, there exists k such that $\lambda = k\lambda_0$; and for each $r_0 \in R^+$, the function $f : \Lambda \to R^+$ defined by

$$\begin{cases} f : \lambda_0 \to r_0 \\ f : \lambda \to kr_0 \end{cases}$$

can be shown to be an isomorphism of $(\Lambda, +, <)$ and $(R^+,+,<)$. Notice that, since the choices of λ_0 and r_0 were arbitrary, this isomorphism can be set up in infinitely many ways.

There is also an operation of multiplication in the set of all functions from $\tilde{D}$ to R^+, but it is easy to verify that the subset Λ of length functions is not closed under this multiplication: the product of two length functions is not a length function.

We return now to the question which we left unresolved: whether or not there are any structure preserving functions from $\tilde{D}$ to R^+; i.e., whether the set Λ is empty or not. You are probably familiar with procedures for setting up suitable length functions: we shall describe two of them. The first is the one usually used in elementary work, and the second is derived from the procedure used by Eudoxus to develop a theory of ratios for segments.

Method I: Length Function in Terms of Selected Unit and Sub-units.

We first select, quite arbitrarily, any rod d_0 in D as a "unit". (Strictly speaking, the unit is the equivalence class of d_0.) Then for any other rod d; we compare d with successive "multiples" of d_0 until we reach a multiple nd_0 (i.e., $d_0 * d_0 * \ldots * d_0$; n terms; clearly $1d_0 = d_0$) such that $nd_0 \lesssim d < (n + 1)d_0$, where the symbol $\lesssim$ stands for "less than or equivalent to". (This requires that D be archimedean, a property which we assume to have empirical justification. It also requires that we have an unlimited number of rods which are equivalent (in length) to d_0. It is convenient to simplify notation by using the same notation d_0; for each of these, in the notation nd_0: strictly speaking we should move to equivalence classes at this point.) Assume next that we are provided with 10 rods d_1, such that $10d_1 \sim d_0$. Adjoin these to the composite rod nd_0 (laid off along d) and obtain a positive integer n_1 $(0 \leq n_1 \leq 9)$ such that $nd_0 * n_1d_1 \lesssim d < nd_0 * (n_1 + 1)d_1$. Assume next, that we are provided with 10 rods d_2, such that $10d_2 \sim d_1$, and continue the procedure to obtain n_2 $(0 \leq n_2 \leq 9)$ such that $nd_0 * n_1d_1 * n_2d_2 \lesssim d < nd_0 * n_1d_1 * (n_2 + 1)d_2$. In this way we can build up the decimal number $r = n.n_1n_2 \ldots$ and we define a function

$$\lambda_0 : \tilde{D} \to R^+ \text{, by } \lambda_0(\tilde{d}) = r .$$

(Clearly, we did not need to use submultiplies with 10 equivalent parts; we could, for example, have used 2 equivalent parts to obtain the non-integer part of r in binary form.)

Remark. Two things about this process are worth pointing out:

1. In the process, we compared the rod d with the "iterated join" of d_0 with identical copies of itself. In order to obtain the number n, and the digits $n_1, n_2, \ldots$, we needed to be able to count. That is, approached in this way, the measurement of length depends on the simpler ability to measure numerosity.

2. In order to build up the decimal number $r = n.n_1n_2 \ldots$, we needed to be able to "divide" the rod d_0 into 10 equivalent parts d_1; d_1 into 10 equivalent parts d_2; and so on. As we have no assurance that the process of matching will ever terminate (i.e., lead to a finite decimal expansion) this implies that, if the process is to be carried out as outlined, then we must be able to obtain "arbitrarily small" rods. That is, for each positive integer m we need to use rods d_m such that $10^m d_m \sim d_0$. Our knowledge of physics suggests that such rods cannot be obtained. (I.e., that matter, unlike the real line, is not "inifinitely divisible".)

We assume that it is a property of the operations used for determining equivalence, and for determining the function λ_0, that λ_0 induces a corresponding function (for which we use the same symbol) on $\tilde{D}$. We now examine this function, to see whether or not it has the properties we demand of a length function. If (as a result of empirical evidence) we were to make a number of additional assumptions concerning, for example, the "algebra" of the "multiplication" nd_0, then we could make some attempt at proving that the λ_0 empirically defined, is an isomorphism of $(\tilde{D}, *, <)$ into $(R^+, +, <)$. We shall have more to say in the next section on the corresponding question which arises in connection with length measurement in geometry; but as far as empirical length-measurement is concerned, whether or not λ_0 is an isomorphism is a question which can only be answered on an empirical basis, taking into account not only the experimental evidence, but also the consequences of the assumptions (hypotheses) suggested by the evidence.

Concerning the question of "ontoness": we observe that, unless r has a terminating decimal (or binary, if that is what we have used) expansion, even the assumption of exact matching will not allow us to actually find r, because of the infinite character of the process. Hence, even if r is rational, an assumed exact procedure would not necessarily disclose this. Moreover, if we admit even the smallest amount of inexactness (say of the order of 10^{-100} cm) then, since every real number interval contains infinitely many rationals and infinitely many irrationals, there is no possibility of seriously considering the question of whether for some $d \in \tilde{D}$, $\lambda_0(d)$ must be irrational. In other words, there is no way of deciding empirically whether or not we "need" real (or at least more than rational) numbers for length measurement, and it is meaningless even to ask the question, unless we can find some way to say precisely what we mean by "need".

The procedure outlined above for establishing the existence of a length function, required an arbitrary choice of "unit". We have defined the unit of a length function to be the (unique if it exists) element of $\tilde{D}$ whose image is the number "1". Thus if we assume that each length function is onto R^+, then there is a 1-1 correspondence of units and functions as described above, and the set of all possible units is clearly $\tilde{D}$ itself. But if each empirically-obtained function is not onto R^+, then some length functions will exist which do not have units, and there will not be a 1-1 correspondence of units with the set of all length functions.

Exercises 2-4

1. Assume that there exists at least one length function λ_0, and prove the assertion just made: that for any $\tilde{d} \in \tilde{D}$, there exists a length function $\lambda_{\tilde{d}}$, such that $\lambda_{\tilde{d}} : \tilde{d} \to 1$.

2. Assume that each length function is onto R^+, and prove that the 1-1 correspondence

$$\tilde{d} \longleftrightarrow \lambda_{\tilde{d}}$$

of $\tilde{D}$ and Λ, is order reversing.

3. A is a set with at least two elements, and with a dense order relation. Prove that there cannot be a 1-1 order preserving map from A to J^+.

4. Prove that for $\bar{k} \in \underline{S}^+$, the function $\bar{\bar{k}} : \Lambda \to \Lambda$ defined by $\bar{\bar{k}}(\lambda) = (\bar{k}\lambda)$ is an automorphism of $(\Lambda, +, <)$; show also that if each λ is assumed onto R^+, then there are no other automorphisms, and $\underline{S}^+$ is isomorphic to the automorphism group of $(\Lambda, +, <)$, under the 1-1 correspondence $\bar{k} \longleftrightarrow \bar{\bar{k}}$.

If each length function is assumed to be onto R^+, then it follows trivially that the number "1" is contained in the range of each length function; i.e., each length function has a unit. Conversely, if we assume that each length function has a unit, then it is not hard to show that each length function is onto R^+. For if r is any positive real number, and if λ is any length function, then $\frac{1}{r}\lambda$ is also a length function; and if $\tilde{d}_1$ is the assumed unit of $\frac{1}{r}\lambda$, we have $(\frac{1}{r}\lambda)(\tilde{d}_1) = 1$, so that $\lambda(\tilde{d}_1) = r$. Hence every positive real number is in the range of λ, and λ is onto R^+. We record this simple but important result as a theorem.

Theorem 2-4.1 Each length function is onto R^+ if and only if every length function has a unit.

Ratios and Ratio Operations: In the following discussion we assume that each length function is onto R^+, so that each length function has a unit. If you worked the second exercise above, you probably discovered that if $\tilde{d}_1$, $\tilde{d}_2 \in \tilde{D}$, and if λ_1, λ_2 are length functions for which $\tilde{d}_1$, $\tilde{d}_2$ are units (i.e., $\lambda_1(\tilde{d}_1) = \lambda_2(\tilde{d}_2) = 1$) then $\lambda_1(\tilde{d}_2) = \frac{1}{(\lambda_2(\tilde{d}_1))}$. (This follows simply from the fact that if $\lambda_1(\tilde{d}_2) = k$, then $\lambda_1 = k\lambda_2$, and hence $\lambda_2 = \frac{1}{k}\lambda_1$.) This suggests that not only is the 1-1 correspondence of functions and units order-reversing, but that, in some sense, there is a reciprocal relationship involved; i.e., that there should be some sort of "ratio", of units, $\tilde{d}_1 : \tilde{d}_2$, and a "ratio" of functions, $\lambda_1 : \lambda_2$, such that

$$\tilde{d}_1 : \tilde{d}_2 = \lambda_2 : \lambda_1 .$$

We could consider that the described process for assigning a "length" to a rod d, using the rod d_0 as unit, established a "ratio" $d : d_0$ (i.e., that $d : d_0 = \lambda_0(d)$) and that, because of the arbitrariness of the choice of d_0, the measurement procedure actually gave a means of determining a ratio (i.e., a real number) $d_1 : d_2$, for each ordered pair of rods (d_1, d_2). Let us assume (on the basis of empirical evidence) that this ratio is the same for equivalent pairs of rods, so that we obtain an (empirical) function

$$\rho : \tilde{D} \times \tilde{D} \to R^+$$

defined by

$$\rho(\tilde{d}_1 , \tilde{d}_2) = d_1 : d_2 .$$

If we continue to assume that each length function is onto R^+, we can show the well-known relationship between these ratios and the similarity factors which relate the corresponding length functions.

Theorem 2-4.2. If each length function is assumed to be onto R^+, and if λ_1, λ_2 are the length functions whose units are $\tilde{d}_1$ and $\tilde{d}_2$, respectively, then. $\lambda_2 = (d_1 : d_2)\lambda_1$.

Proof. From the assumptions made, each two length functions are similar. Hence there exists $k \in R^+$, such that $\lambda_2 = k\lambda_1$. Hence $\lambda_2(\tilde{d}_1) = k\lambda_1(\tilde{d}_1) = k$. But (by the definition of $d_1 : d_2$), $\lambda_2(d_1) = d_1 : d_2$. Hence $\lambda_2 = (d_1 : d_2)\lambda_1$, as required.

Remarks:

1. If $\tilde{d}_1$ = "inch", and $\tilde{d}_2$ = "foot", this is simply the well-known relationship

$$\lambda_{\text{foot}} = \frac{1}{12} \lambda_{\text{inch}}$$

Or, in words, "length in feet equals one twelfth length in inches".

2. If $\lambda_2 = k\lambda_1$, it is reasonable to define the <u>ratio of</u> λ_2 <u>to</u> λ_1 to be the real number k . We write this in the usual way, as $\lambda_2 : \lambda_1$. It then follows from the theorem that $\lambda_2 : \lambda_1 = \tilde{d}_1 : \tilde{d}_2$.

3. It is easy to see that, since all length functions are similar, the ratios of length functions satisfy the relationship:

$$(\lambda_3 : \lambda_2)(\lambda_2 : \lambda_1) = (\lambda_3 : \lambda_1) .$$

Hence, if $\tilde{d}_1$, $\tilde{d}_2$, $\tilde{d}_3$, are the corresponding units, we have

$$(\tilde{d}_2 : \tilde{d}_3)(\tilde{d}_1 : \tilde{d}_2) = (\tilde{d}_1 : \tilde{d}_3) ,$$

which we rewrite in the more usual order: $(\tilde{d}_1 : \tilde{d}_2)(\tilde{d}_2 : \tilde{d}_3) = (\tilde{d}_1 : \tilde{d}_3)$.

If we use the common notation $\frac{\tilde{d}_1}{\tilde{d}_2}$ for $\tilde{d}_1 : \tilde{d}_2$, this property takes the highly suggestive form

$$\left(\frac{\tilde{d}_1}{\tilde{d}_2}\right)\left(\frac{\tilde{d}_2}{\tilde{d}_3}\right) = \frac{\tilde{d}_1}{\tilde{d}_3} ,$$

but we must not assume that this result holds because of numerical "cancellation": the units $\tilde{d}$ are not numbers, even though their ratios are.

In general, whenever we have a set A, and a binary operation on A, with values in R^+ (i.e., a function $\beta : A \times A \to R^+$) which, for all a_1, a_2, a_3 in A, has the "cancellation" property that, $\beta(a_1,a_2)\beta(a_2,a_3) = \beta(a_1,a_3)$, then β is called a ratiofunction, or a ratio operation on A, with values in R^+. It follows that, if we assume that the described procedure for setting up length functions on D leads to length functions which are onto R^+, then the corresponding function

$$\rho : \tilde{D} \times \tilde{D} \to R^+ ,$$

defined by

$$\rho(\tilde{d}_1,\tilde{d}_2) = \tilde{d}_1 : \tilde{d}_2 = \lambda_2(d_1) ,$$

is a ratio operation. We shall see later that there is a natural correspondence of ratio operations and ratio scales.

In many treatments of length measurement it is assumed that, in addition to an equivalence relation, a related order relation, and a related "join", or "addition" operation, there is a related ratio operation, and an empirical process for measuring the values of this ratio operation. Among the properties which are generally assumed (inductively) for this ratio operation

$$\rho : D \times D \to R^+$$

are

(i) If $d_1 \sim d_1'$ and $d_2 \sim d_2'$, then $\rho(d_1,d_2) = \rho(d_1',d_2')$;

(ii) If $d_1 < d_2$, then, for every d_0, $\rho(d_1,d_0) < \rho(d_2,d_0)$;

(iii) $\rho(d_1,d_2)\rho(d_2,d_3) = \rho(d_1,d_3)$;

(iv) $\rho(d_1 * d_2 , d_0) = \rho(d_1,d_0) + \rho(d_2,d_0)$.

In general, a ratio operation which has a property like (iv), with respect to a commutative, associate, binary domain operation like the "join", is called an additive ratio operation. We shall see later that every ratio operation determines an equivalence relation on its domain, and an operation of "addition" on equivalence classes, such that the given ratio operation is additive with respect to the defined "addition". However, in measurement situations there is usually a prior operation (like the join) and we seek a ratio operation which is additive with respect to this.

If the existence of such a ratio operation for length is assumed, then it is easy to use the ratio operation to establish length functions. We shall see how this is done in the discussion below of an alternate (theoretical) procedure for establishing empirical length functions: a procedure which corresponds closely to the classical device used by Eudoxus to overcome the difficulties resulting from an inadequate number system and the consequent absence of a satisfactory theory of length.

The concept of "ratio" is fundamental in a large part of physical measurement, and most of us have a strongly developed intuitive feeling for its basic property, the cancellation property, which we usually use so naturally that we are scarcely conscious that we are making use of it. (E.g., from "A is twice as long as B, and B is three times as long as C", we draw the conclusion "A is six times as long as C" without any thought of the assumed (empirical) properties of ratios (or, equivalently, of length functions) from which the conclusion may be drawn.)

You may have observed that we have defined the term "ratio operation", but we have not yet defined "ratio". As far as we are concerned, the two ideas go together: whenever we have a set A and a ratio operation $\beta : A \times A \to R^+$, then $\beta(a_1,a_2)$ is called the "ratio of a_1 to a_2 (with respect to β)". If β is the "length ratio" function, we might express this as "the ratio of the length of a_1 to the length of a_2". A ratio $\beta(a_1,a_2)$ may be written symbolically as $a_1 : a_2$, or by $\frac{a_1}{a_2}$; but, except in the special case that a_1 and a_2 are themselves numbers, this is not ordinary division: "ratio" is just a binary operation β from a set A to the positive reals, which satisfies the cancellation property: $\beta(a_1,a_2)\beta(a_2,a_3) = \beta(a_1,a_3)$. This operation, like number division, is non-associative and non-commutative. You will recall that number division is usually regarded as a secondary operation, defined in the well-known way in terms of multiplication. (I.e., for $a \neq 0$, we define $\frac{b}{a}$ to be the unique

c such that $ac = b$.) ~~We shall see later that~~ there is a "multiplication" which is similarly related to every ratio operation: a so-called "scalar multiplication" by positive real numbers; but there is not generally any suggestion that this "scalar multiplication" is a more fundamental concept than "ratio", in connection with measurement questions.

We shall explore the relationship of ratio operations, ratio scales, and scalar multiplication more fully in the next section.

Method II: Length Function in Terms of Ratios and a Selected Unit.

Let us imagine that we are back at the point where we have established the length-structure of $(\tilde{D}, *, <)$ as a densely-ordered abelian semigroup, and that we are interested in comparing the ratios (a still undefined term in this paragraph) of ordered pairs of rods. In other words, for d_1, d_2, d_3, $d_4 \in D$, we want to establish a criterion for determining a relation, α, on $D \times D$, which extends our intuitive idea that when d_1 is an integral "multiple" of d_2, and d_3 is the same integral "multiple" of d_4, then

$$(d_1,d_2)\ \alpha\ (d_3,d_4).$$

(The word "multiple" above, is used as before, in the sense of repeated joining: i.e., for n a positive integer, $nd = d * d \ldots * d$ (n terms).)

Given rods d_1, d_2, d_3 and d_4, we assume that the corresponding equivalence classes each have an unlimited number of elements, so that we can compare arbitrary (integral) multiples of any rod, with arbitrary (integral) multiples of any other. If we discover that, for positive integers m, n,

$$md_1 \sim nd_2 \text{ and } md_3 \sim nd_4$$

then we define

$$(d_1,d_2)\ \alpha\ (d_3,d_4).$$

It is assumed that this empirically established relation leads to a corresponding relation in $\tilde{D} \times \tilde{D}$, and that this relation is symmetric, reflexive, and transitive; i.e., it is an equivalence relation. However, it is possible that, for given (d_1,d_2), there is no pair of positive integers m, n, such that $md_1 \sim nd_2$; and, intuitively, it does not seem satisfactory that every such pair should be regarded as inequivalent under α; i.e., that every such pair should constitute an equivalence class with a single element. (For example, if (d_1,d_2) is such a pair, then, intuitively, the pair

$(2d_1,2d_2)$ should be equivalent to (d_1,d_2) .) The Greeks found such "incommensurable" pairs in their geometry, and Eudoxus' brilliant idea was to extend the definition of equivalent ratios by defining

$$(d_1,d_2)\ \alpha\ (d_3,d_4)$$

if, and only if, $md_1 < nd_2$ whenever $md_3 < nd_4$. [You should verify that this really is an extension of the relation α .] We assume that this leads (empirically) to a corresponding equivalence relation on $\tilde{D} \times \tilde{D}$. Observe that, as with the earlier procedure, we can never discover empirically whether or not Eudoxus' idea is actually needed for a theory for physical measurement: as long as we admit that our comparison procedure is necessarily imperfect, then, for any d_1 , d_2 , there will always be positive integers m , n , such that md_1 appears to be equivalent to nd_2 .

At this stage, in addition to the empirical structure of $(\tilde{D}, *, <)$ as a densely-ordered archimedean, abelian semigroup, we have an "equivalent-ratio" structure on $\tilde{D} \times \tilde{D}$. The corresponding geometrical structure was quite sufficient for the purposes of classical geometry, but if we want to obtain real-number-valued length functions, we must go a step further. We observe that the procedure for setting up the equivalent-ratio relation leads directly to a function β from $\tilde{D} \times \tilde{D}$ to R^+, a function which, in effect, gives a positive real value to each ratio. This is defined as follows:

If, for d_1 , $d_2 \in D$ and for some m , n ,

$$md_1 \sim nd_2$$

then we define $\beta : (d_1,d_2) \to \frac{n}{m}$. More generally, for any d_1 , d_2 , we define $\beta : (d_1,d_2) \to r$, where r is the real number determined by the cut $\{\frac{n}{m} : nd_2 < md_1\}$. (In order that this set be a cut, we are assuming that D is empirically archimedean: i.e., for any d_1 , d_2 , there exists an integer p , such that $pd_2 > d_1$.) We use the same symbol, β , to denote the resulting function on $\tilde{D} \times \tilde{D}$.

We assume that the following properties of the function β are consistent with the empirical evidence:

(i) $\beta(d_1 * d_2, d_3) = \beta(d_1,d_3) + \beta(d_2,d_3)$, for all d_1 , d_2 , d_3 .

(ii) $\beta(d_1,d_2)\beta(d_2,d_3) = \beta(d_1,d_3)$, for all d_1 , d_2 , d_3 .

In other words, β is an additive ratio operation. As usual, we also write $\beta(d_1,d_2)$ as a "formal fraction", $\frac{d_1}{d_2}$, and as $d_1 : d_2$.

From this point, it is only a short step to the description of a suitable length function. So far there is nothing to distinguish any particular element of $\tilde{D}$, so again we make an arbitrary choice of some $\tilde{d}_0 \in \tilde{D}$, and define a function

$$\lambda_0 : \tilde{D} \to R^+ .$$

by

$$\lambda_0(\tilde{d}) = \beta(\tilde{d}, \tilde{d}_0) = \tilde{d} : \tilde{d}_0$$

for each $\tilde{d} \in \tilde{D}$.

The verification that λ_0 is a suitable length function is a direct consequence of the many assumptions which we have made, and which (we have suggested) are consistent with empirical evidence. We note that the selected comparison class $\tilde{d}_0$ is the unit which corresponds to λ_0, as we would expect from a comparison of the above procedure with that of Method I.

It is interesting to compare the two procedures which we have outlined for the empirical determination of a length-measurement function. From a practical point of view, one of these procedures requires the existence of an unlimited number of "copies" of a selected unit rod, and the existence (or "construction") of suitable "fractional parts" of the selected unit rod; the other procedure requires that we have an unlimited supply of "copies" of every rod. The first method is closer to the actual procedure used in length-measurement with a ruler (or weighing with a balance). Neither procedure can enable us to resolve the question of whether or not rational numbers are sufficient for empirical measurement (or whether such a question is meaningful). In practice, of course, we usually use only rational numbers as values of empirical measure functions.

Length-functions, Units, and Values. We return now to a discussion of the relationship between length functions, units, and values. We assume that each length function maps $(\tilde{D}, *, <)$ isomorphically onto $(R^+, +, <)$, and note that the following properties appear to be consistent with the empirical evidence:

If $\lambda_1 \longleftrightarrow \tilde{d}_1$, $\lambda_2 \longleftrightarrow \tilde{d}_2$, $\lambda_3 \longleftrightarrow \tilde{d}_3$ under the 1-1 correspondence of functions and units, then

(i) $\dfrac{\lambda_1}{\lambda_2} = \dfrac{\tilde{d}_2}{\tilde{d}_1} = \dfrac{\lambda(\tilde{d}_2)}{\lambda(\tilde{d}_1)}$ for every $\lambda \in \Lambda$.

(ii) $\lambda_1(\tilde{d}_1) = \lambda_2(\tilde{d}_2) = \lambda_3(\tilde{d}_3) = 1$.

(iii) $\lambda_1(\tilde{d}_2) = \dfrac{1}{(\lambda_2(\tilde{d}_1))}$.

(iv) $\lambda_1(\tilde{d}_3) = \lambda_1(\tilde{d}_2) \cdot \lambda_2(\tilde{d}_3)$.

(Of course these properties are not all independent; see exercises below.)

Exercises 2-4 (continued)

(Assume, where necessary, that each length function maps $(\tilde{D}, *, <)$ isomorphically onto $(R^+, +, <)$.)

5. Verify the above properties (i) -- (iv).

6. Using d_1 as a unit, the lengths of d_2, d_3 are 7.6 and 9.5 respectively. What is the length of d_3 if d_2 is used as a unit? If the length of d_4 with d_3 as a unit is 12.8, what is the (length) ratio of d_4 and d_1 ?

7. When d_1 is used as a unit, the length of d is 7 . When d_2 is used as unit, the length of d is 11 . What is the length of d when $d_1 * d_2$ is used as a unit?

8. Show that if d_1, d_2, ... , d_n is a finite sequence of rods, whose lengths are in arithmetic progression when any rod d_0 is used as a unit, then the lengths are also in arithmetic progression when any other unit is used.

(Note: In view of the result of Exercise 8, we say that a sequence of rods is in arithmetic progression whenever their values (under any length function) are in arithmetic progression. We observe that such a statement is unit-free: it does not depend on the choice of units.)

9. Show that if $d_1, d_2, \dots, d_n$ is a sequence of rods whose values are in harmonic progression under any length function, that they are in harmonic progression under any other length function. (We say that such a sequence of rods is in harmonic progression; this is also a unit-free statement.)

10. Show that if a sequence of length functions $\lambda_1, \lambda_2, \dots, \lambda_n$ has the property that, for some particular $d \in D$, $\lambda_1(d), \lambda_2(d), \dots, \lambda_n(d)$ are in arithmetic progression (harmonic progression), then the same is true for the values on every other rod in D. (Such a sequence of functions is said to be in arithmetic progression (harmonic progression); we could have defined these terms directly, using the structure in Λ.)

11. Let $(d_i) = (d_1, d_2, \dots, d_n)$ be a sequence of rods, and let $(\lambda_i) = (\lambda_1, \lambda_2, \dots, \lambda_n)$ be the corresponding length functions. Prove that (d_i) is an arithmetic progression if and only if (λ_i) is a harmonic progression; and that (d_i) is a harmonic progression if and only if (λ_i) is an arithmetic progression. Hence show that (under the 1-1 correspondence of units and functions) the arithmetic mean of two units corresponds to the harmonic mean of the related length function (which gives the harmonic mean of the values, on each element of the domain), and that the harmonic mean of two units corresponds to the arithmetic mean of the related functions. (Cf. Exercise 7 above.)

12. In the context of classical euclidean geometry, devise constructions for finding the arithmetic mean and the harmonic mean of two line segments.

13. If $\beta : A \times A \to R^+$ is any ratio operation, prove that

(a) for all $a_1, a_2 \in A$,

$$\beta(a_1, a_2) = \frac{1}{\beta(a_2, a_1)} ;$$

(b) for all $a \in A$,

$$\beta(a, a) = 1 .$$

14. If λ is a length function which has a unit, and if $\beta : D \times D \to R^+$ is the ratio operation for length, prove that, for any $d_1, d_2 \in D$,

$$\beta(d_1,d_2) = \frac{[\lambda(d_1)]}{[\lambda(d_2)]}.$$

(In other words, the "length ratio" of d_1 to d_2 is the same as the ratio of the numbers $\lambda(d_1)$ and $\lambda(d_2)$, for every λ which has a unit. If length functions are not assumed to be onto R^+, it is not necessary that each length function should have a unit. The above result is still true for such a function, but in order to prove it we must use the property that all length functions are similar. In the next section we shall see that this property can be proved as a consequence of the axioms of classical geometry, without assuming that each length function is onto R^+.)

15. If G denotes the set of all functions from a set A to R^+, there is a "scalar multiplication" of elements of G by positive real numbers, as defined in Section 1-5. In terms of this multiplication, we define a function

$$\varphi : G \times G \to G$$

to be homogeneous of degree α, if there exists $\alpha \in R$, such that

$$\varphi(kg_1,kg_2) = k^{\alpha}\varphi(g_1,g_2)$$

for every k, and every $(g_1,g_2) \in G \times G$. Let φ be a function

$$\varphi : G \times G \to G ,$$

which is obtained from a finite number of the operations (in G) of addition, multiplication, scalar multiplication by positive real numbers, and the formation of powers. Then φ determines a "corresponding" function

$$\bar{\varphi} : R^+ \times R^+ \to R^+ ,$$

such that $[\varphi(g_1,g_2)](a) = \bar{\varphi}(g_1(a),g_2(a))$, for all $g_1, g_2 \in G$, and all $a \in A$. Prove that φ is homogeneous of degree α if and only if $\bar{\varphi}$ is homogeneous of degree α.

16. With the same notation as Exercise 15, let A be the domain of length functions and let $F (\subset G)$ be the subset of all length functions. Let $\varphi: G \times G \to G$ be a homogeneous function of degree 1, formed as in Exercise 15. Prove that if f_1, $f_2 \in F$, then $\varphi(f_1,f_2) \in F$. That is, every such homogeneous function (of degree 1) of two length functions is also a length function. In particular, if λ_1 and λ_2 are length functions, then so are the functions $\sqrt{\lambda_1\lambda_2}$ and $\sqrt{\lambda_1^2 + \lambda_2^2}$. [We remark that a corresponding result holds for every ratio scale.]

Remarks:

1. The results of Exercises 8 - 11 and 14 above carry over to other sets of measure functions whose domains are densely-ordered abelian semigroups, whose ranges are R^+, and which are similarity invariant. (E.g., area, time intervals, mass, etc.)

2. Perhaps you have observed the similarity of the formulas (ii) - (iv) above, relating length-functions, units, and values, to the formulas which relate logarithmic functions, bases, and values; and also to the rules for differentiation in calculus. For example, compare (iv) with the formula

$$\log_a c = \log_b c \cdot \log_a b ,$$

and with the chain rule for the derivative of a composite function. (With a suitable choice of notation we could make corresponding formulas identical.) You might find it interesting to look for the underlying reasons for this similarity.

The facts that

(i) a length function is fully determined when a value is assigned to any element of the domain; (i.e., there is exactly one length function which has a given value on a given domain element);

(ii) any two length functions differ by composition with a positive similarity transformation;

are of very great practical importance. In particular, they imply that a function for length measurement can be completely established on the basis of a selected measure for one object (usually the unit, with measure 1) ; and that, in order to convert from one system to another, we only need to know the length of a single object in both systems: it is not necessary to compare units directly.

Use of the Term "Scale". This word is used in a variety of ways in relationship to measurement questions, and it would probably be hopeless to attempt to fix on a single meaning. For example, "ratio scale", as used by Stevens for the classification of measure functions, refers to an equivalence class of similar functions, such as our class Λ' of length functions. The word "scale" is also used for each function in such an equivalence class: e.g., the metric scale of length. It is also used to denote an object, or machine, employed for obtaining values (i.e., "generating" the function) on some (usually restricted) part of the domain: e.g., a scale used for weighing. Another related use is in such expressions as "scale model" and "the scale of a map": we shall have more to say later about some of these uses.

More About Language. Let us look at the way in which a word like "inch" is used. From our point of view, "inch" may be regarded either as the name for a unit, (i.e., the name of a particular equivalence class of rods) or as the name for the corresponding function: if we assume that there is a 1-1 correspondence of units and functions, then the choice may be regarded as immaterial. To conform to general usage we should regard "inch" as the name of the unit, and refer to the corresponding function as "the inch function" or λ_{in}. Similarly let λ_{ft} denote the foot function, λ_m the meter function, etc. With this convention we list together a few common expressions relating to length measurement, and the corresponding expressions in functional language:

	Common Usage	In Functional Language
(i)	This box is 4 inches long; or the length of this box is 4 inches.	$\lambda_{in}(\text{this box}) = 4$
(ii)	The sum of the lengths of box A and box B is 7 inches.	$\lambda_{in}(A) + \lambda_{in}(B) = 7$
(iii)	Length of A = 7 inches.	$\lambda_{in}(A) = 7$

(iv) $L = 7$.	$\lambda_b(X) = 7$. (The common abbreviation is meaningless unless there is an implied unit, "b", and an implied object, X.)
(v) 12 inches = 1 foot.	$\lambda_{in}(A) = 12$, if and only if $\lambda_{ft}(A) = 1$; or $\lambda_{in} = \overline{12}\lambda_{ft} = 12\lambda_{ft}$; or foot: inch = 12.
(vi) The length of A is 9ft 7 in.	$A = A_1 * A_2$, and $\lambda_{ft}(A_1) = 9$, $\lambda_{in}(A_2) = 7$; or, $\lambda_{ft}(A) = 9.7$ (base 12) ; or, $\lambda_{in}(A) = 12\lambda_{ft}(A)$ $= 97$ (base 12) $= 115$ (base 10) .
(vii) 2 ft > 23 in.	If $\lambda_{ft}(A) = 2$ and $\lambda_{in}(B) = 23$ then $B < A$ (where $<$ is the order relation in D).

We observe that the common usage for (v) and (vii) has a perfectly clear meaning in terms of the "scalar multiplication" (by positive integers) of elements (such as "inch", "foot") in $\tilde{D}$.

This seems an appropriate place to ask whether or not any meaning can be given to the expression "the length of A", where A is an object in the domain of those objects which are length measurable. We can do this most simply by the technical device of defining the <u>length of A</u> to be the equivalence class (in $\tilde{D}$) to which A belongs. This enables us to make all sorts of meaningful statements about lengths of objects in terms of the structure in $\tilde{D}$ -- i.e., without reference to any particular units or length functions. With this meaning, we can interpret a statement such as "the length of A is 7 feet" as equivalent to "7 foot is the equivalence class in $\tilde{D}$ to which A belongs". We can also give a clear meaning to a statement such as "A is longer than B", without reference to functions or units: "A longer than B" means simply that $\tilde{A} > \tilde{B}$ in terms of the order relation established in $\tilde{D}$.

<u>Unit-free Statements</u>. Many of the statements which we can make about length, are true no matter which particular length function (or unit) we use. Such statements are often called <u>unit-free</u> or <u>invariant</u> statements about length. We have already seen examples of these in the exercises above. (E.g., lengths of rods being in arithmetic progression.) Some further examples of unit-free statements are:

(i) Jack is as tall as John;

(ii) Jack is taller than Jill;

(iii) Bill is half as tall as his father;

(iv) this stick is as long as those two sticks combined;

(v) the length of A is the average of the lengths of B, C, and D;

(vi) the perimeter of a square is four times the length of a side;

(vii) the ratio of the lengths of those poles is the same as the ratio of the lengths of their shadows;

(viii) the ratio of the length of this pole to the length of its shadow, is greater at noon than it is at 4 p.m.

All of these statements have clear meanings in terms of the structure in $\tilde{D}$; the final step, to the definition of actual measure functions on $\tilde{D}$, is not needed in order to give meaning to these statements.

The Domain of Λ. The whole description of length measurement in this section has been based on rather vaguely suggested empirical operations on a certain set of real objects. It is clear that the domain to which the suggested procedures could be applied, is "too small", and that other operations must be used if length measurement is to be possible in situations where the primitive procedures described cannot be used. What we seek to do is to "extend the domain", by using procedures which are applicable to an enlarged domain; which lead, if possible, to a corresponding structure on the domain (i.e., a densely ordered, abelian, archimedean semigroup); and which "agree" with our earlier procedures where both are applicable. You will find discussion of some aspects of this question in [2], and in Chapter 1 of [13]. It is worth noting that, in extending the domain, we wish to include objects which no longer correspond to "linear" situations, but which are "curved". This idea will come up again in the next section, where you will see that questions of domain, and extension of the domain, are also important in "mathematical" measurement; i.e., in the discussion of defined measure functions (length, area, etc.) in formal mathematical systems.

Of course, the mathematical and physical ideas cannot really be separated, except at a most rudimentary level. Many procedures for extending the empirical domain involve the use of mathematical theories (e.g., trigonometry). Moreover, as suggested in the preface, lengths of real objects are often arrived at by the use of "models"; i.e., by first "mapping" the object into a suitable mathematical system, and then arriving at the length by a combination of empirical and mathematical processes. For example, the circumference of a wheel is commonly "measured", by measuring the length of its diameter, mapping (implicitly) the wheel into a circle whose diameter has the same length as that of the wheel, and then calculating the length of the circumference of the circle, using the theory of mathematical length-measurement. The justification for this procedure lies in the fact that it leads to results which agree (within the accuracy of the empirical processes involved) with the result which would be obtained by direct physical measurement.

Situations dealing with "model-building" often involve assumptions which can be conveniently indicated by the use of commutative diagrams. For example, in the simple situation of the wheel, we are led to a diagram like that below:

$$\begin{tikzcd}
B \arrow[r, "\mu"'] \arrow[d, "i"'] \arrow[rr, bend left, "\lambda_e"] & D \arrow[r, "\lambda_m"] \arrow[d, "j"] & R^+ \arrow[d, "\overline{\pi}"] \\
W \arrow[r, "\mu"'] \arrow[rr, bend right, "\lambda_e"'] & C \arrow[r, "\lambda_m"] & R^+
\end{tikzcd}$$

B denotes the set of wheel diameters, W the set of wheels, and i the natural map of a diameter to the corresponding wheel. D denotes the set of plane line segments, C the corresponding circles (for which these segments are diameters) in the plane, and j the natural map of a segment to the circle with that segment as diameter. Commutativity in the left rectangle is a simple consequence of assumptions concerning the nature of the model mapping μ. Commutativity in the right rectangle is a consequence of a mathematical result concerning the relationship of the length measures of a circle and any of its diameters. (This result is valid no matter which particular mathematical length measure function λ_m we choose.) λ_e denotes an empirical length function on a domain which includes B and W.

The empirical hypothesis, which justifies the usual method for finding the circumference of the wheel (i.e., select a unit, measure the diameter, and multiply this number by π), is the following:

> If λ_e and λ_m are chosen so that the top "triangle" is commutative (i.e., so that under μ, the unit for λ_e maps into the unit for λ_m), then the bottom triangle is also commutative.

With this hypothesis, it is readily verified that the whole diagram is now commutative, and hence we obtain the usual "formula":

$$\lambda_e(w) = \pi\lambda_e(b) \text{ , where } i(b) = w \text{ .}$$

This is just a very special case of a general situation concerning the use of mathematical models: frequently assumptions (hypotheses) concerning the validity of a model, can be stated as "commutativity" conditions concerning mathematical and empirical relations and functions. The connecting functions between the "empirical world" and the "mathematical world" are usually measure functions. It is not generally possible to prove the validity of the model, (i.e., check all commutativity properties) because of the complexity of the domains of the functions and relations involved. But to prove the invalidity of the model, it is only necessary to find a single case where commutativity fails. This is what is frequently happening when some experiment is devised to test the validity of a certain hypothesis involving the use of mathematical models: commutativity is being checked for a single domain element, with respect to two different "function paths".

Ratio and Comparison. We conclude this section with a comment on a frequently heard statement

(i) "We can only compare like things.",

which has some relationship to the subject of measurement. This statement is often used incorrectly, to support the claim that, for example, it doesn't make sense to make a statement such as

(ii) "The ratio of the number of oranges in this box to the number of cars in the parking lot is 12.3."

From our point of view, the only reasonable meaning to be given to statement (i) is that, when used to compare objects, it is equivalent to something like:

(iii) "We can only meaningfully compare objects in terms of some common measurable attribute; i.e., objects which belong to the domain of some common measure function."

We may then compare them (with respect to the attribute which the function measures) in terms of the "ratio structure" of the domain, or by comparing their functional values in R^+. Observe that if the measure function is similarity-invariant (e.g., length), then it does not matter which of the equivalent measure functions we choose if we are only interested in the ratio of the values.

With the meaning that we have given to statement (i), statement (ii) is meaningful: the attribute in question is numerosity. But a statement such as "the length of this box is greater than the weight of this box" is not meaningful: the box is first referred to as determining an element of the domain of a length function, while in the second reference the box determines an element of the domain of a weight function, and these domains must be considered as disjoint, in spite of the fact that the same material object determines an element of each.

On the other hand, a statement like "the weight of this box in pounds is greater than the length of this box in inches" is a meaningful, even if rather uninteresting, statement about numbers; it is numbers which are being compared, and not the domain elements which determine them through the specified measure functions.

2-5 Length In Formal Mathematical Systems

In order to discuss the concept of length within a formal mathematical system, we must, of course, have a precise description of that system. For example, if our mathematical system is the so-called number line, we can define a mathematical concept of length in a particularly simple way: if A, $B \in R$, $A \neq B$, we may define the segment $\overline{AB}$ by

$$\overline{AB} = \{p : p \in R \text{ , and } A \leq p \leq B \text{ or } B \leq p \leq A\}$$

and we can define a "length" function λ on the set of all segments by

$$\lambda(\overline{AB}) = |A - B| .$$

($\lambda(\overline{AB})$ is usually denoted in geometry by AB : we use this notation where convenient.)

The corresponding situation in the cartesian plane $R \times R$, and in cartesian 3-space, $R \times R \times R$, is well-known, and very little more complicated. Does this mean that the mathematical theory of length is completely simple -- virtually trivial? Not at all: two important questions should immediately occur to you:

(i) What is the justification for assuming (for example) a 1-1 correspondence between points of the line, the plane, and space, and the real numbers, ordered pairs of real numbers, and ordered triples of real numbers, respectively?

(ii) What about length for subsets of space which are not line segments?

Question (i) amounts to asking for a justification for cartesian geometry. Such a justification could be empirical, or it could be given in terms of some other axiomatic treatment of geometry. Question (ii) is the very important question of "extension of the domain".

We look first at mathematical matters related to question (i); before we can attempt an answer, we must be clear about what the question means. Roughly speaking, it asks for a procedure by means of which we may establish the isomorphism of a "given" geometry with cartesian geometry. In order to answer this question we have to be clear about the geometry that is "given", i.e., about our assumptions, or postulates.

Unfortunately, no matter how we proceed, it is much harder to give a full system of axioms for "geometry", than it is for the natural numbers. Thus, although in principle this section is capable of precise presentation, this implies that we should first make clear all of our assumptions. We will compromise a little on this ideal, and hope that the omissions will not obscure the essential ideas.

We shall examine the question of "mathematical length" in three main contexts: that of classical euclidean geometry, that of cartesian geometry, and from the "intermediate" standpoint of the treatment of geometry which was suggested by G. D. Birkhoff, and carried through in [14], [15] and other recent books. It is convenient to follow the terminology of Moise, and refer to this last approach as <u>metric geometry</u>.

Length in Metric Geometry. If we examine a metric treatment of geometry (such as the SMSG treatment) based on "ruler" and "protractor" axioms we find that (if S denotes space) a distance function

$$\alpha : S \times S \to R$$

is postulated. In other words, for each pair of points $A , B \in S$, it is assumed that a distance, $\alpha(A,B)$ is given. (This notation may be abbreviated to AB , but it is more instructive, for some of our purposes, to use the longer notation.) The following properties are postulated for α :

(i) For all $A , B \in S$, $\alpha(A,B) \geq 0$.

(ii) $\alpha(A,B) = 0$ if and only if $A = B$.

(iii) $\alpha(A,B) = \alpha(B,A)$.

A line is a set of points. (This is not a definition: "line" is an undefined concept in metric geometry.) A coordinate system for a line ℓ , is defined to be a 1-1 correspondence of the given line and the real number system, R , (i.e., a 1-1 onto function $\varphi: \ell \to R$). which is related to the distance function, α , as follows:

If $A , B \in \ell$, then $\alpha(A,B) = |\varphi(A) - \varphi(B)|$.

It is then postulated (the so-called "ruler postulate") that, for each line ℓ , there exists a coordinate system. (It is further postulated that if $P , Q \in \ell$, $P \neq Q$, then there exists a coordinate system, φ , for ℓ , such that $\varphi(P) = 0$; and $\varphi(Q) > 0$; this postulate is not really needed: it can be proved as a theorem.) We economize on notation by using the same symbol φ for each of the postulated coordinate functions, one for each line ℓ .

It can be shown (see exercises) that if φ is composed with any rigid motion of R (i.e., a transformation ρ on R of the type $\rho : x \to jx + b$, where $j = +1$ or -1 , $b \in R$) then $\rho\varphi$ is also a coordinate system for ℓ ; i.e., each line has infinitely many coordinate systems, which are related to each other by composition with rigid motions, and each of which is related to the given distance function, as in the ruler postulate.

It is further postulated that each pair of (different) points, A , B , determines (uniquely) a line, $\overleftrightarrow{AB}$, and betweenness is defined by:

C is between A and B (denoted by $A - C - B$) provided that $C \in \overleftrightarrow{AB}$ and $AC + CB = AB$. For $A \neq B$, the segment $\overline{AB}$ is now defined by $\overline{AB} = \{C : C \in \overleftrightarrow{AB}$ and $A - C - B\}$. It follows immediately that $\overline{AB} = \overline{BA}$. Thus a 1-1 correspondence is established between the set of (unordered) pairs of distinct points in S, and the set, D, of all segments in S.

Length Functions. In metric geometry a function

$$\lambda : D \to R^+$$

is defined to be a length function if it satisfies

(i) whenever $A - C - B$,

$$\lambda(\overline{AC}) + \lambda(\overline{CB}) = \lambda(\overline{AB}) ;$$

(ii) λ gives the same value to congruent segments. (See below for definition.)

In metric geometry, the function

$$\lambda_0 : D \to R^+$$

defined by $\lambda_0(AB) = \alpha(A,B)$, satisfies (i) trivially, from the definition of betweenness. We shall see below that it also satisfies (ii), and hence that it is a length function.

In view of the fact that a segment is uniquely determined by its end points, so that the functions α and λ_0 are closely related, you might think that it is an unnecessary luxury to introduce the length function λ_0, in addition to the distance function, α. We have done this for two main reasons:

(i) it makes the treatment more closely parallel to that of the previous section; and

(ii) the domains of λ_0 and α are quite distinct. When it comes to the question of extending the domain of the length functions, it is the domain D which we will wish to extend, and not the domain of α. In other words, we will wish to extend the concept of length, so that it applies to sets of points which are not line segments.

The function λ_0 is a length function for the set D. Thinking along the same lines as for the empirical length functions of the last section, we ask whether or not there are any other mathematical "length functions" for D, and, if more than one, how are different length functions related? We shall see that, as in the previous section, all functions which are similar to λ_0 are also length functions.

Congruence and "Length Structure" for Segments. We can establish a "length structure" in D by making strong use of the given distance function α. We define first a relation "has the same length as " (which we denote by the symbol " $\cong$ ") by

$$\overline{AB} \cong \overline{CD} \text{ if and only if } \alpha(A,B) = \alpha(C,D) .$$

The usual word used in geometry for this relation on segments, is congruent. It is trivial to verify that congruence of segments is an equivalence relation: we denote the set of equivalence classes by $\tilde{D}$. Trivially, the defined function λ_0 gives the same value to congruent segments, and hence is a length function. Any length function for D must give the same value to congruent segments, and hence determine a corresponding (length) function for $\tilde{D}$: it is convenient to use the same name for corresponding functions. We also use the given distance function to define a relation "less than in length" (<) in D (and hence in $\tilde{D}$) by

$$\overline{AB} < \overline{CD} \text{ if and only if } \alpha(A,B) < \alpha(C,D) ,$$

and we prove easily that this gives a (strict total) order relation in $\tilde{D}$.

We look next for some operation in D or $\tilde{D}$ which will be analogous to the join operation of the previous section. We will not be able to carry out physical operations on our segments (such as placing them end to end) but our axioms permit us to do something rather like this: we can use the coordinate structure of each line to prove that, given any two segments $\overline{AB}$, $\overline{CD}$, and any line ℓ , there exist (in infinitely many ways) points E , F , G on ℓ , such that E - F - G , $\overline{AB} \cong \overline{EF}$, and $\overline{CD} \cong \overline{FG}$. It is natural to define

$$\overline{EG} = \overline{EF} * \overline{FG}$$

but of course this does not lead to a binary operation on D , because relatively few pairs of segments will be suitably located so that they can be "joined". However, our main interest is in $\tilde{D}$, and not in D , and any "additive" function on $\tilde{D}$ immediately yields a corresponding length function on D . We can now prove that, although $\overline{EG}$ is certainly not uniquely determined by $\overline{AB}$, $\overline{CD}$, the equivalence class of $\overline{EG}$ is uniquely determined by the equivalence classes of $\overline{AB}$ and $\overline{CD}$. (You should provide this proof.)

This enables us to define a join operation on the set $\tilde{D}$ of congruence classes of segments. If $\tilde{d}_1$, $\tilde{d}_2$ are two such classes, denote the join class by $\tilde{d}_1 * \tilde{d}_2$. (Notice, by the way, that while the join of two rods in the previous section placed them end to end, but (empirically) disjoint, the join of two segments exists only when they are collinear, and when they intersect in a common end point.)

We can now prove, purely within the formal framework of our geometry, that the system $(\tilde{D}, *, <)$ is a densely ordered archimedean, abelian semigroup, and that the join operation and the order relation are connected by the property that: $d_1 < d_2$ if and only if there exists d_3 with $\tilde{d}_1 * \tilde{d}_3 = \tilde{d}_2$. If you provide these (and other) omitted proofs for yourself, using the definitions and postulates given, you will appreciate that their simplicity depends mainly on the existence of the postulated distance function α, and on the existence of the postulated coordinate system, φ, for each line. In other words, these are very powerful axioms. (You will probably appreciate this even more, after you have seen how difficult it is to introduce length functions into classical synthetic geometry.)

Having now established the structure of $\tilde{D}$ as a densely ordered, archimedean, abelian semigroup, we could go further and establish a theory of "ratios" in $\tilde{D}$, by again appealing to the given distance function α: we shall not stop to do this; it does not present any difficulty.

If we now reverse our viewpoint, and ask whether there exist any functions from $\tilde{D}$ to R^+ which preserve the structure of $\tilde{D}$, it is no surprise to discover that the defined length function λ_0 (and, more generally, any length function, λ) yields a corresponding function on $\tilde{D}$ (we use the same symbol, λ_0, for this function) and that this function (and, more generally, any length function, λ) is structure-preserving. (You should verify this: see exercises.) Moreover the coordinate-system postulate ensures that λ_0 is onto, and hence we may use the corresponding result from the previous section to deduce that there exist other isomorphisms of $(\tilde{D}, *, <)$ onto $(R^+, +, <)$, and that these differ from λ_0 (and from one another) by composition with an automorphism (i.e., a positive similarity: see Theorem 2-2.2) of $(R^+, +, <)$.

We denote the set of functions thus obtained by Λ, and use the same notation for the set of corresponding functions on D (which are obtained by composing the natural "classification" function, $D \to \tilde{D}$, with functions in Λ). Each such function, $\lambda : D \to R^+$, is easily shown to be a length function, and every two such functions are similar. The set, Λ, of length functions is an example of a ratio scale.

We might now ask how these new length functions are connected with our geometry? To begin with, we remind you that our first length function, λ_0, was directly derived from our postulated distance function α . Could we reverse the process, and construct other distance functions from all of the other length functions which we have now discovered? I.e., if $\lambda_1 \in \Lambda$, let us define

$$\alpha_1 : S \times S \to R$$

by

$$\alpha_1(A,B) = \begin{cases} \lambda_1(\overline{AB}) & \text{if } A \neq B ; \\ 0 & \text{if } A = B . \end{cases}$$

We know that there exists $k > 0$ such that $\lambda_1 = k\lambda$, and hence $\alpha_1 = k\alpha$. Using this, we can easily prove that α_1 has the following properties (which, we recall, were postulated for α) .

(i) For every $A , B \in S$, $\alpha_1(A,B) \geq 0$.

(ii) $\alpha_1(A,B) = 0$ if, and only if, $A = B$.

(iii) $\alpha_1(A,B) = \alpha_1(B,A)$.

We can define a new notion of "betweenness" corresponding to α_1 , and we can verify that

(iv) $A - B - C(\alpha)$ if, and only if, $A - B - C(\alpha_1)$. (The notation should be self-explanatory.)

We have already seen that, although one coordinate system φ was postulated for each line ℓ , there are infinitely many others, related to φ by composition with rigid motions of R . By direct checking we can verify that, if $k \neq 1$, none of these is related to α_1 in the way that φ is related to α . However, it is easy to verify that, for each line ℓ , the "coordinate" function $\varphi_1 = k\varphi$ (and, more generally, any function derived from $k\varphi$ by a rigid motion) is a 1-1 mapping of ℓ onto R , and that it is related to α_1 in the same way that φ is related to α . That is, for all points $A , B \in \ell$, $\alpha_1(A,B) = k\alpha(A,B) = k|\varphi(A) - \varphi(B)| = |k\varphi(A) - k\varphi(B)| = |\varphi_1(A) - \varphi_1(B)|$.

All of this suggests that we might ask the question: if we develop a new "geometry" by using α_1 , φ_1 , in place of α , φ , how will it be related to our postulated geometry? The answer is, of course, that there will be no discernible difference: every significant theorem is the one geometry is a theorem in the other, so it seems reasonable to say that the geometries on S are isomorphic or equivalent, in a sense that is not too difficult to make precise. We shall not examine this question in detail, but if you have filled in the proofs which we omitted earlier in this section, then you will have done most of the work.

Perhaps the relationship of distance/length and congruence should be made a little clearer. If, as in the metric approach to geometry, we postulate a distance function, then congruence is defined in terms of that function. For reasons of simplicity, congruence is often defined separately for various subsets of space (segments, angles, triangles, and so on) but eventually these are all brought together with a single definition: subsets K_1 and K_2 of space (S) are said to be congruent (written $K_1 \cong K_2$), if and only if there is a rigid motion, or congruence, of S onto S, which maps K_1 onto K_2. A rigid motion is defined to be a 1-1 correspondence which preserves distances. I.e.,

$$\rho : S \to S$$

is a rigid motion, if it is 1-1 and onto, and if for all pairs of points A, B $\in$ S,

$$\alpha(A,B) = \alpha(\rho(A), \rho(B))$$

where α is the distance function. Now if α_1 is another distance function, differing from α by composition with a positive similarity of R, then we have

$$\begin{aligned}\alpha_1(A,B) = k\alpha(A,B) &= k\alpha(\rho(A), \rho(B)) \\ &= \alpha_1(\rho(A), \rho(B))\end{aligned}$$

In other words: ρ is a rigid motion in terms of α, if and only if it is a rigid motion in terms of α_1. It follows that the basic notion of congruence is unaffected by the change of distance function which we have introduced. It can be shown that composition with a positive similarity is the only possible variation for α, if the set (actually a group) of rigid motions of S (and hence the notion of congruence) is to be unchanged.

We can sum up the above by saying that, although the treatment of geometry which we have sketched (in part) postulated a particular distance function $\overset{\circ}{\alpha}$, and a particular coordinate system φ, there are infinitely many other coordinate systems which correspond to the given distance function; and there are infinitely many other equivalent distance functions, each with its own related set of coordinate systems. Moreover these additional distance functions and coordinate systems are related to the postulated functions and to each other by positive similarities. Corresponding to each distance function there is a length function on segments, and the set of admissible length functions is invariant under composition with positive similarities. The basic notions of betweenness and congruence are the same for all equivalent distance functions.

In the Birkhoff treatment of geometry, whenever a "length" is mentioned it is understood that this is the "length" in terms of the postulated distance function, and a related coordinate system. Thus the length function is clearly determined and understood, and need not be specifically identified.

Corresponding to each length function, there is a unit; ~~this is the~~ equivalence class, in $\tilde{D}$, which maps into the real number 1, under the length function. Conversely, given any equivalence class in $\tilde{D}$, there is a length function for which that class is the unit.

The situation is very similar to that described in the section on the empirical idea of length, and many of the results of that section will, of course, carry over. In particular, there are the same relationships between units and length functions and the results of all of the relevant Exercises 2-4 are still true. In fact, the proofs are essentially the same as those which (hopefully) you supplied before.

It should not have surprised you that, although the metric treatment of geometry appeared to favor one particular length function and one particular coordinate function, an identical treatment could be carried out by using any equivalent length function, and any corresponding coordinate function. After all, the geometry of Euclid was carried through without the use of any length function, using only the structure of $\tilde{D}$ as a densely ordered abelian semi-group -- a structure which (not in those terms of course) was derived from a unit-free set of axioms. (See later.)

Before leaving the metric treatment, there are several matters which we should mention:

(i) There are important questions concerning extension of domain: these are not specific to the metric treatment, so we defer consideration until later.

(ii) From the beginnings which we have sketched for a metric treatment, one can proceed to develop the familiar results of geometry, including, in particular, the possibility of setting up a cartesian coordinate system for the whole of space: i.e., a 1-1 mapping β of S onto $R \times R \times R$ with the property that

$\alpha(A,B) = \{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2\}^{1/2}$, where $\beta(A) = (a_1,a_2,a_3)$, $\beta(B) = (b_1,b_2,b_3)$. Details of this development can be found in the references.

(iii) If you examine the many theorems (in a metric treatment of geometry) which involve the concept of length, you will find that the actual values of the postulated distance function are not used in any essential way. In other words, for any segment it is only the congruence class in $\tilde{D}$ which is involved, and this is the same under all admissible length functions.

(iv) Two important theorems which involve length, are Pythagoras' Theorem (whose result is easily shown invariant under a similarity transformation of R^+) and the so-called "triangle inequality", which states that the sum of the lengths of any two sides of a triangle is greater than the length of the third side. Notice that if the "length of a segment" were to be interpreted as the equivalence class in $\tilde{D}$ to which the segment belongs, the latter theorem is still valid in terms of the structure of $\tilde{D}$; but the corresponding interpretation of Pythagoras' Theorem is not satisfactory, because we do not have a multiplication in $\tilde{D}$. This does not mean that Pythagoras' Theorem forces us to use the values of a length function in R^+, but it does mean that, if we wish to avoid the use of values, then we must proceed in some other way. For example, we could consider an appropriate structure in $\tilde{D} \times \tilde{D}$, and this would lead us in the direction of area. This is, of course, the way in which the Greeks saw Pythagoras' Theorem: for them it was a theorem about area.

(v) In view of the impossibility of deciding empirically whether or not we really needed R^+ as a value space for our empirical length functions, we might question the use of the real numbers, rather than the rational numbers, or some other subset of R, in the postulated distance-function, and in the coordinate-system postulate; (and in the angle measure function which is also postulated in metric geometry).

We look briefly at this last question. As part of the systematic development mentioned in (ii), we will have proved Pythagoras' Theorem. From this, we see that, if we had tentatively begun with the rationals Q, we would reach a point where we would require a number whose square is 2, in order to satisfy our distance postulate. (I.e., if every segment has a "length", then the length of the hypotenuse of a right isosceles triangle whose congruent sides have length "1", must be $\sqrt{2}$. This tells us, in particular, that the hypothesis that euclidean geometry is a suitable model for real space, implies that there exist irrationally related "distances" in real space. As remarked earlier, whether this is, in fact, the case, can never be decided empirically, and it might not even be meaningful to ask the question.) As is well-known, the real number $\sqrt{2}$ is not rational. This does not, of course, tell us that we must use the real numbers: it merely says that the rationals are insufficient. We might proceed cautiously, and tentatively try out the surd numbers. (Briefly, the <u>surd field</u> is a subfield of the real numbers: it is the least field which contains the rationals, and which is closed under any finite number of the operations of addition, subtraction, multiplication, division, and square root extraction. It is far from being the whole real field: for example it does not even contain the simple algebraic number $\sqrt[3]{2}$, a fact of considerable importance in the proof that there is no geometric construction for "duplicating the cube".) This would get around the difficulty produced by Pythagoras' Theorem, and it would give us a geometry which is satisfactory for many purposes, and which (with suitable definitions of the terms involved) can be shown to satisfy the postulates of classical synthetic geometry. But if, for example, we were to require the existence of angles whose measures are integral submultiples of existing angles, we would encounter another problem: some angles (e.g., an angle whose degree measure is 60) would fail to have "trisectors". At this stage we might well wish to impose conditions which would require that our image space should at least include the algebraic numbers. (This field contains all those real numbers which are roots of some polynomial equation with integer coefficients: it is intermediate between the surd field and the real field.) It is possible that by putting more and more demands on our geometry we could reach a point where the reals would be needed, but, as stated above, such a requirement is not needed in order to satisfy the axioms of classical synthetic geometry, for which the surd field is quite adequate. Thus the ruler postulate, in requiring a function which is <u>onto</u> the reals (which implies that the distance function is also onto) represents a strengthening of classical geometry. Metric geometry, using the full set of real numbers, is only one of the geometries which satisfy the axioms of classical synthetic geometry. (We shall see later just what additional postulates are needed if synthetic geometry is to be necessarily isomorphic to metric geometry, and to real cartesian geometry.)

Exercises 2-5

1. Show that the rigid motions of R, defined by

$$\rho : x \rightarrow jx + b \ (j = +1 \text{ or } -1 , b \in R)$$

form a (non-abelian) group under composition, and that this group is a subgroup of the affine group $\underline{A}$. (Cf. Section 1-6.)

2. If ρ is a rigid motion on R, and α, φ, denote the postulated distance function and coordinate function, for a line ℓ, show that the composite transformation $\rho\varphi$ is also a coordinate system for ℓ. That is, $\rho\varphi$ is 1-1 onto, and has the property that for A, $B \in \ell$,

$$\alpha(A,B) = |\rho\varphi(A) - \rho\varphi(B)| .$$

3. Verify that the function

$$\lambda : \tilde{D} \rightarrow R^{+}$$

derived from the postulated distance function and the ruler postulate of metric geometry, is an isomorphism of the ordered abelian semigroup $(\tilde{D}, *, <)$ onto the ordered abelian semigroup $(R^{+}, +, <)$.

4. Prove that the set of functions

$$F = \{f : f = \rho\bar{k} , k \in R^{+}\}$$

where ρ is a rigid motion of R, is the group $\underline{A}$ of affine transformations of R.

Length in Synthetic Geometry. In a certain sense length does not appear in classical euclidean geometry (usually referred to as synthetic geometry) so the heading above might be considered inappropriate. But what we are going to do, is show that a notion of length is implicit in synthetic geometry, and that it may be introduced explicitly, so as to give the distance/coordinate structure which is postulated in the metric approach (and from which the fully-coordinate cartesian structure can be developed). A fully detailed treatment would be too lengthy for this book, but we can sketch the main lines of the development. You will see that these have considerable similarity to the empirical procedure which we described for the establishment of a length structure, with segments corresponding to rods, and congruence corresponding to the empirical relation "equivalent with respect to length".

In synthetic geometry, we have the usual incidence structure, with points, lines, planes (as undefined concepts) related by the so-called "incidence axioms"; (e.g., given two different points, there exists exactly one line containing them; given three non-collinear points, there exists exactly one plane containing them; and so on). These are the same as in the metric treatment. But, whereas in the metric approach a distance/coordinate structure is postulated, and the concepts of congruence and betweenness are defined and their properties proved, in the synthetic approach congruence and betweenness are undefined concepts, with postulated properties. Insofar as these postulates involve segments, we shall have to use them in order to establish a concept of length, so we go into some detail here:

In synthetic geometry, congruence of segments is a postulated relation (for which we use the symbol "$\cong$") on the set of all segments. (I.e., for each two segments, either they are congruent or they are not congruent: this is analogous to the assumption of an exact physical procedure for the determination of length-equivalence or non-equivalence for rods.) Segment is defined in the usual way, in terms of the concept of "betweenness". In synthetic geometry, betweenness is a relation on ordered triples of points. We adopt the simple notation $A - B - C$ (read as "B is between A and C"). Thus for each ordered triple of points (X,Y,Z), it is postulated that either $X - Y - Z$ or (not $X - Y - Z$). It should be noted that Euclid did not formulate the idea of betweenness explicitly, but he definitely made use of it implicitly, and used the following betweenness postulates without explicitly stating them:

B-1 If $X - Y - Z$ then X, Y, Z are different collinear points.

B-2 Given three different collinear points, exactly one is between the other two.

B-3 The relation is symmetric in the sense that $X - Y - Z$ if and only if $Z - Y - X$.

B-4 If X and Y are any two points, then there are points Z, W, such that $X - Z - Y$, and $W - X - Y$.

B-5 Any four collinear points can be named in an order X_1, X_2, X_3, X_4 such that $X_1 - X_2 - X_3 - X_4$. (This is a kind of transitivity condition; the notation means that all of the four relations $X_1 - X_2 - X_3$, $X_1 - X_2 - X_4$, $X_1 - X_3 - X_4$, $X_2 - X_3 - X_4$, hold.)

The congruence postulates for segments are:

CS-1 Congruence is an equivalence relation on the set D of all segments.

CS-2 Given a segment $\overline{AB}$ and a ray $\overrightarrow{XY}$, there is a unique point $Z \in \overrightarrow{XY}$ such that $\overline{AB} \cong \overline{XZ}$.

Congruence and betweenness are related in the postulate:

CB-1 If $A_1 - B_1 - C_1$, $A_2 - B_2 - C_2$, and $\overline{A_1B_1} \cong \overline{A_2B_2}$, then $\overline{B_1C_1} \cong \overline{B_2C_2}$ if and only if $\overline{A_1C_1} \cong \overline{A_2C_2}$.

As before, we define a length function for the set D of all segments in space, to be a function

$$\lambda : D \to R^+$$

which has the properties

(i) if d_1, d_2, are segments, and $d_1 \cong d_2$, then $\lambda(d_1) = \lambda(d_2)$;

(ii) if $X - Y - Z$, then $\lambda(\overline{XY}) + \lambda(\overline{YZ}) = \lambda(\overline{XZ})$.

We are now ready to establish a suitable "length structure" in D, (the set of all segments) and in $\tilde{D}$ (the set of congruence classes of segments). Congruence is the equivalence relation which corresponds to the idea "same length". A notion of order is quite simply introduced: we define $\overline{AB} < \overline{XY}$ if there exists Z with $X - Z - Y$ and $\overline{AB} \cong \overline{XZ}$. Using the congruence and betweenness postulates, it is straightforward (see exercises) to prove the following:

(i) The relation $<$ is transitive.

(ii) The relation $<$ is "preserved" under congruence, and hence yields a corresponding relation (for which we use the same symbol) on $\tilde{D}$.

(iii) For any two segments $\overline{AB}$, $\overline{XY}$, exactly one of the following holds, $\overline{AB} < \overline{XY}$; $\overline{AB} \cong \overline{XY}$; $\overline{XY} < \overline{AB}$; hence trichotomy holds for $<$ in $\tilde{D}$, so that $<$ is a strict total order relation in $\tilde{D}$.

We introduce next a join operation $(*)$ in D, and on $\tilde{D}$. If $A - B - C$, we define $\overline{AB} * \overline{BC} = \overline{AC}$. We could develop some properties of $*$ in D, but by now you should see that it will be simplest to go directly to the set of congruence classes, $\tilde{D}$. Given $\tilde{d}_1$, $\tilde{d}_2 \in \tilde{D}$, let $\overline{A_1B_1} \in \tilde{d}_1$, $\overline{A_2B_2} \in \tilde{d}_2$. Then, from the postulates, we can find C_1 such that $A_1 - B_1 - C_1$.

and $\overline{B_1C_1} \cong \overline{A_2B_2}$. We define the join of $\tilde{d}_1$ and $\tilde{d}_2$ to be $\widetilde{\overline{A_1C_1}}$, and denote this join by $\tilde{d}_1 * \tilde{d}_2$.

We must first verify (see exercises)

(i) that this yields a (single-valued) operation in $\tilde{D}$; (i.e., we must prove that the congruence class $\widetilde{\overline{A_1C_1}}$ is unchanged for different choices of segments from $\tilde{d}_1$ and $\tilde{d}_2$).

We should next prove (see exercises below) that the join operation on $\tilde{D}$ has the properties:

(ii) it is associative;

(iii) it is commutative;

(iv) it preserves order, in the sense that

$$\widetilde{\overline{A_1B_1}} < \widetilde{\overline{A_2B_2}} \Longleftrightarrow \widetilde{\overline{A_1B_1}} * \widetilde{\overline{XY}} < \widetilde{\overline{A_2B_2}} * \widetilde{\overline{XY}}$$

for any segment $\overline{XY}$.

Exercises 2-5 (continued)

5. Prove the assertions (i) -- (iii) made above concerning the relation $<$ for segments.

6. Prove the assertions (i) -- (iv) made above concerning the properties of the join operation on the set of congruence classes of segments.

7. If $X_1 - X_2 - X_3 - X_4$, show that $X_4 - X_3 - X_2 - X_1$, and that no other order is possible.

8. Prove that the order relation $<$ in $\tilde{D}$ is dense, and that $\tilde{D}$ contains no least element and no greatest element.

9. If m, n, are positive integers, and $n(\tilde{d})$ denotes the n-fold iterated join, prove that, for all m, n, and $\tilde{d}$,

(a) $(m + n)\tilde{d} = m\tilde{d} * n\tilde{d}$;

(b) $n(\tilde{d}_1 * \tilde{d}_2) = n\tilde{d}_1 * n\tilde{d}_2$;

(c) $(mn)\tilde{d} = m(n\tilde{d}) = n(m\tilde{d})$;

(d) $1\tilde{d} = \tilde{d}$.

10. Prove that $\tilde{d}_1 < \tilde{d}_2$, if and only if $n\tilde{d}_1 < n\tilde{d}_2$ for each positive integer n.

11. Prove that for positive integers m, n, and $\tilde{d} \in \tilde{D}$,

$$m\tilde{d} < n\tilde{d} \text{ if and only if } m < n .$$

By now we have arrived at a structure, $(\tilde{D}, *, <)$, which is a densely-ordered, abelian, semigroup.

In other words, we have derived (from the postulates) for the set of congruence classes of segments, a structure which is similar to the structure developed empirically and inductively for the set of equivalence classes of rods in Section 2-4. It follows that if we ask the questions:

(i) Are there any structure-preserving mappings from $(\tilde{D}, *, <)$ to $(R^+, +, <)$?

(ii) If there is more than one such map, how are they related? ,

then the answer to question (ii) will be exactly as before. That is, if there exists a structure-preserving map, λ, from $(\tilde{D}, *, <)$ to $(R^+, +, <)$, then there exist infinitely many; and if λ is onto, then these are all related by composition with positive similarities. Moreover, if Λ denotes the set of all such structure-preserving mappings, the group $\underline{S}^+$, of positive similarity transformations of R^+, is isomorphic to the group of those automorphisms of Λ which preserve its algebraic structure as an ordered abelian semigroup, $(\Lambda, +, <)$, of functions.

The structure-preserving functions (if any) in Λ, will, of course, yield length functions for segments; and, conversely, any length function will correspond to such a structure-preserving function. If λ is any one of these, we may establish a distance function α, as indicated earlier. From this, if λ is onto R^+, it is not too difficult to set up a suitable coordinate function on each line ℓ, by the following procedure:

(i) Let A, B, $C \in \ell$, with $B - A - C$.

(ii) If X is any point on the ray $\overrightarrow{AB}$ and Y any point on the opposite ray $\overrightarrow{AC}$, define $f : \ell \to R$ by

$$\begin{cases} f(A) = 0 \\ f(X) = \alpha(A,X) = \lambda(\overline{AX}) \\ f(Y) = -\alpha(A,Y) = -\lambda(\overline{AY}) . \end{cases}$$

In order that f be a suitable coordinate function as postulated in the metric treatment of geometry, we must prove that f is onto R, and that for any two points P, Q, $\in \ell$,

$$\alpha(P,Q) = |f(P) - f(Q)| .$$

You should be able to prove these results for yourself (under the assumption, of course, that a length function $\lambda : \tilde{D} \rightarrow R^+$ exists, and that λ is onto).

Existence of Length Functions. The procedures which we might follow, to set up a suitable length function on $\tilde{D}$, are quite similar to those used in the last section to establish such a function empirically on the set of equivalence classes of rods. However, whereas in the discussion of the physical measurement of length we could appeal to empirical evidence to support the assumptions which were needed, here we must stay within the limitations of the axioms of synthetic geometry, and we must prove both the existence of the objects used in our procedure (e.g., integral multiples and submultiples of a given segment) and the correctness of our result.

Method I.

If we consider the mathematical counterpart of our empirical "Method I", we see that we must first select, as a unit, an equivalence class $\tilde{d}_0$ of segments. Let $\overline{A_0B_0}$ be any segment in $\tilde{d}_0$. For any other segment $\overline{PQ}$, we can parallel our empirical process, using the congruence and betweenness postulates, to establish the existence of points $X_1, X_2, \cdots, X_n, \ldots$ on the ray $\overrightarrow{PQ}$, such that $P - X_1 - X_2 - \ldots - X_n - \ldots$, and with $\overline{PX_1} \cong \overline{X_1X_2} \cong \ldots \cong \overline{A_0B_0}$. But, in order to assert that there will exist a positive integer n_0 such that either $X_{n_0} = Q$, or $P - X_{n_0} - Q - X_{(n_0+1)}$, we must add the archimedean postulate to the axioms of classical synthetic geometry. (Remember that the archimedean postulate is equivalent to the statement that, given any two segments $\overline{L_1L_2}$, $\overline{M_1M_2}$, there exists a positive integer n such that $n(\overline{L_1L_2}) > \overline{M_1M_2}$.)

The next step requires the existence of a segment $\overline{A_1B_1}$, such that for some positive integer m, $(m > 1)$

$$m(\widetilde{\overline{A_1B_1}}) = \widetilde{\overline{A_0B_0}} .$$

For any integer $m \geq 2$, we can show the existence of such a "submultiple" of $\overline{A_0B_0}$, by using "parallel projection", as in the well-known construction for subdividing a segment into m congruent parts. If we take $m = 10$, we can continue to mimic our earlier empirical procedure, to establish the existence of successive digits in a decimal number $r = n_0.n_1n_2 \dots$. We then define a function, λ , by

$$\lambda(\overline{PQ}) = r .$$

This procedure is relatively simple to describe, but it is awkward to verify that the "constructed" function, λ , has the desired property of mapping the ordered semigroup $(\tilde{D}, *, <)$ isomorphically into $(R_0^+, +, <)$: If you have ever tried to carry through a development of the real number system, based on decimal "sequences", you will be quite familiar with this particular difficulty!. If you think about it, you will see that it is relatively straightforward to prove that λ gives the same value on congruent segments, and hence carries over to congruence classes, and that λ preserves inequalities. The difficulty comes in proving that λ is additive; i.e., that λ carries joins in $\tilde{D}$ into sums in R^+ : if you attempt to construct such a proof, you will find that you get involved in all the awkwardness of adding decimal fractions from left to right. No doubt a proof could be carried through, but we shall not attempt it because it is simpler, and more instructive, to carry through Method II below in considerable detail.

We remark that the axioms of synthetic geometry, augmented by the archimedean postulate, do not permit us to prove that λ is onto. The best we could do in this direction is to show that the range of λ must contain at least the positive surd numbers. We look at this question again after discussing the second method, merely noting here that the demonstration of the existence of integral multiples and integral submultiples can be extended, to show the existence of all positive rational "multiples" of any segment; and thus, in particular, the existence of arbitrarily "small" and arbitrarily "large" segments.

Method II

In discussing the physical procedure for establishing length functions, there did not appear to be any strong reason for preferring either method. But in the context of synthetic geometry, if we really intend to fill in all of the details the second method has many advantages. As in the physical situation, we can establish a "theory of ratios" for congruence classes of segments. This theory is quite independent of any question of the selection of "units".

We first introduce an "equal-ratio" relation on the set of ordered pairs of congruence classes (i.e., in $\tilde{D} \times \tilde{D}$) . We define the relation, ~ , by

$$(\tilde{d}_1,\tilde{d}_2) \sim (\tilde{d}_3,\tilde{d}_4)$$

if, for positive integers m , n , $m\tilde{d}_2 < n\tilde{d}_1 \Longleftrightarrow m\tilde{d}_4 < n\tilde{d}_3$. It is not difficult to prove that this relation is an equivalence relation: i.e., it is symmetric, reflexive, and transitive. We could (as Euclid did) go ahead and define inequalities between classes of equivalent ratios, and prove many properties, all without the assumption of an archimedean postulate. But these proofs are quite tedious, and this would distract us from our main objective: the construction of a suitable length function on $\tilde{D}$.

We observe that the procedure for defining the above equivalence relation on $\tilde{D} \times \tilde{D}$, suggests a procedure for defining an (absolute) real valued function on the set K of equivalence classes in $\tilde{D} \times \tilde{D}$. For any such class, ρ , and any $(\tilde{d}_1,\tilde{d}_2)$ in ρ , let $C = \{\frac{m}{n} : m , n$ positive integers, and $m\tilde{d}_2 < n\tilde{d}_1\}$. We observe first that if any fraction $\frac{m}{n} \in C$, then all equivalent fractions belong to C (Exercise 10 above) so that we can consider C to be a set of rational numbers. We see immediately, from the definition of the equivalence relation ~ for ratios, that C does not depend on the choice of $(\tilde{d}_1,\tilde{d}_2)$ in ρ . Moreover it appears that (possibly with additional assumptions) the set of rational numbers C might be a cut. If so, it determines a real number r , and we could go ahead and define our desired function by $\varphi : \rho \rightarrow r$. We are now at a critical point in the discussion, so we slow down and fill in some of the details. We draw on the results of Exercises 2-5 (5, 6, 9, 10, 11) above, whose proofs are fairly straightforward.

Theorem 2-5.1. If, in addition to the usual postulates of synthetic geometry, we assume the archimedean postulate, then the set C of rational numbers (defined above) is a cut.

Proof. We must first show that every positive rational belongs either to C , or to its complement in Q^+ , and that neither C nor its complement is empty. That every positive rational belongs to C or to its complement follows from trichotomy in $\tilde{D}$. From the archimedean postulate, there exist positive integers p , q , such that

$$\tilde{d}_2 < p\tilde{d}_1 \; ; \; \tilde{d}_1 < q\tilde{d}_2 \; .$$

Hence $\frac{1}{p} \in C$ and $q \notin C$, so that neither C nor its complement is empty.

We must next show that if $\frac{m_1}{n_1} \in C$ and $\frac{m_2}{n_2} < \frac{m_1}{n_1}$, then $\frac{m_2}{n_2} \in C$. If $\frac{m_2}{n_2} < \frac{m_1}{n_1}$ then $m_2n_1 < n_2m_1$. Hence

$$m_2n_1\tilde{d}_2 < n_2m_1\tilde{d}_2 \qquad \text{(Exercise 11 above)}$$

$$< n_2n_1\tilde{d}_1 \qquad \text{(Definition of } C) .$$

Hence $$m_2\tilde{d}_2 < n_2\tilde{d}_1 \qquad \text{(Exercise 10 above)}$$

and therefore $$\frac{m_2}{n_2} \in C \qquad \text{(Definition of } C) .$$

Finally, we have to show that, if $\frac{m_1}{n_1} \in C$, then there exists $\frac{m_2}{n_2} \in C$, such that $\frac{m_2}{n_2} > \frac{m_1}{n_1}$. Since $\frac{m_1}{n_1} \in C$, we have $m_1\tilde{d}_2 < n_1\tilde{d}_1$. Hence, from the definitions of $*$ and $<$, there exists $\tilde{d}_3$ such that

$$n_1\tilde{d}_1 = m_1\tilde{d}_2 * \tilde{d}_3 .$$

From the archimedean postulate, there exists a positive integer t such that $t\tilde{d}_3 > \tilde{d}_2$. Hence

$$tn_1\tilde{d}_1 = tm_1\tilde{d}_2 * t\tilde{d}_3 \qquad \text{(Exercise 9 above)}$$

$$> tm_1\tilde{d}_2 * \tilde{d}_2 \qquad \text{(Exercise 6 above)}$$

$$= (tm_1 + 1)\tilde{d}_2 \qquad \text{(Exercise 9 above)}.$$

Hence $$\frac{tm_1 + 1}{tn_1} \in C \text{ , and } \frac{tm_1 + 1}{tn_1} > \frac{tm_1}{tn_1} = \frac{m_1}{n_1} .$$

This completes the proof of the theorem. You should note how much we depended on the archimedean postulate.

We now define the function

$$\varphi : K \to R^{+}$$

by $$\varphi(\rho) = r ,$$

where r is the real number determined by the cut C above. We shall use the same symbol, φ, to denote the corresponding function on $\tilde{D} \times \tilde{D}$ and we shall prove that φ is an additive ratio operation, as defined in the last section.

<u>Theorem 2-5.2</u>. The function φ has the following properties: for any $\tilde{d}_0 , \tilde{d}_1 , \tilde{d}_2 \in \tilde{D}$,

(i) If $\tilde{d}_1 < \tilde{d}_2$, then $\varphi(\tilde{d}_1,\tilde{d}_0) < \varphi(\tilde{d}_2,\tilde{d}_0)$;

(ii) $\varphi(\tilde{d}_1 * \tilde{d}_2,\tilde{d}_0) = \varphi(\tilde{d}_1,\tilde{d}_0) + \varphi(\tilde{d}_2,\tilde{d}_0)$;

(iii) $\varphi(\tilde{d}_2,\tilde{d}_1) \cdot \varphi(\tilde{d}_1,\tilde{d}_0) = \varphi(\tilde{d}_2,\tilde{d}_0)$

<u>Proof</u>. In view of the fact that φ is defined by using cuts, you will not be surprised to discover that the proof of this theorem makes full use of the properties of cuts, and that it looks like many of the proofs found in elementary analysis.

<u>Proof of (i)</u>. The monotone property, (i), is most easily proved as a consequence of (ii). Assume that (ii) has been proved, and that $\tilde{d}_1 < \tilde{d}_2$. Then there exists $\tilde{d}_3$, with $\tilde{d}_1 * \tilde{d}_3 = \tilde{d}_2$. Hence, from the additive property (ii) of φ ,

$$\varphi(\tilde{d}_2,\tilde{d}_0) = \varphi(\tilde{d}_1 * \tilde{d}_3 , \tilde{d}_0) = \varphi(\tilde{d}_1,\tilde{d}_0) + \varphi(\tilde{d}_3,\tilde{d}_0)$$

so that $$\varphi(\tilde{d}_1,\tilde{d}_0) < \varphi(\tilde{d}_2,\tilde{d}_0) ,$$

as required.

<u>Proof of (ii)</u>. (Additive Property of φ .) Let $C_{10} = \{\frac{m}{n} : m\tilde{d}_0 < n\tilde{d}_1\}$, $C_{20} = \{\frac{m}{n} : m\tilde{d}_0 < n\tilde{d}_2\}$, and $C_{30} = \{\frac{m}{n} : m\tilde{d}_0 < n(\tilde{d}_1 * \tilde{d}_2)\}$.

Let $$\frac{m_1}{n_1} \in C_{10} , \frac{m_2}{n_2} \in C_{20} .$$

We shall show first that

$$\frac{m_1}{n_1} + \frac{m_2}{n_2} = \frac{m_1 n_2 + m_2 n_1}{n_1 n_2} \in C_{30} :$$

We have $$m_1 \tilde{d}_0 < n_1 \tilde{d}_1 , \text{ and } m_2 \tilde{d}_0 < n_2 \tilde{d}_2 .$$

Hence,
$$\begin{aligned}(m_1 n_2 + m_2 n_1)\tilde{d}_0 &= m_1 n_2 \tilde{d}_0 * m_2 n_1 \tilde{d}_0 \\ &< n_1 n_2 \tilde{d}_1 * n_2 n_1 \tilde{d}_2 \\ &= n_1 n_2 (\tilde{d}_1 * \tilde{d}_2) :\end{aligned}$$

Hence, $$\frac{m_1}{n_1} + \frac{m_2}{n_2} \in C_{30} ,$$

and therefore, from the definition of addition for cuts,

$$C_{10} + C_{20} \leq C_{30} .$$

We shall complete the proof of (ii) by showing that

$$C_{30} \leq C_{10} + C_{20} .$$

This is accomplished by proving that, if

$$\frac{m_1}{n_1} \notin C_{10} \text{ and } \frac{m_2}{n_2} \notin C_{20} ,$$

then $$\frac{m_1}{n_1} + \frac{m_2}{n_2} \notin C_{30} ,$$

so that $$C_{30} \not\leq C_{10} + C_{20} .$$

(Cf. Exercise 2-2.3.)

$$\begin{cases}\dfrac{m_1}{n_1} \notin C_{10} \text{ implies that } m_1 \tilde{d}_0 \geq n_1 \tilde{d}_1 ; \\ \dfrac{m_2}{n_2} \notin C_{20} \text{ implies that } m_2 \tilde{d}_0 \geq n_2 \tilde{d}_2 .\end{cases}$$

Hence
$$(m_1n_2 + m_2n_1)\tilde{d}_0 = m_1n_2\tilde{d}_0 * m_2n_1\tilde{d}_0$$
$$\geq n_1n_2\tilde{d}_1 * n_2n_1\tilde{d}_2$$
$$= n_1n_2(\tilde{d}_1 * \tilde{d}_2) .$$

Therefore
$$\frac{m_1n_2 + m_2n_1}{n_1n_2} = \frac{m_1}{n_1} + \frac{m_2}{n_2} \notin C_{30} .$$

Hence $C_{10} + C_{20} = C_{30}$, and hence φ is additive.

Proof of (iii). (Cancellation Property.) With a similar notation to that used earlier, let
$$C_{21} = \{\tfrac{m}{n} : m\tilde{d}_1 < n\tilde{d}_2\}$$
$$C_{10} = \{\tfrac{m}{n} : m\tilde{d}_0 < n\tilde{d}_1\}$$
$$C_{20} = \{\tfrac{m}{n} : m\tilde{d}_0 < n\tilde{d}_2\} .$$

Then, if $\frac{p}{q} \in C_{21}$ and $\frac{r}{s} \in C_{10}$, we have $r\tilde{d}_0 < s\tilde{d}_1$, $p\tilde{d}_1 < q\tilde{d}_2$, and hence $pr\tilde{d}_0 < ps\tilde{d}_1 < qs\tilde{d}_2$. Hence $\frac{pr}{qs}$ $(= \frac{p}{q} \cdot \frac{r}{s}) \in C_{20}$. Hence, from the definition of multiplication for cuts,
$$C_{21} \cdot C_{10} \leq C_{20} .$$

We shall show equality, by proving the opposite inequality; from Exercise 2-2.4, this will be accomplished if we show that, if $\frac{p}{q} \notin C_{21}$ and $\frac{r}{s} \notin C_{10}$, then $\frac{p}{q} \cdot \frac{r}{s} \notin C_{20}$.

$$\frac{p}{q} \notin C_{21} \text{ implies that } p\tilde{d}_1 \geq q\tilde{d}_2$$
$$\frac{r}{s} \notin C_{10} \text{ implies that } r\tilde{d}_0 \geq s\tilde{d}_1 .$$

Hence
$$pr\tilde{d}_0 \geq ps\tilde{d}_1 \geq qs\tilde{d}_2$$

so that
$$\frac{pr}{qs} \left(= \frac{p}{q} \cdot \frac{r}{s}\right) \notin C_{20} .$$

It follows that $C_{20} = C_{21} \cdot C_{10}$, so that the cancellation property holds, and φ is an additive ratio operation for $\tilde{D}$.

Comment. If you worked through the above proof in detail, you might have felt that the proof would have been simpler if we had had available positive rational "multiples" of segment classes, instead of just integral multiples. It is a fact, as we suggested earlier, that an "algebra" of such positive rational multiples can be developed, and that it can be used in constructing a proof of the above theorem. (The existence of all real multiples cannot be shown unless we make further assumptions: this question of "completeness" is discussed later.) If you want to see how the alternative proof goes, you should work the following exercises:

Exercises 2-5 (continued)

12. By analogy with the classical construction procedures of geometry, prove that, for each positive integer n, and each congruence class $\tilde{d}_1$ of segments, there exists a unique congruence class $\tilde{d}_2$ such that

$$n\tilde{d}_2 = \tilde{d}_1 .$$

13. Define the $\tilde{d}_2$ found in Exercise 12 to be $\frac{1}{n}\tilde{d}_1$, and hence define a congruence class $\frac{m}{n}\tilde{d}_1$, for each fraction $\frac{m}{n}$. Prove that $\frac{m}{n}\tilde{d}_1$ is the same for all equivalent fractions, so that we get a "product" $q\tilde{d}_1$ for each positive rational number q. Prove that this "product" (or "scalar multiplication") satisfies:

(i) $$1\tilde{d} = \tilde{d} ;$$

(ii) $$(q_1 + q_2)\tilde{d} = q_1\tilde{d} * q_2\tilde{d} ;$$

(iii) $$q(\tilde{d}_1 * \tilde{d}_2) = q\tilde{d}_1 * q\tilde{d}_2 ;$$

(iv) $$(q_1q_2)\tilde{d} = q_1(q_2\tilde{d}) ;$$

(v) for every $\tilde{d}$, $q_1 > q_2$ if and only if $q_1(\tilde{d}) > q_2(\tilde{d})$;

(vi) for every positive rational q, $\tilde{d}_1 > \tilde{d}_2$ if and only if $q(\tilde{d}_1) > q(\tilde{d}_2)$.

14. Use the result of Exercise 13 to construct an alternative proof to Theorem 2-5.2.

This is about as far as we can go towards setting up a length function without making some arbitrary choices. Usually, at this stage, one selects one of the equivalence classes (say $\tilde{d}_0$) and decides that this shall be the unit; i.e., we decide, quite arbitrarily, that $\lambda_0(\tilde{d}_0) = 1$, where λ_0 is to be our length function. In this way we get one length function for each choice of $\tilde{d}_0$. Actually, instead of using the value "1" , we could define $\lambda_0(\tilde{d}_0)$ to be any positive real number, say k_0 . Your reaction to this is likely to be that we are being pedantic -- that we could quickly compose such a λ_0 with the similarity transformation $\frac{1}{k_0}$, and get another length function which would take $\tilde{d}_0$ into the number "1" ; and the equivalence class corresponding to $\frac{1}{k_0} \cdot \tilde{d}_0$ would, in fact, be the unit for λ_0 , so why not start with this unit to begin with? The difficulty is that we do not know that there is always an equivalence class corresponding to $\frac{1}{k_0} \tilde{d}_0$; so far we have only shown the existence of rational multiples, and k_0 need not be rational. In fact, as mentioned earlier, there are geometries (such as cartesian surd geometry, in which all coordinates must be surds) which satisfy all of the postulates of synthetic geometry, and for which there are length functions which do not map any equivalence class into the number "1" -- i.e., in terms of the usual usage of the word "unit", they have no unit. This suggests that there might be something significant in the observation that we could be more general by mapping a selected equivalence class $\tilde{d}_0$ into a selected real number k_0 .

The difficulty, of course, goes back to the properties of the ratio operation φ which we have constructed: the properties which we have proved for φ will enable us to construct, (assigning values in R^+ to $\tilde{d}_0$) as many length functions, "based" on $\tilde{d}_0$, as there are elements of R^+ ; but, unless (for fixed $\tilde{d}_0$) φ maps the set of all ordered pairs $(\tilde{d}, \tilde{d}_0)$ onto R^+ , not all of these functions will have units!

We now define

(*) $$\lambda_0(\tilde{d}_0) = k_0 \quad ; \quad \lambda_0(\tilde{d}) = [\varphi(\tilde{d}, \tilde{d}_0)]k_0 .$$

The first part of the following theorem is a simple consequence of our earlier theorem concerning the ratio operation φ . The second part is an exercise in the properties of isomorphisms and cuts.

Theorem 2-5.3.

(i) The function λ_0 is an isomorphism of $(\tilde{D}, *, <)$ into $(R^+, +, <)$, and determines a corresponding length function

$$\lambda_0 : D \to R^+ .$$

(ii) If λ_1, λ_2, are any two isomorphisms (not necessarily obtained from φ) of $(\tilde{D}, *, <)$ into $(R^+, +, <)$, which agree on any element of $\tilde{D}$, then $\lambda_1 = \lambda_2$. Equivalently, two length functions which agree on any segment, are, in fact, the same function.

Proof. The proof of part (i) is quite straightforward, and we leave it to you. The proof of part (ii) is very similar to that of Theorem 2-2.3, so we sketch it only: first, show that every length function preserves order. Then show that, because of the finite additive property for length functions, if $\lambda_1(\tilde{d}_0) = \lambda_2(\tilde{d}_0)$, then $\lambda_1(q\tilde{d}_0) = q\lambda_1(\tilde{d}_0) = q\lambda_2(\tilde{d}_0) = \lambda_2(q\tilde{d}_0)$ for every positive rational q. Then show that if $\tilde{d}$ is any segment class which is not equal to $q\tilde{d}_0$ for any positive rational q, then the monotone character of length functions implies that

$$\lambda_1(\tilde{d}) = [\varphi(d,d_0)][\lambda_1(\tilde{d}_0)] = [\varphi(d,d_0)][\lambda_2(\tilde{d}_0)] = \lambda_2(\tilde{d}) .$$

Corollary 1. There are no other length functions than those functions λ_0 which are obtained, as described above (definition *) from the ratio function φ, by selecting an element $\tilde{d}_0$ (of $\tilde{D}$) and a value $k_0 = \lambda_0(\tilde{d}_0)$.

Proof. Let λ be any length function, not necessarily obtained from φ as above. Let $\tilde{d}_0 \in \tilde{D}$, and let k_0 be the value of λ at $\tilde{d}_0$. Let λ_0 be the length function obtained (by definition *) from the ratio operation φ, and such that $\lambda_0(\tilde{d}_0) = k_0$. Then, from the theorem, $\lambda = \lambda_0$.

Corollary 2. Any length function λ may be expressed as the composite of any length function which has a unit, and a suitable positive similarity transformation of R^+. In particular, every two length functions are similar.

Proof. Let $\tilde{d}_0 \in \tilde{D}$, and let $\lambda(\tilde{d}_0) = k_0$. Let λ_0 be the length function for which $\tilde{d}_0$ is the unit. Then it is easy to prove that $\lambda = k_0\lambda_0$. Hence every length function is similar to λ_0, and hence, because similarity of functions is an equivalence relation, every two length functions are similar.

Remark: You should note that we have now proved that, in a "synthetic geometry" which satisfies the archimedean postulate, every two length functions are similar, and any function which is similar to a length function is a length function. That is, the set of all length functions (for segments) is a ratio scale, whether or not each function is onto R^+ .

We look next at the structural properties of the "restricted" set of those length functions for which there is a unit. (I.e., the set Λ_1 of those length functions whose range includes the number "1" .) The relationship between functions in Λ_1 is given by part (iii) of Theorem 2-5.2. For if $\lambda_0 , \lambda_1 \in \Lambda_1$; and $\lambda_0(\tilde{d}_0)$ [illegible] then for any $\tilde{d} \in \tilde{D}$, by the cancellation property

$$\lambda_1(\tilde{d}) = \lambda_0(\tilde{d}) \cdot \lambda_1(\tilde{d}_0) .$$

In other words, λ_1 is the composite of λ_0 with the positive similarity determined by the number $\lambda_1(\tilde{d}_0)$: the set of all such numbers, for $\lambda_1 \in \Lambda_1$, is not necessarily R^+ , hence, although the functions in Λ_1 are all similar, they do not necessarily constitute a ratio scale.

We state the following theorem, and leave the proof to you.

Theorem 2-5.4.

(i) The range of the function φ is a multiplicative subgroup, Φ^+ , of R^+ ;

(ii) Φ^+ includes the positive surds (and hence, of course, the positive rationals);

(iii) the range of each length function which has a unit, coincides with the range of φ ; i.e., it is the group Φ^+ ;
[Hint: You must show that $\{\varphi(\tilde{d},\tilde{d}_0) : \tilde{d}_0$ fixed, $\tilde{d} \in \tilde{D}\} = \{\varphi(\tilde{d}_1,\tilde{d}_2) : \tilde{d}_1 , \tilde{d}_2 \in \tilde{D}\}$. Because $\varphi(\tilde{d}_1,\tilde{d}_2) = \varphi(\tilde{d}_1,\tilde{d}_0) \cdot \varphi(\tilde{d}_0,\tilde{d}_2)$, you can complete the proof by showing that there exist segments d' , d'' , such that $d_0 : d_2 = d' : d_0$, and $d_1 : d_0 = d'' : d'$. This, of course, follows from well-known geometric constructions.]

(iv) if $\lambda_0(\tilde{d}_0) = \lambda_1(\tilde{d}_1) = 1$, then $\lambda_0(\tilde{d}_1) = \dfrac{1}{\lambda_1(\tilde{d}_0)}$; and

$$\lambda_0(\tilde{d}_1) = \frac{\lambda_0(\tilde{d})}{\lambda_1(\tilde{d})} , \text{ for every } \tilde{d} .$$

The Range of the Length Function. As stated earlier, cartesian geometry with surd coordinates satisfies the axioms of classical synthetic geometry, and also the archimedean postulate. Therefore, the whole of the above theory is applicable to cartesian geometry with surd coordinates. In this case it is easily shown that the range, Φ^+, for the corresponding ratio function φ, is the set of all positive surds. That is we see, from this example, that it is not necessary for the ratio function (for a geometry which satisfies the archimedean postulate in addition to the usual postulates of synthetic geometry) to be onto R^+.

If, for cartesian geometry with surd coordinates, we were to set up our ratio function φ, and then move to a length function λ by defining $\lambda(\tilde{d}_0) = \pi$ for a selected congruence class $\tilde{d}_0$, then all length values would be surd multiples of π. It can be shown that these numbers are all transcendental, and hence that this λ would not (in the ordinary sense) have a unit. But it would definitely be a length function.

For any particular length function λ_0, derived from a unit $\tilde{d}_0$, the range of λ_0 is just the range of the ratiofunction φ restricted to the subset K_0 of ordered pairs $(\tilde{d},\tilde{d}_0)$, with $\tilde{d}_0$ fixed. From Theorem 2-5.4, this is the same as the range Φ^+ of φ. We can ensure that this range is R^+, only by adding an additional "completeness" postulate. We could put this in the form: $\varphi(K) = \Phi^+ = R^+$, but this would be a very strange postulate for synthetic geometry, involving explicitly, as it does, the set of positive real numbers. The following is a suitable geometric form for a completeness postulate which will ensure that $\Phi^+ = R^+$:

Cantor-Dedekind Completeness Postulate. If $\{d_n\}$ is any sequence of segments such that $d_{n+1} \subset d_n$ for every n, then $\bigcap_n d_n \neq \emptyset$. (The notation $\bigcap_n d_n$ means the intersection of all of the sets d_n.)

(Moise calls a synthetic geometry with this additional property "complete in the sense of Dedekind"; other writers refer to it as the Cantor postulate.)

It is not too difficult (using properties of cuts and our previous discussion concerning the existence of all positive rational "scalar multiples" of any segment) to show that the assumption of the archimedean postulate, and the Cantor-Dedekind postulate, implies the existence of all real positive "scalar multiples" of any segment. (See exercises below.) It will follow immediately that $\Phi^+ = R^+$, and hence that every admissible length function is onto R^+. If we take any one of these functions, and use it (as indicated earlier) to set up a coordinate system for every line, then it will follow that every such coordinate function will be onto R. That is, we will obtain a distance function for $\mathcal{S}$, and a coordinate function for each line, which satisfy the relevant postulates of the metric treatment of geometry.

It can also be shown that any synthetic geometry which satisfies the archimedean postulate, and the Cantor-Dedekind completeness postulate, is isomorphic to the metric geometry of Birkhoff, and to real cartesian geometry. This question is discussed in [14].

Exercises 2-5 (continued)

15. Assume the archimedean postulate and the Cantor-Dedekind postulate. Let $\overline{AB}$ be any segment, and let λ be the length function for which $\overline{AB}$ is the unit. Then (Exercises 12, 13 above) corresponding to every positive rational number q, there is a unique point $Q \in \overrightarrow{AB}$, such that $\lambda(\overline{AQ}) = q$. Let r be any positive real number. Then it is a simple property of the real numbers, that there exist sequences of rational numbers (x_i), (y_i) such that, for each i, $x_i < x_{i+1} \leq r$; $y_i > y_{i+1} > r$; and $y_i - x_i < \frac{1}{i}$. If (X_i), (Y_i) are corresponding sequences of points of $\overrightarrow{AB}$, with $\lambda(\overline{AX_i}) = x_i$, $\lambda(\overline{AY_i}) = y_i$, prove that

    (a) $\overline{X_{i+1}Y_{i+1}} \subset \overline{X_iY_i}$ for each i.

    (b) $\bigcap_i \overline{X_iY_i}$ is a single point, P.

    (c) $\lambda(\overline{AP}) = r$.

    (d) λ is onto R^+.

16. Prove Theorem 2-5.4(i), that Φ^+ is always (i.e., without the assumption of completeness) a multiplicative subgroup of R^+.

17. Prove that Φ^+ is always closed under addition, and therefore an (additive) semigroup.

18. Prove Theorem 2-5.4(ii), that Φ^+ always contains the positive surds.

19. We can use the ratio operation $\varphi: \tilde{D} \times \tilde{D} \to R^+$, and the result of Theorem 2-5.4(iii) to define a "scalar multiplication" of elements of $\tilde{D}$ by elements of the "semi-field" Φ^+, as follows: if $\varphi(\tilde{d}_1, \tilde{d}_2) = r$, define $r\tilde{d}_2 = \tilde{d}_1$.

(a) Show that this scalar multiplication is properly defined (i.e., exists, and is single valued) for every $r \in \Phi^+$, and every $\tilde{d} \in \tilde{D}$, and satisfies

(i) $1\tilde{d} = \tilde{d}$;

(ii) $(r_1 + r_2)\tilde{d} = r_1\tilde{d} * r_2\tilde{d}$;

(iii) $r(\tilde{d}_1 * \tilde{d}_2) = r\tilde{d}_1 * r\tilde{d}_2$;

(iv) $(r_1 r_2)\tilde{d} = r_1(r_2\tilde{d})$;

(v) for every $\tilde{d} \in \tilde{D}$, $r_1 > r_2$ if and only if $r_1\tilde{d} > r_2\tilde{d}$;

(vi) for every $r \in \Phi^+$, $\tilde{d}_1 > \tilde{d}_2$ if and only if $r\tilde{d}_1 > r\tilde{d}_2$.

(b) If the Cantor-Dedekind completeness postulate holds, then $\Phi^+ = R^+$, and the scalar multiplication is defined for all elements of R^+.

Scalar Multiplication in The Domain, and Common Scientific Language. We point out that the scalar multiplication is only defined, as a single valued operation, for equivalence classes in $\tilde{D}$. However it is frequently useful to represent equivalence classes by particular elements from those classes, and write, for example, $d_1 = rd_2$, when $\varphi(\tilde{d}_1, \tilde{d}_2) = r$. We do this implicitly in such everyday language as "this stick is three times as long as that one". The existence (by assumption, or proof) of a scalar multiplication (by positive real numbers) for the common physical measures, also underlies such everyday usage as:

$$7 \text{ ft.} + 4 \text{ ft.} = (7 + 4)\text{ft.} = 11 \text{ ft.}$$

For if "ft." is just another name for a "unit" class (say $\tilde{d}$) , this is just the scalar multiplication property

$$7\tilde{d} + 4\tilde{d} = (7 + 4)\tilde{d} = 11\tilde{d} .$$

Other common statements which you can readily interpret in terms of this scalar multiplication, are:

$$1 \text{ ft.} = 12 \text{ ins.};$$
$$8 \text{ ft.} = 8\ (12 \text{ ins.}) = 96 \text{ ins.};$$
$$3(7 \text{ ft. } 1 \text{ in.}) = 21 \text{ ft. } 3 \text{ ins.}$$

But we are not yet ready to interpret the "multiplication" represented by the statement:

$$3 \text{ ft.} \times 4 \text{ ft.} = 12 \text{ ft.}^2 .$$

We have dealt with the question of length in synthetic geometry in considerable detail, because of its inherent interest, because there are some facts involved (particularly questions of range) which do not seem to be well known, and because the amount of work needed to set up mathematically-based length and coordinate functions from the axioms of synthetic geometry, is not generally appreciated. You can find an alternate treatment in [14] , which uses a least upper bound property instead of cuts. The total amount of work involved is about the same.

Links With Other Parts of Mathematics. We have shown that the domain of the length functions (i.e., the set of all segments) can be given a structure which leads to the ordered semigroup $(\tilde{D}, *, <)$. Because the order structure can be defined in terms of the join operation, we can regard the order structure as secondary, and concentrate our attention on the semigroup $(\tilde{D}, *)$. We also saw that a "scalar multiplication" could be defined in $\tilde{D}$, first by positive rational numbers, and, eventually, by any number in Φ^+ , the common range of all length functions. The set Φ^+ is a sort of "semi-field"; i.e., a group under multiplication and a semigroup under addition. The set Φ of real numbers which consists of the elements of Φ^+ , their negatives, and zero, is an ordered field. If the Cantor-Dedekind completeness postulate is assumed, then $\Phi^+ = R^+$, and $\Phi = R$.

If you are familiar with the notion of "a vector space over a field", the structure of $\tilde{D}$ should remind you of the structure of a vector space, except that

(i) $\tilde{D}$ is only a semigroup, and not a group, under the join operation;

(ii) the associated scalar system is not a field, but only the "semi-field" Φ^+ of all positive elements of the ordered field Φ .

A structure which is like a vector space, but whose associated scalar system is only a ring Z , is called a Z-module. This suggests that a suitable name for the structure of $\tilde{D}$, with respect to the join operation and the defined scalar multiplication, would be Φ^+-semimodule. If completeness is assumed, this is then an R^+-semimodule. We assume completeness in what follows.

For this vector-space-like, but more general, structure, the notion of linear functional can be defined in the usual way as a function into the associated scalar system, which preserves "addition", and scalar products; and we can easily show that the length functions are just the linear functionals for the R^+-semimodule $(\tilde{D}, *, R^+)$. The set of all such linear functionals (i.e., length functions) is the ratio scale Λ, and this may be given a "dual" structure in the same way as for a vector space: addition is just functional addition, and the scalar multiplication in Λ is just that which is defined for real-valued functions generally. In this way the system $(\Lambda, +, R^+)$ is also an R^+-semimodule, which is the conjugate, or dual, of $(\tilde{D}, *, R^+)$. This is probably the simplest example of the concept of dual space in relation to linear spaces. As every element of $\tilde{D}$ is the unit of some length function, we therefore see that the well-known "duality of functions and units" in relation to length measurement (and, of course, other ratio scales) is just a very simple example of the standard duality of linear algebra.

If Φ^+ is a proper subset of R^+, a scalar multiplication for $\tilde{D}$ by all positive real numbers is not defined; but we still regard length functions as additive positive-real-valued functions on $\tilde{D}$. In this case the set of all length functions is a kind of generalized dual to $(\tilde{D}, *, \Phi^+)$. In general, there will be an isomorphism of $(\tilde{D}, *, \Phi^+)$ into the dual space of $(\Lambda, +, R^+)$, and this will be onto if $\Phi^+ = R^+$; this isomorphism is defined just as in vector space theory: $\tilde{d} \rightarrow g$, where $g : \Lambda \rightarrow R^+$ is the linear functional such that $g(\lambda) = \lambda(\tilde{d})$.

Remark Concerning Units and Scales. Both in the empirical treatment of length and in the mathematical treatment, we ended up with a set of admissible length functions, any two of which are related by composition with a positive similarity transformation. Thus, if λ_1, λ_2 are admissible functions, there is a positive real number, k, such that $\lambda_2 = k\lambda_1$. Therefore, for any two elements x, y, in the common domain of λ_1, λ_2, we have

$$\frac{\lambda_2(x)}{\lambda_1(x)} = \frac{\lambda_2(y)}{\lambda_1(y)} = k$$

and therefore

$$\frac{\lambda_2(x)}{\lambda_2(y)} = \frac{\lambda_1(x)}{\lambda_1(y)}.$$

Conversely, if λ_1 and λ_2 are any two functions with a common domain and with positive real values, such that

$$\frac{\lambda_1(x)}{\lambda_1(y)} = \frac{\lambda_2(x)}{\lambda_2(y)}$$

for every pair of domain elements x and y ; then $\frac{\lambda_2(x)}{\lambda_1(x)} = \frac{\lambda_2(y)}{\lambda_1(y)} = k > 0$, so that $\lambda_2 = k\lambda_1$.

This almost trivial use of the simple fact that, for positive real numbers a , b , c , d , $\frac{a}{b} = \frac{c}{d}$ if and only if $\frac{a}{c} = \frac{b}{d}$, is important for all sets of measure functions which are related by positive similarities (i.e., which are ratio-scales, or similarity-invariant, in terms of the classification suggested in Chapter 1); it is this relationship (in either form) which is used in all simple calculations involving changes of units and scales.

You might find it useful to verbalize each of the two equivalent forms above:

(i) $\frac{\lambda_1(x)}{\lambda_1(y)} = \frac{\lambda_2(x)}{\lambda_2(y)}$, can be expressed as: the ratio of the length measures of any two objects, is the same for all length functions. (This number is, of course, the ratio, or "relative magnitude", of x and y .)

(ii) $\frac{\lambda_2(x)}{\lambda_1(x)} = \frac{\lambda_2(y)}{\lambda_1(y)}$, can be expressed as: the ratio of the length measures of an object under two given length functions, is the same for all objects. (This common ratio is, of course, the ratio of the two functions.)

The equivalence of these statements for similarity-invariant measures may be summarized as: measure functions (on the same domain) are related by positive similarities, if and only if they "preserve ratios". In other words, (as pointed out earlier) there is a natural 1-1 correspondence of ratio operations and ratio scales.

Ratio Operations, Ratio Scales, and Scalar Multiplication. We have seen above that all of these ideas enter into a theory of "length", and that there are many connections between them. The situation is similar for the other elementary "scalar" measures of the physical sciences, so it might be useful to spend a little time looking at the relationship of these ideas a little more closely.

In the framework of classical geometry (augmented by the archimedean postulate, and possibly by the Cantor-Dedekind postulate) we were able to introduce the basic ideas about "length", entirely in terms of the axioms of the geometry. You will recall that we defined a domain structure of equivalence (just the postulated congruence relation) and a binary operation of "joining", or "combining" domain elements (segments). This led to a semigroup structure on the set of equivalence classes of domain elements. Length functions were defined as those which, in effect, preserved this structure. In the process of proving the existence of length functions, we established an additive ratio operation (suitably related to equivalence) and constructed length functions from this ratio operation, by using "ratios" with respect to fixed domain elements as units. We were able to show that all such functions were similar; that any function similar to such a function was also a length function; and that any length function must be similar to such a function. Thus we proved that the set of all length functions is a ratio scale.

We also saw that we could work either directly from the ratio operation, or from the derived ratio scale, to define a "scalar multiplication" of domain classes by positive real numbers, and that the structure of the set of equivalence classes of domain elements was then an R^+-semimodule. (Or, if we did not assume completeness, a sort of "incomplete" R^+-semimodule, in the sense that scalar multiples were not always defined; but, where defined, we had the usual properties of a scalar multiplication.) When we now looked again at our length functions, we found that they had the additional property (not actually required by their definition) that they also "preserved" scalar multiplication, in the sense that $\lambda(rd) = r\lambda(d)$, whenever the scalar product rd was defined.

In the construction of a suitable measure theory for a particular physical "attribute", we do not usually have such a well-developed formal (axiomatic) system such as geometry, to use as a model space. The question then arises as to what sort of a theoretical picture we should assume as a result of empirical evidence. In practice it is useful to use all of the related ideas of ratio operations, ratio scales, and scalar multiplication, but our experience with "length" suggests that we do not need to assume all of these ideas directly as they are certainly not independent; if we assume some of them then we can define the others and exhibit their inter-relationship in a mathematical framework.

In any such approach, we usually assume that we have empirical procedures which correspond to the following assumptions;

(a) There exists an equivalence relation on the domain with respect to the attribute in question.

(b) There exists a binary operation of combining domain elements, which carries over to equivalence classes, and which gives the set of such classes a structure like an abelian semigroup, but not necessarily "complete". (There are empirical/philosophical problems concerning closure.)

(c) There exists an order relation, which also carries over to equivalence classes, and which is suitably related to the assumed binary operation.

(d) There exists a ratio operation which is properly related to the equivalence relation, which is additive, and which (because of the connection between "addition" and order) preserves order. (In the establishment of this ratio operation, it is highly likely that we shall have introduced a scalar multiplication by positive integers, and possibly also by positive rational numbers.)

At this point we have enough assumptions to define the relevant measure functions (as functions with values in R^+, which respect equivalence and which are "additive") and to prove their existence (by direct construction from the assumed ratio operation, exactly as for length functions). We can also prove, as for length, that any function which is similar to one obtained from the ratio operation, will be a suitable measure function, but we cannot prove the converse (that all suitable measure functions are similar) unless we make some further assumptions. One form of such an assumption (which would leave nothing to prove) would be to assume directly that there are no other suitable measure functions than the set of those functions which are similar to the functions which we obtained from the assumed ratio operation. That is, we would be assuming that the set of measure functions formed a ratio scale.

This conclusion could be reached in another way, by first introducing the notion of scalar multiplication into the domain, and then modifying the definition of our measure functions, to require that they "preserve" scalar multiplication as well as being additive. In this approach, the basic assumption is that of an additive ratio operation, from which the rest follows. We sketch the steps briefly, leaving you to fill in such details as have not been provided earlier in this chapter:

1. Assume that we have an additive ratio operation

$$\rho : A \times A \to R^+$$

for a set A, which has a binary operation of "combination", which we denote by "*".

2. Define (or assume) a relation "~" on A , such that $a_1 \sim a_2$ if and only if $\rho(a_1,a_2) = 1$;

3. Prove that

   (a) $\sim$ is an equivalence relation;

   (b) if $a_1 \sim a_1'$, $a_2 \sim a_2'$, then

$$\rho(a_1,a_2) = \rho(a_1', a_2') ;$$

   (c) if $a_1 \sim a_1'$, $a_2 \sim a_2'$, $a_3 \sim a_3'$, then

$$\rho(a_1 * a_2 , a_3) = \rho(a_1' * a_2' , a_3');$$

   [(b) and (c) imply that ρ determines a corresponding additive ratio operation on equivalence classes].

4. Introduce a scalar multiplication in the set $\tilde{A}$ as follows: if $\rho(a_1,a_2') = r$, define $r\tilde{a}_2 = \tilde{a}_1$. Then prove that, where defined, this scalar multiplication has the usual properties; i.e., (i)-(iv) of Exercise 2-5.19.

5. Define a suitable measure function $f : A \to R^+$ to be a function which not only preserves equivalence and is additive, but which also "preserves" the scalar multiplication in the sense that $f(r\tilde{a}) = rf(\tilde{a})$.

   [Here we use the same notation (f) for the function on A and for the derived function on $\tilde{A}$.]

6. As usual, prove that those functions which are derived directly from the ratio operation (by fixing a domain class as "unit") satisfy all of the requirements of 5 , and that these functions are all similar.

7. If now $f : A \to R^+$ is any function (not necessarily derived from the ratio operation) which "preserves" the scalar multiplication, then, for any $a \in A$, and fixed $a_0 \in A$, let $\rho(a,a_0) = r$. Then $a = ra_0$, and

$$f(a) = f(ra_0) = rf(a_0)$$

$$= f(a_0) \cdot f_0(a) ,$$

   where f_0 is the measure function for which a_0 is the unit, so that $r = \rho(a,a_0) = f_0(a)$. That is

$$f = [f(a_0)]f_0 .$$

   so that f is similar to f_0 . Thus there are no other measure functions in addition to those which belong to the unique ratio scale which is determined by the (similar) measure functions which have units. [As before, we can easily show that all of these functions are additive on $(\tilde{A}, *)$.]

We remark that the preservation of scalar multiplication thus shows up as a stronger requirement than the preservation of "addition". In the framework of classical geometry we were able to show (by rather tricky arguments involving the topological completeness of the real number system) that additivity implies that scalar multiplication is preserved; but where we do not have a formal framework in which to carry out such a proof, it is probably simplest to require the preservation of scalar multiplication as part of the definition.

An entirely equivalent approach (the difference is essentially linguistic) is to require the preservation of ratios: if $f_1, f_2 : \tilde{A} \to R^+$ are measure functions obtained from an additive ratio function, it is trivial to prove that for any a_1, a_2 in A,

$$\frac{f_1(a_1)}{f_1(a_2)} = \frac{f_2(a_1)}{f_2(a_2)} = a_1 : a_2$$

We could now require of every measure function, $g : A \to R^+$, that, in addition to preserving equivalence and being additive, it preserves ratios in the sense that

$$\frac{g(a_1)}{g(a_2)} = a_1 : a_2$$

From this assumption it is easy to prove that g must be similar to f_1 and f_2.

Remarks:

1. We emphasize that, in mathematical models relating to physical measurement, it is usually assumed (often implicitly) that a scalar multiplication of domain elements is defined for all $r \in R^+$. This is equivalent to the assumption that each measure function is onto R^+, and to the assumption that the domain is an R^+-semimodule.

2. In the previous section we commented that a ratiofunction was related to a scalar multiplication much as ordinary number division is related to number multiplication. We can now make this a little clearer. In the above discussion we showed how an additive ratio operation determined a scalar multiplication. On the other hand, if we have a notion of scalar multiplication

(by positive real numbers) in a set A, then we can define a ratio operation in (not necessarily on) A, by defining: $\rho(a_1,a_2) = r$ if $a_1 = ra_2$. If we assume that every element in A is some positive scalar multiple of every other element, then we can show that ρ is a ratio operation on A in terms of our definition of ratio operation. [You should prove that the cancellation property follows directly from the scalar multiplication property $r_1(r_2a) = (r_1r_2)a$.]

Length in Cartesian Geometry. We have already referred to real cartesian geometry. This is usually developed out of synthetic geometry or metric geometry, by proving that a 1-1 correspondence (a coordinate function) can be set up (in many related ways) between points of space and the cartesian product $R \times R \times R$. When developed from synthetic geometry, the range question is generally slurred over, but, it is usually assumed (implicitly) that a coordinate function is onto. This is equivalent to the assumption of the archimedean and Cantor-Dedekind postulates. This difficulty does not arise in the metric approach: the postulated coordinate functions for each line are onto R, and it follows easily that a coordinate function for space is onto $R \times R \times R$.

If a coordinate function for space is required to preserve distance (in terms of a postulated or constructed distance function for each line, and the usual distance function in $R \times R \times R$), then it can be shown to be unique up to a rigid motion of $R \times R \times R$.

Of course cartesian geometry can be established directly, without any appeal to metric or synthetic geometry, by defining points to be ordered triples of real numbers in the cartesian product $R \times R \times R$, with distance defined by

$$d(X,Y) = ((x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2)^{1/2},$$

where $X = (x_1,x_2,x_3)$, $Y = (y_1,y_2,y_3)$. The usual structures of geometry (lines, planes, segments, rays, etc.) can be defined purely algebraically, and it can be shown that these satisfy the postulates of both metric and synthetic geometry. Such a program is only found in the more advanced books on analytic geometry, but a very similar treatment is usually given in purely algebraic terms in elementary books on linear algebra, in which the spaces studied are finite dimensional vector spaces over the real field.

Length and The Classification of Measure Functions. We conclude this section with a few remarks concerning the relationship of our treatment of "length", to the scheme for classifying measure functions which we sketched in Section 1-7. In particular, we want to remind you that we have not attempted to define the concept of length in any formal sense; and that the definition which we adopted for "length function" was, to a considerable extent, arbitrary.

Recall that we have assumed that there is an identifiable set of elements (the domain) which possess the attribute of length. In a formal system, like a geometry, we could identify this domain precisely, but, of course, this identification is only "relative", as the formal system itself is built on undefined terms. On the length domain we were able to establish many structural elements: an equivalence relation; an order relation which carried over to equivalence classes; an "additive" join operation which yielded a corresponding binary operation on the set of all equivalence classes, and which was also related to the order relation; a resulting operation of subtraction in (but not on) the domain of equivalence classes; a ratio structure; and a scalar product structure. All of these related structural concepts had counterparts in the positive real number system, and the simple definition which we adopted for "length function" (it should be (or yield) an additive function on equivalence classes, with values in R^+) gave us length functions which actually "preserved" all of the "length structure" of the domain. Moreover (with what appeared to be justified assumptions in the empirical case) the set of all "length functions" turned out to be a ratio scale.

Let us temporarily call our earlier length functions "ratio length functions". We might then (in the spirit of Section 1-7) define "nominal length functions" to be functions on the length domain, with values in R^+ (or possibly even in R) which preserve the equivalence relation. (I.e., which give the same value to length-equivalent domain elements.) We could define "ordinal length functions" to be those nominal length functions which also preserve the order relation. And we could define "difference length functions" to be those ordinal length functions which also preserve the difference structure. The "ratio length functions" are then those "difference length functions" which are "additive", and we have a hierarchy of types of length function. In practice, of course, we are only interested in the smallest set, the ratio scale. The functions in this category have all of the properties of those in the "weaker" categories, and more. We certainly would not claim that our definition of "length function" was, in any sense, the "right" definition. All we assert is that, on the basis of considerable experience, it is probably the most useful.

With suitable assumptions we could investigate the structures of the several categories of "length functions", but, as these are of no very great interest, we shall not do so. (In essence, they are those functions which can be obtained from length functions by composition with suitably defined permutations, strictly monotone functions, and affine functions.) We shall have more to say about the last-named category in the next section.

2-6 Coordinate Systems

A coordinate system is usually a function (or a set of functions) which measures some attribute of the space to which it applies. Thus coordinate systems can be considered to be measure functions in the general sense.

Even for the simple space of elementary geometry, there are many different kinds of coordinate systems: cartesian coordinates, polar coordinates, cylindrical coordinates, spherical coordinates, and so on. All of these have at least one thing in common: they serve to "identify" or "locate" the points of the space. For this purpose they are usually 1-1 mappings on the set of points of space, into some appropriate value space: in general, this is not simply the real number system.

It is not our intention to undertake a thorough study of coordinate systems, but we shall take a brief look at one of the simplest of these, from the point of view of measure functions which we have adopted in this book. Specifically, we shall take a closer look at the notion of a coordinate system for a line, a notion which has already been mentioned several times.

In the metric treatment of geometry, it is postulated that each line shall possess a coordinate system. This is defined to be a 1-1 function from the line onto the real numbers, which is related in a specific way to a postulated distance function for the space as a whole.

In synthetic geometry, no such distance or coordinate functions are postulated, but we have shown that the congruence-betweenness structure, which is postulated, enables us to prove the existence of length/distance functions, and suitably related coordinate functions; and that, provided that we assume the archimedean and the Cantor-Dedekind postulates, these behave as postulated in the metric treatment.

Let us look briefly at the notion of a coordinate function for a line (in synthetic geometry) from the measurement point of view. Let ℓ be a fixed line. First of all, we should expect that any suitable coordinate function should serve to name, or identify, points of the line. (I.e., it should be 1-1.) Secondly, the postulates of synthetic geometry include a notion of betweenness: we would expect that a suitable coordinate function should preserve this. Finally, certain subsets of the line (segments) have a congruence structure, and we might well expect that a suitable coordinate function should "preserve" this, in some way.

If we think of the real number system, R , as a possible value space for a suitable coordinate function for ℓ, we can use the structure of R to formulate the notion of a coordinate function: an admissible coordinate function is a function $f : \ell \to R$ such that

(i) f is 1-1;

(ii) betweenness of points on ℓ is preserved in the sense that for $A , B , C \in \ell , A - B - C \Longrightarrow f(A) - f(B) - f(C)$, where the notation on the right has the obvious interpretation. (Observe that, as a consequence of this requirement, the image under f of a segment $\overline{AB}$ must be contained in the closed interval $[f(A) , f(B)]$ of real numbers);

(iii) for segments $\overline{AB} , \overline{CD} \in \ell , \overline{AB} \cong \overline{CD} \Longrightarrow |f(A) - f(B)| = |f(C) - f(D)|$.

Notice that all of these requirements are formulated within the postulated concepts of synthetic geometry: they do not involve (explicitly) the concepts of distance and length.

From our earlier work, we know that there exist functions which satisfy these requirements: as we saw earlier, if we go through the process of setting up a length function, $\lambda : D \to R^+$, on the set D of all segments of space, then we can use this (in infinitely many ways) to set up "coordinate functions" $f_\lambda : \ell \to R$ for the line ℓ , and these functions can be shown to satisfy the requirements (i) --- (iii). Moreover, for each fixed f_λ we can obtain other suitable (i.e., satisfying the so-called ruler postulate) coordinate functions by composing f_λ with rigid motions $x \to ax + b$ ($a , b \in R$; $a = +1$ or -1).

We can verify directly that, if f is any admissible coordinate function in the sense that it satisfies (i) --- (iii), then the function gf obtained by composing f with any affine transformation

$$g : x \to ax + b \quad (a , b \in R ; a \neq 0)$$

is also admissible. It is natural to ask, whether or not all admissible coordinate functions can be obtained in this way, from the particular coordinate functions, f_λ , which were derived from length functions.

Let us assume that our space also satisfies the archimedean and the Cantor-Dedekind postulates. Then each f_λ , and each gf_λ is a 1-1 function from ℓ onto R . If now f is any other admissible coordinate function (not necessarily constructed from a length function) then the composite function

$$h = f\, f_{\lambda}^{-1} : R \rightarrow R$$

is a 1-1 function which preserves betweenness, and which preserves equality of differences (i.e., if a, b, c, $d \in \ell$, with $|a - b| = |c - d|$, then $|h(a) - h(b)| = |h(c) - h(d)|$. We comment that the absolute values could be dropped, because the preservation of betweenness ensures that h is either monotone increasing or monotone decreasing; thus the above condition is equivalent to: if $a - b = c - d$, then $h(a) - h(b) = h(c) - h(d)$.

If you now go back to the proof of Theorem 2-2.1, you will find that these properties of h are sufficient to ensure that h is an affine transformation of R. In other words, with the assumptions that we have made, a coordinate system for a line is an affine-invariant measure function, in terms of the classification of measure functions which we described in Chapter 1; and every admissible coordinate system can be obtained (by composition with a suitable affine transformation of R) from any particular coordinate system, f_{λ}, which is derived from a particular length function, λ.

In some measurement situations involving affine-invariant measure functions, we wish to restrict the admissible functions to those which, in some sense, preserve "orientation", or "direction". We shall discuss this idea briefly in the next section, but we mention here that, if such an additional requirement is made of a coordinate function, then the appropriate composition group of functions is the positive affine group of those affine transformations $(x \rightarrow ax + b)$ of R for which $a > 0$.

This is as far as we wish to go in this direction at the present time. We shall return to the discussion of coordinate systems again in the next chapter, in connection with angular coordinates and polar coordinates.

Transformations of Coordinates and of Domain: Scale Models. This seems to be an appropriate place to say something about transformations of the domain (usually called point transformations) and transformation of coordinates, and about the related question of scale models. The discussion is quite incomplete.

Let us first take the viewpoint that our space S is the space of metric geometry, and that a (cartesian) coordinate system for S is a 1-1 function

$$\rho : S \rightarrow R \times R \times R \; (= E, \text{ say})$$

which is related to a distance function in S as follows: Let $X = (x_1, x_2, x_3)$ and $Y = (y_1, y_2, y_3)$, be points in E, and let $\delta : E \times E \rightarrow R^{+}$

denote the "distance" function in E given by $\delta(X,Y) = ((x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2)^{1/2}$. For each coordinate function ρ_i , let α_i denote the corresponding distance function for S ; i.e., the function which is related to the maps $\rho_i \times \rho_i$ and δ , as indicated by the commutative diagram:

$$\begin{tikzcd}
 & E \times E \arrow[dd, "\delta"] \\
S \times S \arrow[ur, two heads, "\rho_1 \times \rho_1"] \arrow[dr, "\alpha_1"'] & \\
 & R^+
\end{tikzcd}$$

We now ask ourselves: what is the relation between those coordinate functions which correspond to the same distance function for S ? Since each coordinate function is 1-1 onto, each pair of coordinate functions determines a 1-1 correspondence on E ; i.e., corresponding to two coordinate functions ρ_1 , ρ_2 , there is a 1-1 correspondence $\beta_{12} : E \to E$ which makes each of the following diagrams commutative:

$$\begin{tikzcd}
 & E \arrow[dd, leftrightarrow, "\beta_{12}"] \\
S \arrow[ur, leftarrow, "\rho_1"] \arrow[dr, leftrightarrow, "\rho_2"'] & \\
 & E
\end{tikzcd}
\qquad
\begin{tikzcd}
 & E \times E \arrow[dd, leftrightarrow, "\beta_{12} \times \beta_{12}"] \\
S \times S \arrow[ur, leftrightarrow, "\rho_1 \times \rho_1"] \arrow[dr, leftrightarrow, "\rho_2 \times \rho_2"'] & \\
 & E \times E
\end{tikzcd}$$

If we are now given a coordinate function ρ_1 , and a "corresponding" distance function, α_1 , for S (in the sense that $\alpha_1 = \delta(\rho_1 \times \rho_1)$), then our question amounts to asking for the set of those transformations $\beta_{12} : E \to E$ which are derived from "equivalent" coordinate functions ρ_2 (i.e., coordinate functions for which $\alpha_1 = \delta(\rho_1 \times \rho_1) = \delta(\rho_2 \times \rho_2)$. This set of transformations, which can be shown to be independent of α_1 (i.e., unit-free), is the so-called group of <u>rigid motions of E</u> . It is fairly easy to see that these rigid motions are just those transformations β of E , which leave the "distance function" $((x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2)^{1/2}$ unchanged, for each pair of points $X = (x_1,x_2,x_3)$, $Y = (y_1,y_2,y_3)$, of E. That is, if $\beta(X) = X' = (x_1', x_2', x_3')$ and $\beta(Y) = Y' = (y_1', y_2', y_3')$ then $(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 = (x_1' - y_1')^2 + (x_2' - y_2')^2 + (x_3' - y_3')^2$. You can find details in [7] .

We can now take a different point of view: instead of looking at changes in the coordinate functions, we can look at 1-1 transformations (so-called "point transformations", as distinct from "coordinate transformations") of the geometric space S onto itself, and ask ourselves what point transformations of S are "distance preserving". From our earlier discussion, you will recall that these are just the congruences (also called "rigid motions") of S.

If we have any point transformation $\gamma : S \to S$ (not necessarily a congruence), and a fixed coordinate function $\rho : S \to E$, then γ determines in a natural way a transformation $\beta : E \to E$, where $\beta = \rho\gamma\rho^{-1}$. (Recall that all of these functions are 1-1 onto, so that the compositional inverses exist.) In other words, β is determined so that the diagram below is commutative:

$$\begin{tikzcd}
S \arrow[r, leftrightarrow, "\rho"] \arrow[d, leftrightarrow, "\gamma"'] & E \arrow[d, leftrightarrow, "\beta"] \\
S \arrow[r, leftrightarrow, "\rho"'] & E
\end{tikzcd}$$

On the other hand, if β is a 1-1 correspondence of E, then, for each coordinate function ρ, there is a point transformation γ of S which is determined by β so as to make the above diagram commutative. Thus each coordinate function ρ, determines a 1-1 correspondence of the point transformations of S and the 1-1 correspondences of E. It can be shown that the set of those point transformations of S which correspond to the group of rigid motions of E is independent of ρ, and that it is the group of congruences of S.

Suppose now that α_1 and α_2 are distance functions for S, and that ρ_1, ρ_2, are corresponding coordinate functions, as described earlier. Then α_1 and α_2 differ by a positive similarity, $\bar{k}$, of R^+ (i.e., $\alpha_2 = \bar{k}\alpha_1$). Moreover there is a transformation β on E, such that $\rho_2 = \beta\rho_1$. It follows that, for any X, Y, in E,

$$\delta(\beta X, \beta Y) = k\delta(X,Y) .$$

This situation is pictured in the following commutative diagram.

$$\begin{tikzcd}
 & E \times E \arrow[r, "\delta"] \arrow[dd, leftrightarrow, "\beta \times \beta"] & R^+ \arrow[dd, leftrightarrow, "\bar{k}"] \\
S \times S \arrow[ur, leftrightarrow, "\rho_1 \times \rho_1"] \arrow[dr, leftrightarrow, "\rho_2 \times \rho_2"'] \arrow[urr, bend left, "\alpha_1"] \arrow[drr, bend right, "\alpha_2 = k\alpha_1"'] & & \\
 & E \times E \arrow[r, "\delta"] & R^+
\end{tikzcd}$$

The set of all transformations, β derived in this way is a group, called the <u>extended</u> <u>similarity</u> <u>group</u> on E . Clearly this group contains the rigid motions of E .

We now turn the picture around, and consider the "dual" situation: for a given coordinate function ρ , what is the set of those point transformations of S which (as described above) correspond to the extended similarity group on E ? This is the set (actually a group) of those point transformations γ of S , each of which has the effect of multiplying all distances by some $k > 0$. The situation is as pictured in the following commutative diagram:

$$\begin{tikzcd}
S \times S \arrow[r, two heads, leftrightarrow, "\rho \times \rho"] \arrow[d, leftrightarrow, "\gamma \times \gamma"] \arrow[rr, bend left, "\alpha"] & E \times E \arrow[r, "\delta"] \arrow[d, leftrightarrow, "\beta \times \beta"] & R^+ \arrow[d, leftrightarrow, "\bar{k}"] \\
S \times S \arrow[r, two heads, leftrightarrow, "\rho \times \rho"'] \arrow[rr, bend right, "\alpha"'] & E \times E \arrow[r, "\delta"'] & R^+
\end{tikzcd}$$

This group of point transformations of S (which includes the congruences) is, of course, the group of similarity transformations of classical geometry. (Similar figures are those which correspond under a similarity transformation of S.) It can be shown that this group is independent of the choice of ρ; i.e., similarity is a unit-free equivalence relation.

If α_1, and $\alpha_2 = k\alpha_1$, are distance functions for $S \times S$, and γ is a k-fold similarity of S, we can summarize the above in the following commutative diagram:

$$\begin{tikzcd}
S \times S \arrow[r, "\alpha_1"] \arrow[d, leftrightarrow, "\gamma \times \gamma"'] \arrow[dr, "\alpha_2"] & R^+ \arrow[d, leftrightarrow, "\bar{k}"] \\
S \times S \arrow[r, "\alpha_1"'] & R^+
\end{tikzcd}$$

Similarity (point) transformations of S have a simple relationship to the matter of "scale" models. We may consider an object in the physical world as mapped into S by a "model function". Generally this function is suppressed, with the object regarded as actually being a subset of S, rather than being in 1-1 correspondence through a certain "model function", with a subset of S. (Another way of looking at this, is to regard physical space, with "points" of an object corresponding to the "points" of physical space which they "occupy", as a metric geometry. That is, physical space is assumed to have an empirical structure which makes it isomorphic to the formal mathematical system S, so that we may treat physical space as a metric geometry, and physical objects as subsets of this geometry. When this is done, the length/distance functions of the mathematical system are usually given the same names (i.e., inch, foot, meter) as the empirical functions used in measuring the physical distances involved. It is possible that this suppression of the "model function" is partly responsible for the common misconception that physical space actually is a metric geometry.)

For the sake of simplicity, we suppress the "model" function and treat physical space as if it is a metric geometry S. We may then describe physical objects as being <u>similar</u>, if there is a point transformation of S, of the type which we have called a k-fold similarity, which maps one onto the other. The result which is pictured in the last commutative diagram, then states that the effect (on the length measure of an object) of changing the length scale (i.e., the length function, not the unit) by a factor k , is indistinguishable from the effect of a k-fold similarity on the domain.

It should be noted that our definition of similar objects is broader than the common usage of the term "scale model". For the latter there is usually an additional restriction, that "orientation" be preserved, to rule out those similarities which change left and right handedness. The distinction is, of course, the same as that between rigid motions (congruences) in the mathematical sense, and the corresponding more restricted physical idea of a rigid motion, which permits only those transformations which are composed of translations and rotations.

While the above discussion has been concerned mainly with length, there are important relationships between similarity transformations of domain and other common measurement functions, including, of course, angle measures, area, and volume. We return to this question later.

2-7 Directed Segments, Orientation, and Vector Measures.

There are many purposes for which the length functions which we have described are, in a sense, too crude. For example, given a pair of (different) points A , B , in space S , we might wish to have a concept of "distance", which distinguishes between the "distance" from A to B , and the "distance" from B to A . That is, we might look for a "distance" function, on S , which preserves the distinction between the ordered pairs (A,B) , (B,A) . Our first thought is that we could do this quite simply, by defining one of these "distances" to be the negative of the other: i.e., by using as value space the whole real number system, instead of just R_0^+ . This immediately raises the question as to which of the two orders should get the positive "distance" value, and which the negative value; and whether or not there is some useful way of making such a choice "uniformly" throughout space. This immediately involves us in questions of orientation, which we discuss briefly below: a full consideration of orientation is beyond the scope of this book.

Instead of considering ordered pairs of points, and "oriented distances", let us look at the equivalent situation with directed segments. A directed segment is a segment, together with an ordering for its end points. (Other equivalent definitions could be given.) Thus each directed segment determines a unique segment, but there are two different directed segments corresponding to each segment. The notation $\vec{AB}$ is often used to denote the directed segment $\{\overline{AB}, (A,B)\}$, but we have already used the notation $\overrightarrow{AB}$ to denote a ray. We therefore adopt the symbol $\overrightarrow{A,B}$ to denote the directed segment "from A to B". Thus $\overrightarrow{A,B} = \overrightarrow{C,D}$ if and only if $A = C$ and $B = D$.

If we now ask whether there are functions, f , defined on the set $\vec{D}$ of all directed segments, and with values in R , which have the properties:

(i) $f(\overrightarrow{A,B}) = -f(\overrightarrow{B,A})$ for every $\overrightarrow{A,B}$ in $\vec{D}$;

(ii) the function λ_f , derived from f by defining $\lambda_f(\overline{AB}) = |f(\overrightarrow{A,B})|$, is a length function for the set D of all segments;

then the answer is clearly "yes": we can take any length function λ_f on D , and make an arbitrary choice of value $(\pm \lambda_f(\overline{AB}))$ for $f(\overrightarrow{A,B})$, and then define $f(\overrightarrow{B,A})$ to be $-f(\overrightarrow{A,B})$. Of course this is not very satisfactory, but it emphasizes that there certainly exist appropriate functions if we ask no more from them than that they satisfy (i), (ii) above.

In addition to (i), (ii), we might look for other conditions which we would like such a function to satisfy. For example we might ask that:

(iii) for all directed segments on a given line, the value of f should have the same sign for all directed segments which "point in the same direction" (see below);

(iv) for all directed segments which are on parallel lines, and which "point in the same direction", f should have the same sign;

(v) for all directed segments which are "sufficiently close together", f should have the same sign.

If (v) is interpreted to mean that, in particular, f should give the same sign to directed segments $\overrightarrow{A,B}$, $\overrightarrow{A,C}$, such that the angle $\angle CAB$ is "sufficiently small", then, intuitively, this condition cannot be satisfied along with condition (iii). For (intuitively) the directed segment $\overrightarrow{A,B}$ could then be continuously rotated, without changing its sign, to the position $\overrightarrow{A,B'}$ "opposite" to $\overrightarrow{A,B}$; conditions (iii) and (i) would then require that $f(\overrightarrow{A,B'}) = f(\overrightarrow{B,A}) = -f(\overrightarrow{A,B})$, while condition (v) would require that $f(\overrightarrow{A,B'}) = f(\overrightarrow{A,B})$.

Condition (iii) (alone) can certainly be satisfied: all we need to do is to show that the set of all directed segments on a line can be partitioned into two equivalence classes by an appropriate relation of "same direction"; see exercises below. Condition (iv) can also be satisfied, along with condition (iii).

Exercises 2-7

1. Let $\overrightarrow{A,B}$ and $\overrightarrow{C,D}$ be directed segments on a line ℓ. We define a relation $\uparrow$ on the set $\vec{D}_\ell$ of all directed segments on ℓ, by $\overrightarrow{A,B} \uparrow \overrightarrow{C,D}$ if and only if $\overrightarrow{AB} \cap \overrightarrow{CD}$ is a ray. Prove that $\uparrow$ is an equivalence relation on $\vec{D}_\ell$, and that there are exactly two equivalence classes with respect to $\uparrow$.

2. If $\overrightarrow{A,B}$, $\overrightarrow{C,D}$ are two directed segments on a line ℓ, then (from the postulates of synthetic geometry: see Section 2-5) the points A, B, C, D, (at least two of which must be distinct) can be named in an order X_1, X_2, X_3, X_4, such that $X_1 - X_2 - X_3 - X_4$. (If only two or three of these points are distinct, this must be interpreted accordingly.) If they are so named, with $A = X_{i(A)}$, $B = X_{i(B)}$, $C = X_{i(C)}$, $D = X_{i(D)}$), prove that $\overrightarrow{A,B} \uparrow \overrightarrow{C,D}$ if and only if, $i(A) - i(B)$, and $i(C) - i(D)$, have the same sign.

3. If $f : \ell \to R$ is a coordinate function for a line ℓ, and if $\overrightarrow{A,B}$ and $\overrightarrow{C,D}$ are directed segments in ℓ, prove that $\overrightarrow{A,B} \uparrow \overrightarrow{C,D}$ if and only if $f(A) - f(B)$ and $f(C) - f(D)$ are either both positive or both negative.

The two equivalence classes of Exercise 1 above, lead to the concept of orientation: we orient the line by selecting one of these equivalence classes. Each class is called a direction. Thus there are two possible directions, and each is completely determined by any of its elements; i.e., by a directed segment.

If we choose one of the two directions, on a fixed line ℓ, as a "positive" direction, and the opposite direction as the "negative" direction, then a function $f : \vec{D}_\ell \to R$ ($\vec{D}_\ell$ denotes the set of directed segments on ℓ) which satisfies (i) --- (iii), can be established in the obvious way. Moreover the concept of direction can be extended to parallel lines in a fairly straightforward way. We shall not go into details of this extension.

Sometimes these ideas are used to ascribe "negative length" to certain directed segments, but it is preferable not to use the word length in connection with measure functions (generalized length functions) on directed segments. It is better to regard the objects under discussion (or, rather, certain equivalence classes of them, under the relation of "same direction and congruent as segments") as vectors, and to look for measure functions which are relevant to the "vector space" structure of the set of such vectors. (In order that this set be a vector space, it must, of course, have a "null vector" or "zero vector"; the class of "degenerate" directed segments $\overrightarrow{A,A}$ provides this particular vector.) For such "vector measures", the appropriate value space is often a so-called "real euclidean vector space", of appropriate dimension. For the restricted set of such vectors on a fixed line, the corresponding 1-dimensional vector space is isomorphic to the real number system, and this is why it is possible to have functions like f above, which are linear isomorphisms from the 1-dimensional vector space $[\vec{D}_\ell]$ (whose elements are equivalence classes of directed segments in $\vec{D}_\ell$) to the real number system R .

We do not wish to get seriously involved in questions of vector measures, which quickly lead into questions of linear transformations and matrices. (There are many excellent books which deal with these ideas.) We do however point out that our earlier join operation for segments can be appropriately modified, and that the set of vectors $[\vec{D}_\ell]$ is an abelian group under this operation. Moreover (generalizing the corresponding situation for segments) there is a "scalar multiplication" of vectors, by elements from an appropriate number field. (If we assume the archimedean and Cantor-Dedekind postulates, this is the field of all real numbers.) These structural properties (an abelian group which is closed under scalar multiplication from an appropriate field) are characteristic of vector spaces. The vectors (equivalence classes of similarly directed congruent segments) in S do, of course, form a vector space, with vector addition, by the "parallelogram rule", corresponding to the modified join operation in $[\vec{D}_\ell]$.

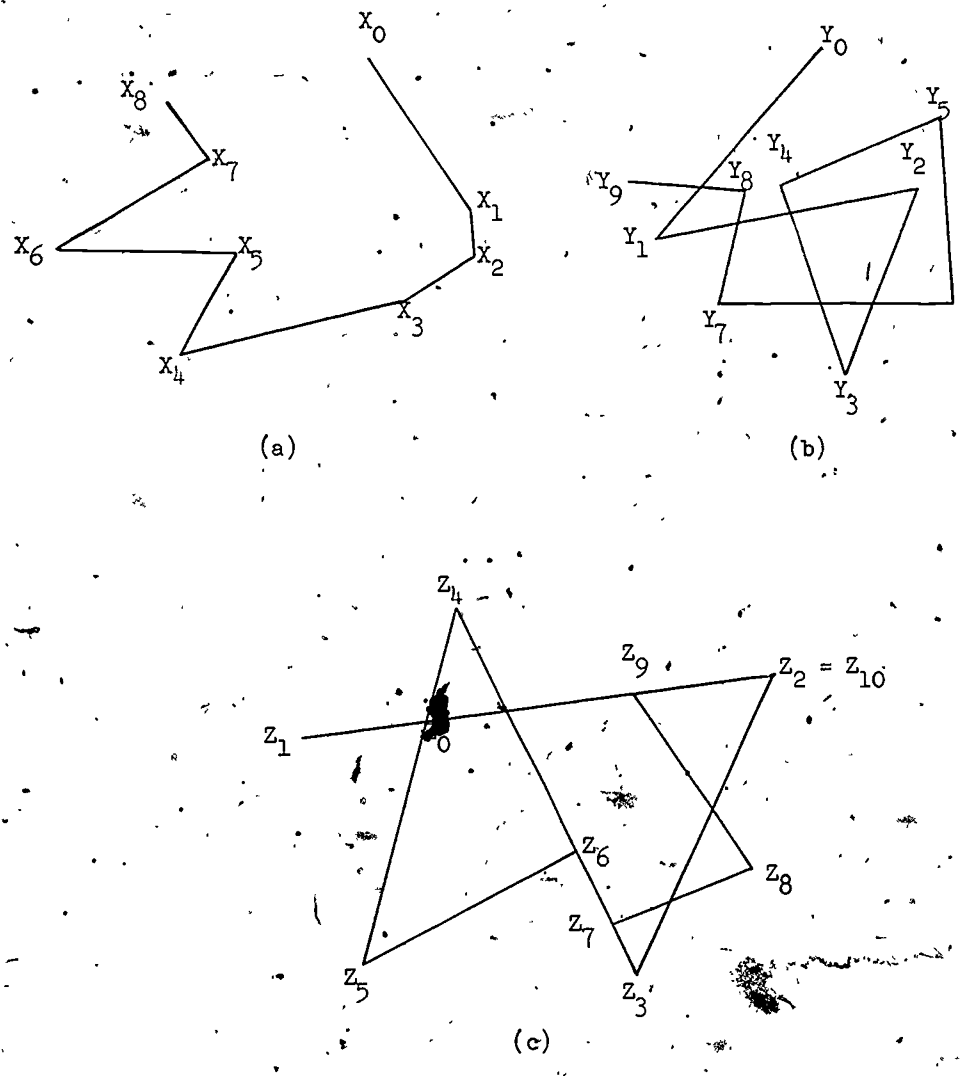
2-8 Extension of the Domain for Length Functions: Curve Length

The foregoing discussion of length functions, both from an empirical and from a mathematical standpoint, has been confined to the length of segments and their physical counterparts. And, even with this restriction, our empirical discussion paid no attention to the fact that the empirical counterparts of all segments could not, in fact, be manipulated in the way which we briefly (and rather vaguely) indicated for the establishment of empirical length-measurement relationships and functions. In fact, at the empirical level there are important practical problems to be solved in defining length functions for a domain which includes all "rods" (i.e., the physical counterparts of all segments). These important questions are properly discussed as part of physics, and we do not concern ourselves with them here, except to point out again that it is not really possible to solve them in a purely empirical framework. For example, the measurement of very large distances involves not only assumptions concerning the behavior of light, but also assumptions concerning the nature of physical space. These assumptions usually involve mathematical "models", and mathematical ideas of distance.

What we are concerned with under the above heading, is the extension of the domain of length functions from segments and their empirical counterparts, to include also such objects as pieces of string, highway distances, perimeters of polygons, circumferences of circles, and so on. Here again, empirical length measurement usually involves both practical questions and mathematical questions. In order to keep the discussion reasonably brief we shall not discuss empirical length measurement further. (The mathematical questions which are relevant to the empirical development, will be mostly answered in a discussion of the corresponding questions of domain extension for length functions in a formal mathematical system.)

From a mathematical point of view, the question of extending the domain of length functions proceeds in two more or less distinct directions, which eventually make contact again at a more advanced level: one of these extensions is concerned with the length of objects called "curves", and the other is concerned (initially) with the generalization of length functions on segments, to a larger class of subsets of the real line. (This is part of the subject known as "measure theory"; see, for example, [11] .) Each of these generalizes our earlier treatement, in the sense that, for each generalization, the domain of an appropriate measure function includes the set of all segments (or something isomorphic to this set), and each function agrees with one of the length functions defined above, on their common domain, the set of all segments. Anything approaching a full treatment of either generalization is beyond the scope of this book, but we can at least give you some idea of the direction in which each leads.

Broken Segments. Before discussing curves and curve length in general, we consider the class of broken segments. (These are often called broken lines, but this terminology is not in keeping with the way in which the words "line" and "segment" are now used.) An elementary broken segment (where there can be no misunderstanding, we abbreviate this term to "broken segment") is a finite sequence of segments $\overline{X_0X_1}$, $\overline{X_1X_2}$, $\overline{X_2X_3}$, ... , $\overline{X_{n-1}X_n}$, which, in addition to the "chain" property suggested by the terminology, satisfy the further condition that each two of them have at most one point in common. Clearly a segment may be regarded as a broken segment. Diagrams (a) and (d) below illustrate elementary broken segments, but diagram (c) does not:



(a)

(b)

(c)

A broken segment is simple, if the only intersections of its constituent segments are at the end points of successive segments. Thus (a), above is simple, but (b) is not. Every segment is a simple broken segment. If b is a broken segment, we denote by $\bar{b}$ the set which is the union of its constituent segments.

We consider the problem of extending length functions from the domain D of segments, to the domain D_B of broken segments. Our natural impulse is to take a length function, λ, for D, and then extend this function to D_B by defining its value on a broken segment to be the sum of the values of λ on the constituent segments. We shall see that, in a certain sense, this is the only reasonable thing to do: if we set down those properties which we feel that every acceptable length function for broken segments must have, we shall find that the only functions with these properties are those which are obtained by adding the values of some λ on the constituent segments. This deduction is quite straightforward, so we give it in some detail as a further illustration of our general approach to measure questions: we try to identify a domain, with some structural properties related to the attribute in question (in this case, length) and then look for functions which have some specified properties in relation to the structure of the domain.

Those simple properties which we might reasonably require of every length function for broken segments are:

(i) If $\lambda_B : D_B \to R^+$ is such a function, then λ_B must agree with one of our length functions (λ, say) on the subdomain D, of segments. Moreover each λ for D should have such an extension to D_B.

(ii) If b_1, b_2, are broken segments, with $\bar{b}_1 \cong \bar{b}_2$, then $\lambda_B(b_1) = \lambda_B(b_2)$.

(iii) If b', b'', are "piecewise congruent", then $\lambda_B(b') = \lambda_B(b'')$; that is, if $b' = (b'_1, b'_2, \ldots, b'_n)$ and $b'' = (b''_1, b''_2, \ldots, b''_n)$, and there is a 1-1 correspondence of the segments b'_i with congruent segments b''_j (order is not important), then we require that $\lambda_B(b') = \lambda_B(b'')$.

We verify easily that, if λ is any length function for D, and if λ_B is defined for each broken segment by adding the values of λ on the constituent segments, then λ_B satisfies (i), (ii) and (iii). We shall prove the converse. Conditions (ii) and (iii) can be combined to define a relation $(\sim)$ on D_B: we define $b_1 \sim b_2$ if there are broken segments b'_1, b'_2, with $\bar{b}_1 \cong \bar{b}'_1$, $\bar{b}_2 \cong \bar{b}'_2$, such that b'_1 is piecewise congruent to b'_2. It is easy to prove that $\sim$ is an equivalence relation. (Intuitively, broken segments are equivalent if they have piecewise congruent subdivisions, where a subdivision is obtained by the "insertion" of a finite number of additional vertices at interior points of constituent segments.)

Let $\tilde{D}_B$ denote the resulting set of equivalence classes of broken segments. Conditions (ii) and (iii) require that λ_B must have the same value on each broken segment in a particular equivalence class, so we may consider λ_B as defined on $\tilde{D}_B$. Congruent segments belong to the same equivalence class, so we have a mapping $f : \tilde{D} \to \tilde{D}_B$. (As before, $\tilde{D}$ denotes the set of congruence classes of segments.)

We show next that:

(a) f is onto; i.e., every equivalence class of broken segments contains a segment;

(b) f is 1-1; i.e., if two segments belong to the same equivalence class, then they are congruent.

In other words, f is a 1-1 correspondence. The proof of (a) and (b) is a straightforward application of the join operation for segments:

(a) If $b = (b_1, b_2, \dots, b_n)$ is a broken segment, then (from Section 2-5) the join class $\tilde{b}_1 * \tilde{b}_2 * \dots * \tilde{b}_n$ is a congruence class of segments, each of which is piecewise congruent under subdivision, to b, and hence equivalent to b. Hence f is onto.

(b) If two segments are equivalent, then they are congruent, or they have piecewise congruent decompositions (or both). But a segment is the join (in some order) of the segments in its decomposition, and the join operation is uniquely defined on congruence classes. Hence the two segments are congruent.

Now that we know that each equivalence class of broken segments contains exactly one equivalence class of segments, condition (i) requires that if $b = (b_1, b_2, \dots, b_n)$ is a broken segment, and if c is a segment with $c \sim b$, then $\lambda_B(b) = \lambda(c)$, where λ is the length function λ_B/D for segments. Since $c \sim b$, c and b have piecewise congruent decompositions. But as we saw above, any segment is the join (in some order) of the segments in its decomposition, and equivalent broken segments have congruent joins. Thus the congruence classes $\tilde{c}$ and $\tilde{b}_1 * \tilde{b}_2 * \dots * \tilde{b}_n$ are the same. But λ carries joins into sums, hence $\lambda_B(b) = \lambda(c) = \sum_{i=1}^{n} \lambda(b_i)$, which is what we set out to prove.

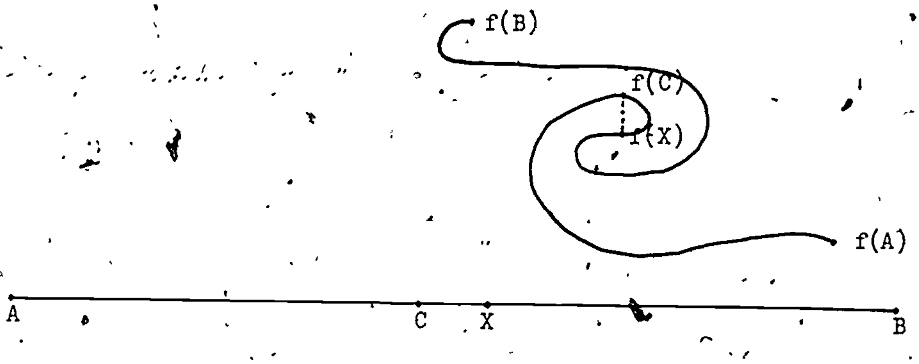
Of course all of this is rather "obvious", but we have given it in some detail, because the analogous idea (piecewise congruence under decomposition) turns out to be very useful in the discussion of angle measures and "generalized angles". Moreover, there are very interesting related problems connected with the so-called elementary theory of area for polygonal regions, and with the non-existence of such an elementary theory for the volumes of polyhedra.

Rectifiable Curves. The intuitive idea of a curve is simple enough, but the intuitive idea is quite inadequate for a discussion of curve length: in order to discuss curve length seriously, we must first clarify our ideas about what is, or should be, a "curve".

Our first approach to the concept might be to try to formalize the physical idea of taking a straight piece of string (a segment) and bending it (in space) in a quite arbitrary way, but without stretching or breaking. That is, we might try to define a curve as a 1-1 continuous image (see below for a discussion of the notion of continuity) of some line segment $\overline{AB}$, under a function f which, in some sense, "preserves length". In view of the fact that the only notion of length which we have to work with, is the length of a segment or broken segment, we could try to convey this idea of length preservation by imposing a "local" condition at each point C of $\overline{AB}$, corresponding to the intuitive idea of "continuous deformation without stretching". This condition would probably take the form that, for each point C of $\overline{AB}$,

$$\lim_{X\to C} \frac{\alpha(f(X), f(C))}{\alpha(X, C)} = 1, \qquad \text{(see diagram)}$$

where α is the distance function in S, and the limit idea would need to be made precise.



In view of our intuitive feeling that "the straight line is the shortest distance between two points", we would expect that, if f is a "length-preserving" function, then, for all $X \in \overline{AB}$, $\alpha(f(X),f(C)) \leq \alpha(X,C)$; and that, the ratio of these distances should be arbitrarily close to 1 when X is close enough to C. This suggests that we should formalize the above limit by the requirement that the least upper bound of the ratio of these distances should be 1. In other words, summing up, this line of thought would lead to a definition of a curve as a 1-1 continuous image of some line segment $\overline{AB}$ under a function f which satisfies

$$\sup_{(X \in \overline{AB})} \left\{ \frac{\alpha(f(X)\ ,\ f(C))}{\alpha(X,C)} \right\} = 1$$

for each $C \in \overline{AB}$.

At first you might think that this approach would be very useful, as it seems to correspond exactly to our intuitive idea of obtaining a curve by a length-preserving distortion of a segment, and it appears to have the additional advantage of leading directly to the definition of curve length. (For such a "curve", we would, naturally, define its length to be that of the segment which appears in its definition.) Moreover segments themselves are clearly "curves" under this definition, and their "curve" lengths are easily seen to be the same as their segment lengths.

Unfortunately this is not the usual way in which objects which we wish to call curves, arise in mathematics: the objects which we wish to consider as curves usually arise from continuous functions (sometimes, but not generally, 1-1) defined on segments of the real line, and it turns out to be more useful to treat the questions of curve definition and curve length separately. In the discussion below of these questions we will see that, with the definitions which we eventually adopt, not all curves can be assigned "lengths": those which can, will be called <u>rectifiable curves</u>. It will be true that those rectifiable curves which correspond to 1-1 continuous functions, will be "curves" in the sense tentatively discussed above (i.e., images of continuous 1-1 "length-preserving" functions on <u>some</u> segment), but this will not be true of curves generally.

In several places above we have mentioned the word "continuous", as applied to a function from a segment to space S . You are probably familiar with the notion of a continuous real-valued function of a real variable, and with the usual $\varepsilon - \delta$ definition, which you can find in any good elementary calculus text, such as the SMSG "Calculus". The definition merely makes precise the intuitive idea that all "sufficiently close" points should map into points which are "arbitrarily close", and it carries over, with very little modification, to functions defined on R or on a segment of R with values in the plane, or in space. We shall not give the definition formally, but we point out that, although the usual definition involves the distance functions of R and S , the notion of continuous function from a segment of R to S is actually independent of any particular distance function in either of these spaces, and depends only on the so-called "topology" of the spaces concerned. Because of this, it is not surprising that topological ideas enter into the discussion of curve length. Again, a full discussion would take us beyond the scope of this book, but even for our limited discussion it will be useful to use the notion of topological transformation: a topological transformation (or homeomorphism) is a 1-1 onto function which is bi-continuous (i.e., continuous in each direction). We need to use the intuitively obvious fact that any segment is homeomorphic to any other segment, in an infinite number of ways. (Intuitively, we map the end points of one segment onto the end points of the other, in either order, and complete the mapping by stretching or shrinking the interior quite arbitrarily provided that we keep the correspondence 1-1.).

In a second attempt to come to grips with the notions of curve and curve length, we might well decide to separate them and to define a curve as the range of a 1-1 continuous function from a segment into S ; i.e., as a homeomorphic image of a segment in S . This would correspond to the idea of a string which is arbitrarily bent, shrunk, stretched, or twisted, in whole or in part, with the 1-1 condition corresponding to the intuitive idea that no two points of the string could occupy the same position at the same time. We could carry out an entirely satisfactory discussion of curve length for such "curves", leading to a generalization of the notion of length to a much larger domain than the set of broken segments. But we would then find that our idea of curve was still not sufficiently general, and that we would want to extend the idea of curve still further, to include not merely topological (homeomorphic) images of segments, but also to include arbitrary continuous functions on segments. (Notice that we did not say the ranges of arbitrary continuous functions on segments: see comment below.)

The idea which we will want to formalize, is that a curve should correspond to the path traced out by a point which moves continuously in a given time interval (or "segment" of time). This physically-motivated idea of "curve" turns out to be the most fruitful one, and it includes our earlier tentative idea as a special case. Therefore, to keep the discussion reasonably brief, we shall formalize this more general idea, and discuss the corresponding notion of curve length in this context. While our treatment will be (or can be made) completely mathematical, we shall refer frequently to the motivating idea of a curve as the path of a moving point.

If x , y ; $(x \neq y)$ are real numbers, we denote by $[x,y]$ the segment (or interval) of real numbers, $[x,y] = \{r : r \in R,\ x \leq r \leq y \text{ or } y \leq r \leq x\}$. We now define a curve to be a continuous function from some real interval [x,y] to space S . A simple curve is a curve which is 1-1 on the interior of the interval; this corresponds to our more restricted notion as discussed above. Notice that a curve can be a constant function: i.e., a function whose range is a single point.

Comment: The subject of curves and curve length, is poorly treated in many Calculus texts. Many of them avoid giving any definition of curve, while others (including some good ones) define a curve to be the image, in S , of a continuous function whose domain is an interval of real numbers. This would be satisfactory if we restricted attention to simple curves, but it is not satisfactory as a general definition of curve.

You will probably find it strange, at first, that we define a curve to be a function, rather than the image of a function. But the definition allows for the idea of a moving point which might return to the same position many times, retrace part of its route, reverse direction, stand still throughout the whole interval, and so on; and the notion of curve length, which we shall introduce, will correspond to the "length of the path traveled". Thus it would not be satisfactory (insofar as our length definition is concerned) to define the curve to be merely the image set in S : many different curves, with different lengths, will have the same image set. Actually you are almost certainly familiar already with the notion of a curve as a function; e.g., the so-called "sine curve". You are probably accustomed to thinking of the sine curve as the graph of the sine function (a function from R to R , whose range is the interval [-1,1]) and certainly not as the range in R (i.e., the interval [-1,1]) of the sine function. And, as you will recall from our earlier discussion, if a function is defined as a certain set of ordered pairs in the cartesian product of its domain and its image space, then the function becomes identical with its graph.

From a physical point of view, defining a curve to be a function is equivalent to defining it to be the graph of the function in "space × time". This graph is a 1-1 continuous image of an interval of R , in the cartesian product of R with S , and this is quite different from the range of the curve function in S . If we were to restrict consideration to simple curves, as far as length is concerned we could get away with defining the curve to be the range: it will turn out that, although different simple curves (functions) have the same range, simple curves which have the same range will all have the same length. For this reason we will use the word "segment" ambiguously, to denote both a geometric segment, and any one of the simple curve functions having this segment as its range. The same remark applies to other simple curves, including, of course, elementary broken segments.

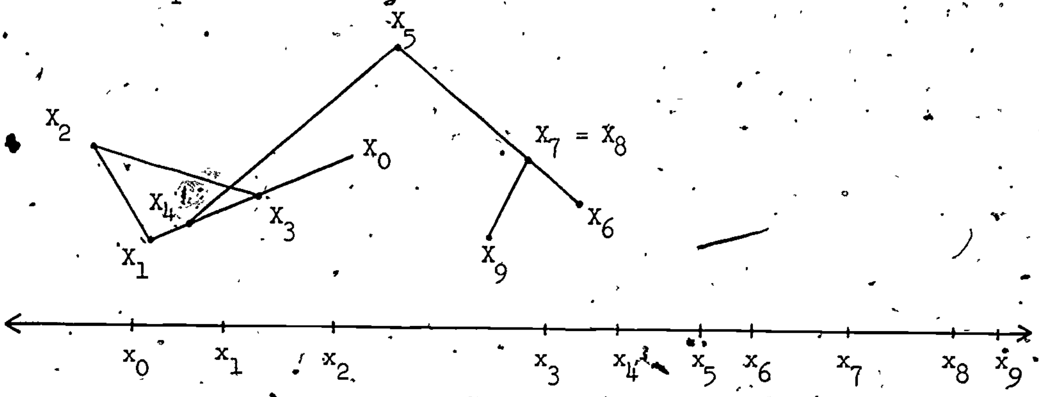
We turn now to a discussion of the introduction of length functions in the set Γ of all curves (as defined above). We observe first that we can "imbed" the set D of all segments of S into Γ by the simple device of using coordinate mappings: for any segment $\overline{AB}$ of S, if ℓ is the line containing $\overline{AB}$, there is a 1-1 onto coordinate mapping

$$\rho : \ell \to R$$

such that $\rho(\overline{AB})$ is a segment $[a,b]$ of R, with $\rho(A) = a$, $\rho(B) = b$, and $\alpha(A,B) = |a - b|$. The segment $\overline{AB}$ can now be "identified" with the function which is inverse to ρ, and restricted to (i.e., defined on) the domain $[a,b]$. In this sense, we can consider that Γ contains D, the set of all segments of S. An easy extension of this idea enables us to imbed (in infinitely many ways) the set of elementary broken segments in Γ.

The idea which is used in an attempt to extend the length functions with domain D, to all (or as much as possible) of Γ, goes back to the early Greek mathematicians, who used it successfully to define and calculate the lengths of many simple curves. The idea is to "approximate" the curve by broken segments.

We define a <u>broken segment</u> to be a curve which is "piecewise linear". That is, $f : [a,b] \to S$ is a broken segment if (assuming $a < b$) there are numbers $a = x_0 < x_1 < x_2 < \dots < x_n = b$, such that for each $i = 1, 2, \dots, n$, the restricted function $f|[x_{i-1}, x_i]$ is constant, or is a segment. This means, of course, that $f/[x_{i-1},x_i]$ is either constant or 1 - 1, and that the image of $f/[x_{i-1},x_i]$ is either a point, or a segment in the geometrical sense. But this does not mean that the overall image of f on $[x_0,x_n]$ is an elementary broken segment. The diagram below, in which X_i denotes $f(x_i)$, gives the idea of a broken segment:



The diagram cannot exhibit the function itself, but only its range. The diagram illustrates the fact that the definition allows for a broken segment (curve) whose image may be quite different from that of a simple broken segment, and which might not fit in with your intuitive picture of what a curve should be. This degree of generality is really necessary if a curve is to be the formalization of the idea of the path of a moving point: you should have no difficulty in envisaging such a "motion", as represented by the diagram.

If curve length is to correspond to the "length of the path traveled", then the appropriate definition for the <u>length of the broken segment</u>

$$f : [x_0,x_n] \to S ,$$

under a length function related to the distance function α for S, is clearly,

$$\begin{aligned}\lambda(f) &= \alpha(f(X_0),f(X_1)) + \alpha(f(X_1),f(X_2)) + \ldots + \alpha(f(X_{n-1}),f(X_n)) \\ &= X_0X_1 + X_1X_2 + \ldots + X_{n-1}X_n\end{aligned}$$

where we use the usual notation $X_{i-1}X_i$ to denote the length (under the relevant length function) of the segment $\overline{X_{i-1}X_i}$. (If $X_{i-1} = X_i$, then $X_{i-1}X_i$ is, of course, 0.) Thus λ is defined on the set of broken segments (which can be assumed to include the set D of all segments, and the set D_B of elementary broken segments), and its values are nonnegative real numbers.

The verification of the following statements is somewhat lengthy, but not difficult:

(i) On the subset D_B, (which includes D) λ "agrees" with the particular length function for D_B which corresponds to the distance function α.

(ii) There is a 1-1 correspondence between the length functions for D_B and their extensions to the set of all broken segments; i.e., each length function has a unique extension.

(iii) If $f : [x_0,x_n] \to S$ is a broken segment, and $\sigma : S \to S$ is a congruence (rigid motion) then σf is a broken segment, and $\lambda(f) = \lambda(\sigma f)$.

(iv) It is possible to establish a "length structure" on the set of all broken segments, with a suitable equivalence relation (including the congruence condition of (iii) and an appropriate generalization of "piecewise congruence under subdivision", and a condition which expresses formally -- by means of a monotone function on the domains of the curve functions -- the idea of traversing.

the same path, but not necessarily in the same direction or at the same "rate"); an inequality relation; and a join operation on the equivalence classes. These agree with the earlier relations and operations on D , and with them the set of equivalence classes of broken segments becomes an ordered abelian semigroup, which, with the zero element excluded, is naturally isomorphic to the semigroup $\tilde{D}$. Moreover, it can be shown that the length function defined above for broken segments agrees with the earlier isomorphism of $\tilde{D}$ onto R^+ . In this sense the extension is unique.

(v) If λ_1 , λ_2 , (with $\lambda_1 = k\lambda_2$) are length functions on $\tilde{D}$, then their extensions to the set of all broken segments have the same relationship; i.e., they are related by composition with the same positive similarity $\bar{k}$ of R_0^+ .

The next step is to attempt to further extend the domain to the set D_c of all curves. We can proceed directly, using the intuitive idea of approximating a curve by broken segments, and define a length function λ for curves, by defining the <u>length of a curve</u> (with respect to λ) to be the least upper bound (if it exists) of the set of lengths (under a fixed length function λ_0) of all approximating broken segments. (Eventually it is convenient to use the same symbol for a length function for broken segments and for its unique extension to the set of curves.) The definition is motivated by two simple requirements:

(i) the length of the curve itself should be at least as large as the length of any approximating broken segment. (This is derived from the intuitive idea of a straight line as the shortest distance between two points); and

(ii) in the set of all approximating broken segments there should be some whose lengths approximate the desired curve length arbitrarily closely.

It turns out that, with the general definition which we are using for curve, there are some curves for which (under each underlying distance/length function on S/D) the set of the lengths of all approximating broken segments is unbounded, and hence there is no least upper bound. Those curves for which such a bound (and hence a least upper bound) exists (under any, and hence under every distance function) are called <u>rectifiable curves</u>. The set D_r of all rectifiable curves is the domain of our extended length functions. This set includes the set of all broken segments; and, for a given underlying segment-length function, the curve length of a broken segment is easily shown to be its length as a broken segment. As for broken segments, curve length is invariant under congruence or piecewise congruence (using a finite number of "pieces"); it is monotone, in the sense that (with the obvious definition for subcurve) the length of a subcurve is $\leq$ the length of the curve; it is additive with respect to a naturally defined join operation for curves; and the "chain rule" for changing units, still applies.

We could try to establish a "length structure" on the domain of curves itself, without using the definition of length, and attempt to prove that the defined length functions are the only homomorphisms which preserve this structure, and which extend the earlier length functions. Something can be done along these lines, but the treatment is quite lengthy, and it cannot avoid involvement with topological ideas which might be unfamiliar to you. Moreover the rather complicated details tend to obscure the essential simplicity of the ideas involved, which, we repeat, are completely motivated by consideration of the physical idea of the distance traveled by a moving point-object during a finite time-interval.

In view of the almost universal acceptance in our culture of the statement "a straight line is the shortest distance between two points", we remark, that, for the geometric space S, it is indeed true that, of all the curves which "join" two points of S, the line segment has least length. (This is a unit-free statement. You should try to prove it for yourself, using the familiar "triangle inequality" of geometry.) But this is not a very profound observation, because (as remarked above) our definition of curve length was partly motivated by the desire to achieve just this situation.

Having defined length functions on the domain of rectifiable curves, the next step is to consider how we might actually calculate the values of these functions for particular elements of the domain. As indicated above, the Greeks succeeded in doing this in a few cases, using rather intuitive methods. But, as you are probably aware, the most common computational device in current use involves calculus methods, especially the evaluation of certain definite integrals. You can find details of such methods and calculations in most books on calculus, and we do not intend to discuss them here. We do, however, comment, that calculus methods for length calculation are applicable to very few (in a certain sense, almost none) of the rectifiable curves. But, fortunately, these few include most of those of greatest importance for the empirical concept of curve length.

In case you wish to pursue the question of curve length, it is only fair to point out that very few calculus texts contain much of the material which we have presented (even if somewhat sketchily) above. Moreover (as remarked above) many of them discuss curves without giving a definition, or they define (incorrectly) a (space) curve to be the range of a continuous function in space, and a broken segment to be a point set, rather than a function. And, almost universally, they fail to mention such important matters as length invariance under congruence and the fact that all curve-length functions are similar.

While on this question of the definition of a curve, we should perhaps make a few comments concerning "plane curves" and "line curves", and some associated language problems. These are not serious, but they might confuse you if you are not forewarned about them. We have defined space curves to be continuous functions of a real number interval $[x,y]$ into S, and we have indicated that these functions may be identified with their graphs in the "four dimensional" product space, $R \times S$. In a natural way, such a graph is the range of a 1-1 continuous function from $[x,y]$ to $R \times S$.

If we wish to be consistent, we must define a plane curve to be a function from $[x,y]$ to a plane P, and this function (i.e., its graph) is the range of a 1-1 continuous function from $[x,y]$ to $R \times P$. This product can be corresponded with S in a straightforward way, so that we can picture a plane curve as a point set in space. To be further consistent, we would have to call a continuous function from $[x,y]$ to a line ℓ a line curve (a terminology which is rarely, if ever, used), and note the correspondence of line curves with certain subsets of $R \times \ell$. This situation is very familiar: if ℓ is given a coordinate system, we usually "picture" a line curve $f : [x,y] \to \ell$ by means of its "graph", $\{(t, f(t)) : t \in [x,y]\}$, in the "plane" $R \times \ell$. This graph is also referred to as a "plane curve". Actually the line curve f does determine a plane curve as we have used the term: namely the simple plane curve

$$F : t \to (t, f(t))$$

so this confusion in language is not too serious. But most of our plane curves (even most of the simple plane curves) cannot be derived in this way from the graphs of line curves, so it is important not to think of plane curves merely as the graphs of continuous real valued functions.

2-9 Extension of the Domain For Length Functions: Measure Theory

Under this heading we shall study measure functions of a rather special kind. The branch of mathematics known as measure theory is a direct descendent of the (mathematical) theory of length, and of the related simple theories of area and volume, which we shall discuss later. We shall restrict our attention mainly to the theory of linear measure, as applied to subsets of a line. This is a generalization of the theory of length as applied to line segments. Measure theory is largely a development of the present century, with its beginnings associated with such names as Borel and Lebesgue. It has a fundamental place in the study of real function theory, and in probability theory.

Roughly speaking, our objective is to extend the domain of the length functions from the set of all segments to include all, or as many as possible, of the subsets of a line. To keep matters simple, we will consider that our line has a coordinate system, and we "identify" points of the line with their coordinates. In other words, we consider the so-called "real line", whose points are the real numbers, and we seek to extend the domain D of the "natural" length function, $\mu : D \to R^+$, for which $\mu([a,b]) = |a - b|$. (Other length functions (related to μ by similarities) could be used as a starting point, leading to corresponding similarity-related measure functions. Such scale changes follow the familiar "similarity" pattern, and do not present any additional difficulties.)

We observe immediately that is not sufficient merely to ask whether the function μ can be extended to a larger domain: we can easily extend μ by giving arbitrary values to those subsets which are not segments. But this, of course, does not fit in with our idea of generalizing the notion of length: the generalization we seek should have some reasonable properties in relation to a "linear measure structure" of the set 2^R of all subsets of the real line R. In other words, we might approach the domain extension question in the same spirit as we approached the original problem of length measurement, and seek to first give a measure structure to 2^R, and then ask whether there are functions which preserve this structure. We shall not carry through such a treatment in detail, but this is the spirit in which we approach the problem.

We recall that the algebraic structure of the set 2^R of subsets of R, includes the partial order relation of inclusion; the commutative and associative operations of union and intersection, each of which distributes over the other; the notion of complement (we denote the complement (in R) of a set A, by A') and the related notion of difference. ($A - B = \{a : a \in A, a \notin B\} = A \cap B'$.)

If we think of our problem in physical terms, as the problem of assigning an appropriate "weight" to every subset of the line, we might well look for a function μ, from 2^R to the set R_0^+ (the nonnegative reals), which has some or all of the following properties (X, Y, W, denote subsets of R).

(i) μ is defined on the whole of 2^R.

(ii) $\mu([a;b]) = |a - b|$, for all $a, b \in R$.

(iii) If $X = \emptyset$, then $\mu(X) = 0$.

(iv) If $\mu(X) = 0$ then $X = \emptyset$.

(v) (Monotonicity). If $X \subset Y$ then $\mu(X) \leq \mu(Y)$.

(vi) If $X \subset Y$, $X \neq Y$ then $\mu(X) < \mu(Y)$.

(vii) If $\mu(X) < \mu(Y)$, and $W \cap X = W \cap Y = \emptyset$, then $\mu(X \cup W) < \mu(Y \cup W)$.

(viii) (Additivity). If $X \cap Y = \emptyset$, then $\mu(X \cup Y) = \mu(X) + \mu(Y)$.

(ix) Congruent sets should have the same measure, and similar sets should have corresponding "similar" measures; more precisely, if for real numbers a, b, $a \neq 0$, we define $[aX + b] = \{y : y = ax + b, x \in X\}$, then $[aX + b]$ is similar to X (with a similarity factor $|a|$), and $[aX + b]$ is congruent to X when $a = 1$ or -1; the suggested requirement is: $\mu([aX + b]) = |a|\mu(X)$.

(x) Property (viii) implies finite additivity; i.e., for any finite collection of pairwise disjoint sets, the measure of the union should be the sum of the measures. We might wish to extend this to imply countable additivity (a countable set is one which is finite, or which is denumerable; a denumerable set is one which can be put into 1-1 correspondence with the natural numbers); i.e., we might require that if X_n is a sequence of pairwise disjoint sets, then $\mu(\bigcup_n X_n) = \lim \sum_n \mu(X_n)$, at least when the series on the right is convergent. (As an example, you might think of the sequence of pairwise disjoint "left closed, right open" intervals;

$$X_1 = [0, \tfrac{1}{2}) = \{x : 0 \leq x < \tfrac{1}{2}\}$$

$$X_2 = [\tfrac{1}{2}, \tfrac{3}{4}) = \{x : \tfrac{1}{2} \leq x < \tfrac{3}{4}\}$$

$$\vdots$$

$$X_n = \left[\frac{2^{n-1} - 1}{2^{n-1}}, \frac{2^n - 1}{2^n}\right) = \left\{x : \frac{2^{n-1} - 1}{2^{n-1}} \leq x < \frac{2^n - 1}{2^n}\right\},$$

whose union is the "left closed, right open" interval $[0,1)$.)

We can easily see that these suggested properties are not all independent, but we shall not attempt to reduce them to an independent set of properties.

We examine some of the implications of these properties. First of all, even in terms of physical plausibility, we must exclude property (i): it does not seem reasonable to expect to assign a (finite) measure to such infinite sets as the real line itself, or to any ray of the line. More significantly, the assumption that a finite measure could be assigned to R would contradict the implication of properties (iii), (v) and (viii), as applied (for example) to the union of the disjoint unit segments $[0,1]$, $[2,3]$, $[4,5]$, ... , which is, clearly, a proper subset of R : No matter what real number $\mu(R)$ we might assign as a measure for R , there is some integer n (representing the measure of n disjoint unit intervals) which is greater than $\mu(R)$, and this would contradict monotonicity. Of course we could think of modifying properties (ii), (vi) or (viii), but it seems far more reasonable to discard property (i), and agree that we should not expect to include all subsets of R in the domain of our measure function.

We might now try to decide which subsets of R we should include in our domain, and which we should not include, but we will not be able to answer this question directly. As sets to be included, we would almost certainly want all finite unions of segments. (You might think that we should be more cautious, and include only the union of pairwise disjoint segments. But, if you observe that the union of two non-disjoint segments is again a segment, you will see that any finite union of segments can be expressed as a finite union of pairwise disjoint segments.) As sets to be excluded from the domain, you might think at first that, in addition to rays and half lines, we should include all unbounded sets. (A bounded set is a set which is contained in some segment of the line.) But this would turn out to be too drastic.

Without actually deciding what the domain must be, we can consider some properties which (now that we are agreed that the domain cannot be the whole of 2^R) the domain should have. We denote the (undetermined) domain by M , and refer (informally at present) to those subsets of R which belong to M as measurable sets (of the line). Some properties which we might like the collection M to possess, are the following:

M(i) M is closed under finite unions. (Clearly, this is related to the finite additivity of μ .)

M(ii) M is closed under differences of sets (and hence, as can be proved, under finite intersection as well.)

M(iii) M contains D , the set of all segments of R .

M(iv) M is closed under those point-transformations which are determined by similarity transformations of R . (This includes, of course, the congruences of R .)

A non-empty collection of sets which are subsets of some single set, and which satisfy M(i) and M(ii) , is called a Boolean ring (usually abbreviated to ring) of sets. A Boolean algebra of sets is a ring which satisfies the stronger condition of being closed under complementation. A σ-ring (σ-algebra) of sets, is a ring (algebra) of sets, which is closed under the operation of forming countable unions. Boolean rings and algebras are common structures for the domain of a measure function. If K is any (non-empty) collection of subsets of the line, then there are rings of sets which include K . (E.g., 2^R .) The intersection of all such rings is again a ring. It is called the ring generated by K .

Instead of worrying further, at this stage, about the domain, let us look at some of the implications of the other suggested properties for μ :

Property (ii), ($\mu([a,b]) = |a - b|$) , must be retained if our measure function is to be a generalization of the "natural" length function.

Property (iii), ($\mu(\emptyset) = 0$) , will follow, if we require finite additivity. (I.e., $\mu(X) = \mu(X \cup \emptyset) = \mu(X) + \mu(\emptyset)$. Hence $\mu(\emptyset) = 0$.)

Property (iv), ($\mu(X) = 0$ implies that $X = \emptyset$) , seems plausible from an empirical standpoint, but accepting it would imply that every set consisting of a single point would have a non-zero measure. Moreover, these single-element sets are all congruent, and the congruence property is one which we will want to retain if at all possible. Thus if the measure of each single-point set is p , $p > 0$, the archimedean property of the positive real numbers tells us that there is a positive integer m such that $mp > 1$. Thus for some finite subset W of m points in the interval [0,1] , we would have (from the finite additivity property (viii)) $\mu(W) > \mu([0,1])$. This is contrary to the monotonicity property (v). It seems more important to retain the additivity and monotonicity properties, so we give up property (iv), and agree that each set consisting of a single point shall have measure 0 . This implies (from the additivity property) that every finite set of points must have measure zero; and if we decide to retain countable additivity, it will imply that every denumerable set has measure zero. (This includes the set of all integers, and the set of all rational numbers, but, of course, not the set of all real numbers, or even the set of all real numbers in a given interval.) Notice that if we retain countable additivity and thus give measure zero to the set of integers, we will have an example of an unbounded set which is contained in our domain of measurable sets. (There are more significant examples of unbounded measurable sets.) You should observe also that if we assume property (ii) for μ, (a closed segment has its "natural" measure), and if we further assume that all finite sets belong to M and have measure zero, then properties M(ii), M(iii) imply that M also contains all open segments, and all "semi-open" segments, and that

$$\mu([a,b]) = \mu([a,b)) = \mu((a,b]) = \mu((a,b)) = |b - a| .$$

(As before, the use of a parenthesis instead of a square bracket indicates that the corresponding end point is excluded, leaving the segment "open" at that end.)

Property (v), (monotonicity), will follow from additivity, and the fact that if $A \subset B$, then $B = A \cup (B - A)$, and $A \cap (B - A) = \emptyset$.

Property (vi), ($X \subset Y, X \neq Y$ implies that $\mu(X) < \mu(Y)$), is clearly related to the discarded property (iv), and once we have agreed that some non-empty sets must be given measure zero, property (vi) must also be given up.

Property (vii), ($\mu(X) < \mu(Y)$ and $X \cap W = Y \cap W = \emptyset$ implies that $\mu(X \cup W) < \mu(Y \cup W)$, for measurable sets X, Y, W), is not an independent property: it is a consequence of (viii).

Property (viii) is the finite additivity property. This is, clearly, quite basic, and we retain it. Associated with this property (now that we have agreed that $M \neq 2^R$) is condition M(i), that M is closed under finite unions.

Property (ix) is the extension of the congruence and similarity properties which already apply on the sub-domain of segments. Our physical intuition suggests that we should try to retain these properties if possible.

Property (x) is the countable additivity property. If we retain the congruence property, and if we limit μ to finite values (see below for comment on this) then we cannot demand that M be closed under countable unions; (the whole real line can be expressed as a countable union of segments, and we have excluded R from M). Thus the countable additivity condition, if retained, will have to be expressed in such a form as:

if X_n is a sequence of pairwise disjoint measurable sets for which $\lim \sum_n \mu(X_n)$ exists, and whose union is in M, then

$$\mu(\bigcup_n X_n) = \lim \sum_n \mu(X_n) .$$

We now consolidate the modified properties suggested for μ and M, again making no attempt to give a minimal list of independent properties. (X, Y, W denote sets in the collection M.)

1. M is a Boolean ring of subsets of R.
2. M contains all (closed) segments, all open segments, all semi-open segments, and all countable sets.
3. M is closed under similarity transformations of R.
4. $\mu([a,b]) = \mu((a,b)) = \mu([a,b)) = \mu((a,b]) = |a - b|$.
5. $\mu(\emptyset) = 0$.
6. If $X \subset Y$ then $\mu(X) \leq \mu(Y)$.
7. If $X \cap Y = \emptyset$ then $\mu(X \cup Y) = \mu(X) + \mu(Y)$.
8. $\mu([aX + b]) = |a|\mu(X)$; $(a \neq 0)$
9. If X_n is a sequence of pairwise disjoint sets of M whose union is in M, then

$$\mu(\bigcup_n X_n) = \lim \sum_n \mu(X_n) .$$

We might now ask ourselves: are there rings M of subsets of R, and functions μ, which have these properties? The answer is that there are. For example, if M_0 is the smallest ring of sets (i.e., the intersection of all such rings) which satisfies conditions 1 , 2 , 3 , then there is a unique measure function on M_0 which satisfies the remaining conditions. We shall not attempt to prove this, but you can find the proof of a very similar result in Chapter 2 of [11].

To some extent this result is unsatisfactory, because it does not give any indication of how much further μ could be extended beyond this minimal ring M_0, nor does it give any information on the question of unmeasurable sets, which, from what we have said so far, you probably associate with the question of unboundedness. In order to give you a larger (but still elementary) picture of the subject of measure theory on the line, we shall outline the theory of Lebesgue measure as applied to bounded subsets of the line, giving only the definitions and the statements of the main results. (You might like to try to prove these for yourself: proofs can be found in many books on real function theory; e.g., [16].) The restriction to bounded subsets is not essential; but in order to remove it we would need to consider infinite values for μ, in the so-called "extended real number system"; and this additional complication is not essential to the ideas which we hope to convey.

We might now reconsider our objective in relation to bounded sets, and ask whether there exists a countably additive function, μ, with values in R_0^+, which is defined on the set B of all bounded subsets of the real line, which satisfies $\mu([0,1]) = 1$ (and hence, as can be proved, $\mu([a,b]) = |a - b|$), and which assigns the same value to congruent sets. (We are not dropping the monotonicity condition: as we saw earlier, it is a consequence of the other conditions.) The answer to this question is "no", so it becomes necessary to relax some of these conditions in order to get a measure function. If we relax the condition of countable additivity to finite additivity, then it can be shown that a suitable function does exist, but it is not unique. (This is quite a difficult theorem.)

[We remark, parenthetically, that the situation is similar for the corresponding generalized area problem for bounded subsets of the plane; but, somewhat surprisingly, the corresponding generalized volume problem (with finite additivity) for bounded subsets of space has no solution; there are no "generalized volume" measures defined on the domain of all bounded subsets of space, which satisfy the congruence condition, and which are finitely additive.]

For the applications of measure theory in mathematics, it turns out to be more fruitful to retain the condition of countable additivity and accept some restriction on the domain, rather than settle for finite additivity, which does not require any further restriction of domain. This leads to the Lebesgue theory of measure. In order to describe this theory, we need a few simple ideas concerning open and closed subsets of the real line. These sets can be considered as generalizations of open and closed intervals, with which you are already familiar.

A set X in R is an <u>open interval</u>, if there exist points $a_1, a_2 \in R$ such that $X = (a_1,a_2) = \{x : x \in R$, and $a_1 < x < a_2$ or $a_2 < x < a_1\}$. (Clearly, if $X \neq \emptyset$, then $a_1 \neq a_2$.)

A set Y in R is a <u>closed interval</u>, if there exist points $b_1, b_2 \in R$ such that $Y = [b_1,b_2] = \{x : x \in R$, and $b_1 \leq x \leq b_2$ or $b_2 \leq x \leq b_1\}$.

A set G in R is <u>open</u> if, given any $g \in G$, there exists an open interval, X, such that $g \in X \subset G$.

A set F in R is <u>closed</u> if its complement in R is open.

The following properties of open and closed subsets of the real line are left for you to prove as exercises:

(i) An open interval is an open set.

(ii) A closed interval is a closed set.

(iii) The empty set $\emptyset$, and the whole line R, are each both open and closed; $\emptyset$ is the only bounded set which is both open and closed.

(iv) Any union of open sets is open, and any intersection of closed sets is closed.

(v) Any finite intersection of open sets is open, and any finite union of closed sets is closed.

(vi) Every open set is the union of a countable collection of pairwise disjoint open intervals.

(vii) If G is open, F closed, then $G - F$ is open, and $F - G$ is closed.

Motivated by our earlier discussion, we can define a measure function m, on the collection B_0 of all bounded open sets, in an entirely natural way. We define $m(\emptyset) = 0$, and, for an open interval (a,b), we define $m((a,b)) = |a - b|$. If now G is any open set, then (from property (vi) above) $G = \bigcup_i \delta_i$, where $\{\delta_i\}$ is a countable collection of pairwise disjoint open intervals. Moreover, if G is bounded, it is easily shown that the sequence (σ_n) of numbers defined by:

$$\sigma_n = \sum_{i-1}^{n} m(\delta_i)$$

is bounded, and hence has a least upper bound, $\sup \{\sigma_n\}$. It seems natural to define

$$m(G) = \sup \{\sigma_n\} ,$$

(In discussion of measure questions this least upper bound is often denoted by $\sum_i m(\delta_i)$; if the number of "component intervals" δ_i is finite this is understood to be the usual sum, and if the number is countable it is the limit of the sequence σ_n of partial sums: i.e., the least upper bound of $\{\sigma_n\}$.)

It can be shown that this measure function, defined on the set of bounded open intervals, is countably additive and monotone on this domain. This domain is not, however, a ring, nor does it even include the closed intervals.

We can extend m so that its domain includes the collection B_C of all bounded closed sets, as follows: If F is a bounded closed set, let $[a,b]$ be the smallest (i.e., intersection of all) closed interval which contains F. We then use the fact that $(a,b) - F$ is open, to define

$$m(F) = |a - b| - m((a,b) - F) .$$

[Actually a and b must belong to F, so that $(a,b) - F = [a,b] - F$.] Because $B_O \cap B_C = \{\emptyset\}$, m is thus defined on $B_O \cup B_C$. It can now be shown that

(i) for any closed interval $[a,b]$, $m([a,b])$ is the "natural" length of the interval;

(ii) the measure of a finite union of pairwise disjoint closed intervals is, as we should expect, the sum of their lengths;

(iii) any bounded finite set of points has measure zero;

(iv) the measure of any bounded closed set is nonnegative;

(v) m is finitely additive on the collection B_C of all bounded closed sets, and countably additive on the collection B_O of all bounded open sets;

(vi) m is monotone on the collection of all bounded open or closed sets;

(vii) the measure of a bounded closed set F is the greatest lower bound of the measures of all of the bounded open sets which contain F.

(viii) the measure of a bounded open set G is the least upper bound of the measures of all of the bounded closed sets contained in G.

The above properties indicate that the function m might be considered to be a reasonably satisfactory extension of the idea of length to the collection $B_O \cup B_C$ of all bounded subsets of R which are either open or closed. But we observe that this collection is not yet a ring, and it does not even contain semi-open intervals, so we must try to extend the function m further than this. The idea which Lebesgue used to achieve this extension (an idea which is not too far removed from the idea of approximating a curve by broken segments) was to use the measures on closed and open sets to get "inner" and "outer" measures for every bounded set, and then to define a bounded set to be measurable if its inner and outer measures were the same. This idea makes all bounded open and closed sets measurable (and gives them the measures described above) but the collection of such bounded measurable sets turns out to be far larger than $B_O \cup B_C$, and it has the desired property of being a ring. We shall outline Lebesgue's treatment:

Let E be a bounded set of R. We define the (Lebesgue) outer measure, $m^*(E)$, of E to be the greatest lower bound of the measures of all bounded open sets G which contain E. (Cf. (vii) above.) I.e.,

$$m^*(E) = \inf_{G \supset E} \{m(G)\} .$$

We define the (Lebesgue) inner measure, $m_*(E)$, of E to be the least upper bound of the measures of all bounded closed sets F which are contained in E. (Cf. (viii) above.) I.e.,

$$m_*(E) = \sup_{F \subset E} \{m(F)\} .$$

The following properties of m^* and m_* are easily proved:

(i) m^* and m_* are nonnegative real valued functions.

(ii) The domain of each function m^*, m_*, is the set B of all bounded subsets of R; clearly this domain B is a ring.

(iii) If $E \in B_O \cup B_C$ (i.e., E is bounded, and either open or closed) then $m^*(E) = m_*(E) = m(E)$.

(iv) For every $E \in B$,

$$m_*(E) \leq m^*(E) .$$

(v) If E_1, E_2, are congruent bounded sets, then

$$m^*(E_1) = m^*(E_2) \text{ and } m_*(E_1) = m_*(E_2) .$$

(A corresponding property holds for similar sets.)

(vi) Both m^* and m_* are monotone; i.e., if E_1, $E_2 \in B$, with $E_1 \subset E_2$, then

$$m^*(E_1) \leq m^*(E_2) \text{; and } m_*(E_1) \leq m_*(E_2)$$

(vii) If E is bounded, and E is the union of a countable collection of bounded sets E_n (not necessarily pairwise disjoint) then

$$m^*(E) \leq \sum_n m^*(E_n) .$$

(viii) If E is bounded, and E is the union of a countable collection of pairwise disjoint bounded sets E_n , then

$$m_*(E) \geq \sum_n m_*(E_n) .$$

(ix) If E is contained in a bounded open set G , then

$$m(G) = m^*(E) + m_*(G - E) = m_*(E) + m^*(G - E) .$$

Remark: You might have noticed the absence of any additivity property for either m^* or m_* . We shall see later that neither function is even finitely additive on B , so, in this sense, neither is a satisfactory generalization of a length function to the whole of B .

This is far from a complete list of the properties of the inner and outer measure functions, but it should give you some idea of the main properties. As remarked above, m^* and m_* agree on $B_0 \cup B_C$, but (as we shall see below) they do not agree on the whole of B . Neither of these functions is finitely additive on B , but each is countably additive on $B_0 \cup B_C$.

If we restrict attention to the collection B_L of those bounded sets for which m^* and m_* give the same value, then the inequalities (iv), (vii) and (viii) above, combine to show that, if E is a bounded set which is the union of the countable collection of pairwise disjoint sets E_n from B_L, then

$$m^*(E) \leq \sum_n m^*(E_n) = \sum_n m_*(E_n) \leq m_*(E) \leq m^*(E) .$$

Thus all of the inequalities in this sentence are, in fact, equalities, and therefore E also belongs to B_L . Thus the functions m_* and m^* agree on B_L , and each is countably additive on B_L . This is the motivation for the definition of measurable set: a bounded set E is said to be (Lebesgue) measurable if $m^*(E) = m_*(E)$. For the collection B_L of bounded measurable sets, the following can be proved:

(i) B_L is a ring.

(ii) B_L contains $B_0 \cup B_C$, and hence contains the ring generated by $B_0 \cup B_C$.

(iii) The function

$$m : B_L \rightarrow R_0^+$$

defined by:

$$m(E) = m^*(E) = m_*(E) ,$$

for each $E \in B_L$ is countably additive on B_L ; it agrees with the "natural" length function on segments; and it gives the same value to congruent sets.

(iv) B_L includes all bounded sets which belong to the σ-algebra generated by $B_0 \cup B_C$; however, there are sets in B_L which cannot be obtained in this way.

(v) Every countable bounded set belongs to B_L , and has measure zero.

(vi) Every set of outer measure zero belongs to B_L , but a set of outer measure zero is not necessarily countable.

(vii) B_L is closed under similarity point-transformations of R , and similar sets have corresponding "similar" measures.

Remark: As stated earlier, the restriction of boundedness is not necessary if we are willing to use as value space the extended nonnegative real number system. (This includes a number ∞ with suitable properties.) Most modern treatments of measure theory proceed in this way. (An elementary treatment can be found in an appendix to Chapter 3 of [16].) If this is done, then it can be shown that the collection M_L of all Lebesgue-measurable sets includes (as a proper subset) the σ-algebra generated by the collection of all open and closed sets. The sets in this σ-algebra are called Borel sets: each Borel set can be obtained from open sets (and hence from open segments) by a countable number of the operations of union, intersection, and complementation.

In a certain sense, any measurable set can be approximated arbitrarily closely by open or closed sets, and a measurable set is "almost" a Borel set. More precisely, if E is measurable, and $\varepsilon > 0$, then

(i) there exists an open set G such that $G \supset E$ and $m(G - E) < \varepsilon$;

(ii) there exists a closed set F such that $F \subset E$ and $m(E - F) < \varepsilon$;

(iii) there exists a Borel set H such that $E \subset H$ and $m(H - E) = 0$.

This is about as far as it is reasonable to go in an elementary book such as this. We shall, however, give you an informal description of the "construction" of a non-measurable bounded set. The existence of such a set is equivalent to the fact that m^* and m_* do not agree on all bounded sets; and it enables us to prove that neither m^* nor m_* is even finitely additive.

For convenience we describe a non-measurable set in terms of the points of a circle in the plane. You can "translate" the description to give a non-measurable subset of the line, by "breaking" the circle and unrolling it onto a semi-open segment. Use of the circle merely avoids involvement with modular arithmetic and with piecewise congruence.

Let C be a circle with unit circumference (i.e., the length of the circle, considered as a simple curve, is 1). Let $\sim$ be the relation on the points of the circle, given by $P_1 \sim P_2$ if the length of either arc $\widehat{P_1P_2}$ is rational. (Clearly if one arc is rational then so is the other; we will refer to such points as being "rationally separated".) You can verify easily that $\sim$ is an equivalence relation on C. Let $\{C_\alpha\}$ be the set of equivalence classes determined by the relation $\sim$. (There is an uncountable infinity of such classes, but we do not need to use this fact.) Then

$$\cup_\alpha C_\alpha = C \text{ ; and } \alpha_1 \neq \alpha_2 \text{ implies that } C_{\alpha_1} \cap C_{\alpha_2} = \emptyset .$$

Let K_0 be a set of points of C which contains exactly one point from each set C_α (the existence of such a set K_0 depends on the so-called "axiom of choice", an axiom of set theory) and let K_q be the set of those points of C which are obtained from K_0 by a positive (i.e., counter-clockwise) rotation ρ_q, where q is rational, $0 \leq q < 1$, and where the rotation ρ_q is "measured" by the arc length q through which each point of K_0 moves. We list the following properties of the sets K_q, with brief comments on the proofs of these properties.

(i) The number of sets K_q is countably infinite. [The set of all rationals satisfying $0 \leq q < 1$ is a countably infinite set.]

(ii) If $q \neq r$, K_q is congruent to K_r. [K_q is the image of K_r, under a rotation of the plane of C about the center of C; a rotation of the plane is a congruence transformation.]

(iii) If $q \neq r$, then $K_q \cap K_r = \emptyset$; i.e., the sets in the collection $\{K_q\}$ are pairwise disjoint.

[If this were not true, then it would mean that some point P of C belongs to both K_q and to K_r. Hence there exist points P_q, P_r of K_0 which map into P under the rational rotations ρ_q, ρ_r, so that P_q, P_r are rationally separated from P, and hence from each other. Thus they belong to the same equivalence class. This contradicts the assumption that K_0 contains exactly one point from each equivalence class.]

(iv) $\bigcup_{q\in[0,1)} K_q = C$.

[Any point X of C belongs to exactly one of the equivalence classes of the set $\{C_\alpha\}$. Let $X \in C_\alpha$. Then from the definition of K_0, there exists exactly one point $Y \in C_\alpha$ (not necessarily different to X) such that $Y \in K_0$. The points X and Y belong to the same equivalence class, and hence they are rationally separated. Hence X is the image of Y under a corresponding rational rotation. That is, X belongs to one of the sets in $\{K_q\}$, and hence $\cup K_q = C$.]

We now have the circle C expressed as the union of a countably infinite number of pairwise disjoint sets, each of which is congruent to K_0. Without too much difficulty, we can "translate" back to the semi-open line segment [0,1) and exhibit this set as a countably infinite union of disjoint sets, each "essentially congruent" to the set K of [0,1), which corresponds to K_0. If m^* and m_* were to agree on K, then K would be Lebesgue measurable (i.e., belong to B_L) and so would all of the "essentially congruent" sets. Moreover each of these would have the same measure as K. It follows from the countable additivity of the Lebesgue measure on B_L, and from the archimedean property of the real numbers that either

(i) $m(K) = 0$, which implies $m([0,1)) = 0$;

or (ii) $m(K) > 0$, which implies $m([0,1)) > 1$ (and, in fact, greater than any real number).

Each of these contradicts the fact that $m([0,1)) = m([0,1]) = 1$. Hence the set K is not measurable.

It is not too difficult to prove the following statements about K, and about non-measurable sets generally:

(i) Every measurable set of positive measure, contains a non-measurable subset.

(ii) The complement of every non-measurable bounded set in an interval which contains it, is also non-measurable.

(iii) If K' is the complement of K in $[0,1]$, then $m^*(K) + m^*(K') > 1$, and hence $m_*(K) + m_*(K') < 1$; more generally, these inequalities hold for every non-measurable subset of $[0,1]$.

(iv) As a consequence of (iii), neither the inner nor the outer Lebesgue measures (on the collection B of all bounded sets) is finitely additive.

We remarked previously, that there exist (not uniquely) finitely additive, congruence-invariant, generalizations of length, which are defined on the whole collection B of bounded subsets of the line. To prove this statement and exhibit such a function is beyond the scope of this book, but we remark that, without too much difficulty, the following properties of any such function μ can be proved:

(i) $\mu(K) = 0$, where K is the non-measurable set constructed above.

(ii) If G is a finite union of open intervals, then $\mu(G) = m(G)$ (where m is the Lebesgue measure).

(iii) We have seen that every bounded open set G is a countable union of pairwise disjoint open intervals δ_k: for all such G, $\mu(G) \geq m(G)$.

(iv) If we add the additional requirement that, for every open set G, as in (iii), $\mu(G) \leq \Sigma(\delta_k)$, then μ agrees with the Lebesgue measure function, m on the collection B_L of all bounded Lebesgue-measure sets.

This is as far as we shall go in the treatment of linear measure on subsets of the line. From the brief sketch which we have given you can see that this is a fascinating and subtle subject, with results which are certainly not intuitively obvious. There are similar theories for area measures on subsets of the plane, and for volume measures in space, and these include results which are even more surprising. Beyond this, there are theories of linear measure for subsets of the plane, (first developed by Caratheodory in 1914) and theories of area measure for non-plane sets. Not surprisingly, the more general linear measure of Caratheodory applies to the set of rectifiable simple curves; (e.g., the graphs of continuous functions are measurable sets in this theory). Moreover the Caratheodory linear measure of the image of a rectifiable simple curve, coincides with its length. (Here we understand that both the linear measure function and the curve length function are extensions of the same length function for segments.)

Chapter 3

THE MEASUREMENT OF ANGLES, AREA, AND VOLUME

3-1 Introduction

Our treatment of angle measurement will be limited to a discussion of plane angles. The measurement of angles has many features in common with the measurement of length, but there are also differences. Three fundamental differences are:

(i) The range of an angle-measure function, on the most natural domain of "simple angles", is an open initial segment (or interval) $(0,r)$ of the positive real numbers, whereas the range of the most natural domain of elementary length-measurable objects (segments and their empirical counterparts) is, as we have seen, the whole set R^+ of positive reals.

(ii) Similar angles are also congruent, a statement which is, of course, not true of segments. This implies that the values of angle measurement functions should be unchanged under a similarity transformation of the domain. (In particular, corresponding angles in different scale models of the same object, are all congruent, and hence have the same measure.)

(iii) Angle measurement is related to length measurement, in that an angle measurement function (the so-called radian measure function) can be defined in terms of the concept of length. In this sense, if length measurement is taken as a primary (fundamental) measure then angle measurement is a secondary (derived) measure. (These ideas will be made precise later.)

In a book such as this, where we are taking an overall view of measure functions generally, as well as a detailed view of some particular measure functions, it is quite natural (in an approach to the subject of angle-measurement) to compare and contrast the situation with our prior treatment of length, and to make the treatments as similar as possible. If we do this, one of the first things that we notice is that there are surprising differences of language. These differences are not only historically interesting, but there is little doubt that they affect our thinking, and that they are responsible for some of the awkwardness, and difficulty, which is involved in the study of angles and angle measurement.

In our treatment of angle measurement we pay particular attention to these language questions, not because we are proposing that you immediately adopt a new language, but because we feel that an awareness of these language differences will help you to a better understanding of the real (i.e., non-linguistic) problems involved. Moreover, by using terms which are parallel to corresponding terms in the theory of length measurement, we can draw on much of the development already given for length.

Before becoming involved in details, we list some of the differences and similarities in the terminologies of length measurement and angle measurement. The discussion is rather informal, in the sense that we use a number of terms not yet formally introduced in this book.

1. In length measurement (and, similarly, in most measurement situations: e.g., area, mass, time, intelligence) we have a word, in this case "length", for the attribute being measured. As we have seen, there are certain difficulties connected with the possibility of actually defining these attributes, but the prevalence of such words suggests the probability that there are advantages in having names for the attributes. For angle measurement we might consider that "angularity" is such a word, but the mere existence of such a word is not enough; we must not only have it, but we must use it. (While we talk of the length of a stick, of a segment, of a curve, and so on, how often (if ever) do we refer to the angularity of an angle or of a rotation?) You might be inclined to argue that angle measure (or angular measure) is a suitable expression to substitute for length, but as you will recall from the previous chapter, there are good reasons for being able to distinguish between the length of an object and the length measure of an object. Thus we were able to define the length of an object to be the length-equivalence class to which it belongs (e.g., the congruence class for segments; the class under an empirically determined equivalence relation for "rods"), and we were able to define the length measure or linear measure of an object to be a set of ordered pairs, each consisting of a length function and its value on the given object. (Recall that any one of these ordered pairs determines the length, and this is what is involved when we say that, for example, the length of a particular object is 6 feet.) The parallel situation for angularity would be to define the angularity of an object to be a

certain equivalence class, and the angularity measure, or angular measure, of an object to be a set of ordered pairs, each consisting of an angular measure function, and a value. In what follows we use the terms "angularity" and "angular measure" in this way, so as to be able to bring out the similarity between length and angularity, and between linear measure and angular measure.

2. We have no single name for an object in the common domain of empirical length measurement functions; for simple mathematical length-measurement functions, the objects are segments, but when the domain is extended the objects may be curves, or they may be rather general point-sets. In the case of angularity-measurement, whether empirical or mathematical, an object in the domain is almost always called an angle. Sometimes we use the word "rotation", but then we usually complicate matters by talking about the angle of rotation, and then referring to the measure of this angle. Clearly there is no reason why we should not regard rotations as being elements of the domain of angular-measure functions, and talk directly about the angular measure of a rotation. Angles (i.e., simple angles, see below) would then correspond to certain simple rotations, just as segments correspond to certain simple curves.

3. The related terms "length", and "distance" for linear measures have their parallels for angular measures. We shall use the term "angular distance" as corresponding to "distance". At the simplest level, the objects which constitute the domain of an "angular distance" function are, of course, pairs of rays with a common end point.

4. Just as there are good reasons for separating the related ideas of distance function and coordinate function, there would seem to be good reasons for doing something similar with respect to angles. Thus, restricting our attention to the set of all coplanar rays with a fixed common vertex, we have an angular distance function on pairs of rays, and related angular coordinate functions. There is no need (at least at the elementary level) to introduce "negative angles". Negative numbers enter through the use of a symmetric interval of real numbers (e.g., $[-\pi,\pi]$) as the range of an angular coordinate function, but there are no negative numbers in the range of an angular measure function. This is parallel to the use of the set of all real numbers (but not just the positive reals) for coordinate functions on the line, and the relationship between coordinate functions and distance functions is quite similar for both angular and linear measurements. From a

mathematical point of view, this similarity is directly related to the separation properties of lines and planes: a line with a point removed is the union of two disjoint convex sets (half-lines), while a plane with a line removed is also the union of two disjoint convex sets (half-planes).

Of course we need to consider both positive and negative numbers in a discussion of the measurement of "directed angles" and "sensed rotations", but these bear the same relationship to the notions of angle and rotation as oriented (directed) segment bears to segment, and there is no more reason to introduce negative angles (with negative angularity), than there is to introduce negative segments (with negative length). In fact, as we pointed out in Section 2-7, the idea that is really involved here is that of a vector. The set of all directed segments on a line, with the same initial point, is a 1-dimensional vector space over the real numbers. For sensed rotations the situation is similar: the set of all sensed plane rotations at a point, with a common initial ray, is also a 1-dimensional vector space over the reals. Moreover, by introducing appropriate equivalence relations, the equivalence classes of all directed segments on a line, and of all sensed plane rotations at a point, become 1-dimensional vector spaces over the reals. In discussions of "angles of any size" it is these vector rotations, or directed angles, that we are concerned with, and the process of giving angular measures to these "angles", using the full set of real numbers as range, is nothing more than the process of setting up isomorphisms from the vector space of sensed plane rotations to the vector space of real numbers, a vector space which is easily shown to be isomorphic to every 1-dimensional vector space over the reals.

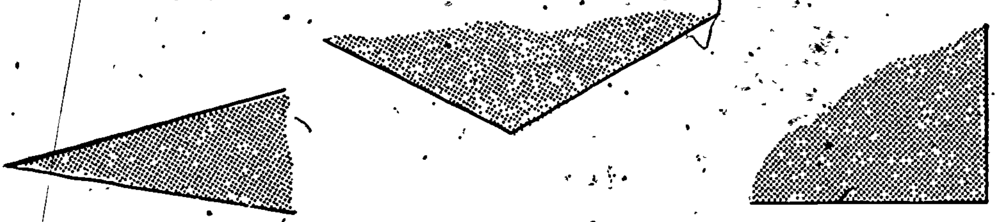
To sum up: If we are to preserve the parallelism of linear and angular measures, we should consider rotations as generalized angles, and hence as elements of the domain of angular-measure functions, but we should not consider sensed rotations as elements of this domain. Sensed rotations, and directed angles, are better considered as vector quantities, for which appropriate vector measure functions are mappings into vector spaces of appropriate dimension.

3-2 Angle Measurement From An Empirical Standpoint

You will have noticed that we have not yet given a definition of angle. From an empirical standpoint there is no more need for the use of a single defined term "angle", for an element in the domain of an angularity-measurement function, than there is for the use of a single word in the corresponding situation for length. You will recall that we did introduce the word "rod", which was used as abbreviation for "element of the domain of a simple empirical length function". In this section we will use the noncommittal word "wedge", in much the same way. In the following section, in which we deal with the corresponding mathematical ideas, we will need to define "angle", just as we needed to define "segment" in the formal treatment of linear measurement in a geometrical context.

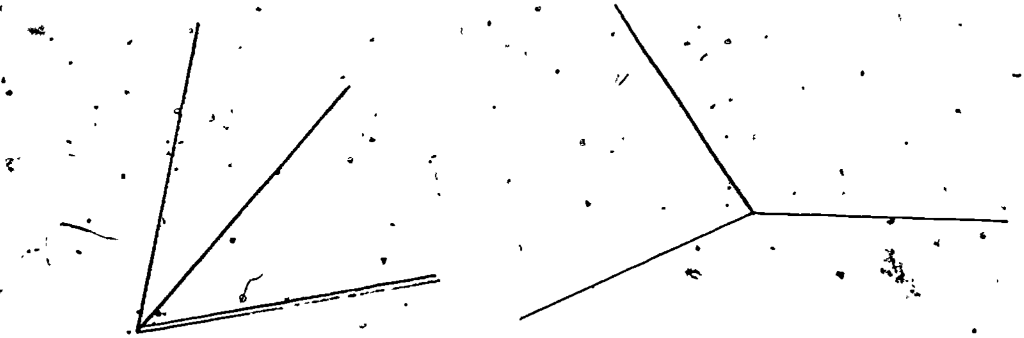
From an empirical standpoint, we must be able to recognize a set A of objects which possess "angularity", and we have to establish empirical procedures for comparing the "amounts of angularity" which these objects have. ("Angularity" will not need to be defined.) These procedures will lead to the establishment of a certain angularity structure on the set A, or on a set $\bar{A}$ of equivalence classes of A, and we will look for functions which map $\bar{A}$ isomorphically into the positive reals.

In the set A of objects which possess angularity, we might include such things as wedges, ordered triples of non-collinear points, pairs of rods which are joined at one end, certain pencil marks on pieces of paper, and certain chalk marks on chalkboards. We use the vague word "wedge", to describe any of these objects, and, at least initially, we assume that our wedges are bounded by pairs of "rays", which are not collinear (and hence, of course, not coincident). We shall have more to say about this exclusion in the mathematical discussion of angles and angularity. We also restrict our notion of wedge, so that each point of a wedge lies either on, or on one side of, the line determined by one of its boundaries. If we wish, these restrictions can be modified later to extend our domain. For convenience we can picture elements of our domain like this:



As in the case of length measurement, we assume that there is an empirical procedure for comparing our wedges, and that this procedure is "perfect" in the sense that it enables us to establish an equivalence relation ("same angularity", denoted by $\sim$) and an inequality relation ("less angularity", denoted by $<$) on the domain, and that the relation $<$ carries over to the set $\tilde{A}$ of equivalence classes of A .

We assume next that there is an empirical join operation for wedges (a binary operation which may be iterated), and that joins are plane objects, as suggested by the diagrams.



This join operation (*) has the property that the join of two wedges need not be (or correspond to) a wedge. We resist the temptation to modify our notion of wedge, so as to make the set of wedges closed under the join operation, but we do enlarge set A to a new set (B say) which includes not only the set A of all wedges, but also the set of all finite joins of wedges. We assume that the empirical equivalence and inequality structure can be extended to this enlarged domain. (It does not require too much imagination to see how this might be done empirically. If we think of our wedges as rigid pieces of thin cardboard, in comparing one finite join of wedges with another we are still involved in a visual or tactile comparison process, and the fact that our joins might "spiral around" many times doesn't complicate the comparison procedure in any essential way: we can count directly the number of times the join spirals around. But, as we shall see, this empirical simplicity does not extend readily to the mathematical theory.)

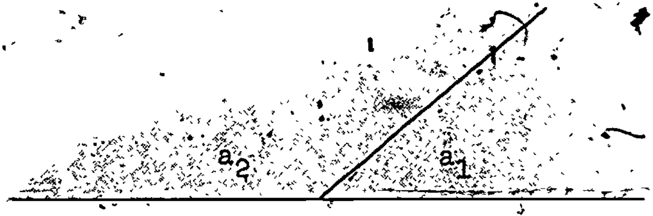
Without stretching our imaginations too far, we can assume that our empirical procedures suggest a structure for $\tilde{B}$ (the set of equivalence classes of our enlarged domain B under the extended equivalence relation) which is, exactly as in the case of length and rods, a densely ordered, abelian, archimedean semigroup. However, as far as our subdomain of wedges is concerned, there are significant differences: the set A of wedges is not closed under the join operation, and the archimedean property has to be modified ~~to something~~ like: "given any wedge a, there is a positive integer n such that the iterated join na is not a wedge." The order relation on the subset $\bar{A}$ is still dense, and the set of wedges has no least element and no greatest element.

We can establish an empirical structure-preserving measure function from the extended domain $\mathcal{B}$ to the positive reals by either of the approaches previously used for setting up similar functions for length measurement. The "unit" approach (in which we first select any wedge as a unit) has the minor advantage that we could describe it on the domain of wedges without going out of the domain. In the alternative approach we first set up a ratio structure, associating a positive real number with each ordered pair of elements of the domain. This involves the comparison of arbitrary integral multiples of wedges; thus the modified archimedean property ensures that we could not carry out this procedure on the restricted domain of wedges.

Clearly, the situation is very similar to that which we discussed for the empirical measurement of length. Let us assume that we obtain, by this empirical procedure, a set of angular measure functions, which, with respect to the structures involved, are isomorphisms. Let

$$\eta : \mathcal{B} \to R^+$$

be such a function. Then η will have the (empirical) property that if a_1, a_2 are any two wedges with the property that, when we form the join, the remaining rays are collinear (this implies, of course, that we have a suitable test for collinearity),



then

$$\eta(a_1) + \eta(a_2) = p$$

where p does not depend on a_1 and a_2. Moreover the function η

restricted to $\bar{A}$, will map the incomplete structure of $\bar{A}$ isomorphically into the interval $(0,p)$; p will be (as far as can be determined empirically) the least upper bound of the range of the restricted function $\eta|\bar{A}$; and 0 will be the greatest lower bound.

As in the case of length, we can show easily that the functions obtained by composing these empirically determined functions with positive similarities of R^+ (intuitively: multiplying all values by the same positive real number) are also structure preserving. The complementary question, "must all structure-preserving functions be related in this way?", can only be answered affirmatively if we make further assumptions (such as the assumption that the functions η and $\eta|\bar{A}$ are onto R^+ and $(0,p)$ respectively) and clearly (as in the case of length) there is no possible way of deciding empirically whether such assumptions are, in some sense, necessary. For empirical purposes the set of all rational numbers in an open interval $(0,p)$ will be quite adequate for the measurement of angularity on the domain of wedges. In the next section we shall see that the corresponding question for angular measurement in (synthetic) geometry is related to the Cantor-Dedekind completeness property: if we assume this property as a postulate, then the angle measurement functions for simple angles will be onto an open interval $(0,p)$ of the positive reals.

[Cf. Exercise 2-2.5, which showed that if p , q , are positive reals, then the only isomorphism of the open interval of reals $(0,p)$ onto the open interval $(0,q)$, which preserves (where relevant) addition, also preserves order; and that it is the similarity transformation $x \to \frac{q}{p}x$. This result can be generalized in the manner of Exercise 2-2.19, and the remark following that exercise.]

The structure of the domain A of wedges can be established empirically in still another way, by using a procedure that has much in common with the method used for establishing a length structure on the set of broken segments. We shall describe this procedure, because it turns out to correspond closely to the simplest way of handling the comparable question in the context of synthetic geometry. We anticipate slightly, by commenting that one of the main difficulties in handling the question of angle measurement in synthetic geometry, is that the previously suggested empirical procedure for forming and comparing wedge joins is awkward to formalize without using the concept of rotation; and this concept is surprisingly hard to introduce into the formal structure of geometry. While we wish to consider this intuitively simple concept eventually, it is desirable to avoid it in a first approach to the problem of angle measurement in formal geometry. This can be done if we mimic the empirical procedure which we describe below. If you are able to follow the empirical procedure, it is not hard to "translate" the various steps in order to obtain the corresponding formal mathematical procedure.

Sums of Wedges. If you refer back to the method which we used in Section 2-8 to set up an equivalence relation on the set of broken segments, and to obtain a structure for this set as an ordered semigroup, you will see that the fact that the constituent segments in a broken segment were joined end-to-end, really didn't play any role. Because of the fact that we were working with equivalence classes, we could just as well have considered that we were working with objects which were just finite sets of segments (not necessarily joined), and that we defined such sets to be equivalent if there was a piecewise congruence of their elements, allowing for decomposition. On such a domain, a formal "join" of two members (each a finite set of segments) could be defined by using the union. Questions of disjointness could be handled without real difficulty, and we would obtain a corresponding formal join of equivalence classes. This operation would be commutative and associative, it would be properly related to an order relation, and so on.

Let us describe a similar procedure for wedges, keeping in mind that the big advantage in working with extended joins of wedges (as before) or with "formal sums" of wedges, is that we obtain an enlarged domain, which is closed under the relevant binary operation. Thus we obtain all of the advantages of working with a semigroup, rather than with an "incomplete" structure, in which joins of wedges need not exist. In this way, the treatment of angular measurement on the enlarged domain becomes virtually identical (except for language) with our treatment of length measurement, leading to angular measure functions on the enlarged domain. From these the necessary measure functions for wedges (and angles) are obtained by merely restricting the domain.

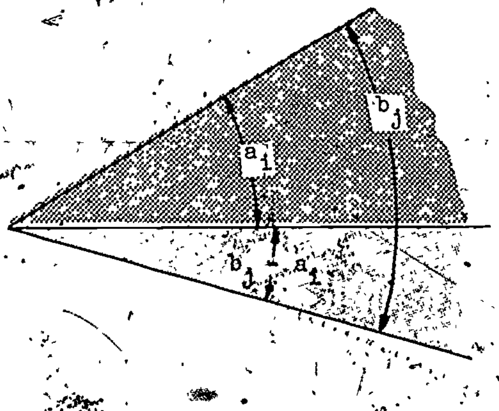
Let a_i, b_j, ..., denote wedges. By a formal sum of wedges, we mean simply a finite set of wedges.

$$W = \{a_1, a_2, \ldots, a_n\} .$$

Such formal sums are, of course, equal, if and only if they contain the same elements. Because a formal sum is simply a set, the same wedge cannot be repeated in a formal sum. But equivalent wedges may be contained in the same formal sum, and this is sufficient for our purposes.

If $a_i \in W$, and if $a_i \sim a_i' * a_i''$, with neither a_i' nor a_i'' in W, then the formal sum, W_1, which is obtained from W by substituting a_i' and a_i'' for a_i, is called a <u>decomposition</u> of W. Any formal sum obtained from W by a finite number of such steps, is called a <u>finite decomposition</u> of W. We define two formal sums W_1, W_2, to be equivalent $(W_1 \approx W_2)$, if there are finite decompositions W_1', W_2', of W_1, W_2, respectively, such that there is a 1 - 1 correspondence of the elements of W_1' and W_2', with corresponding elements equivalent as wedges. If a simple wedge is considered as a formal sum, then $\approx$ agrees with $\sim$, in the sense that, for wedges a, b, $a \sim b$ if and only if $\{a\} \approx \{b\}$. We assume that $\approx$ is (empirically) an equivalence relation which, in a sense, "extends" the equivalence relation for wedges. We can think of an empirical procedure for comparing two formal sums $W_a = \{a_i\}$, and $W_b = \{b_j\}$, as follows:

If there are any common terms in W_a, W_b, discard these first. Let W_a', W_b' be the residual sums. Take any wedge, a_i, from W_a', and any wedge, b_j, from W_b', and compare them as wedges. Exactly one of the relations: $a_i \sim b_j$, $a_i < b_j$, $b_j < a_i$ holds. If the first relation holds, discard a_i and b_j, pull a second wedge from each set, and compare them, etc.; if either of the other relations holds (say $a_i < b_j$), find a b_j' such that $a_i * b_j' \sim b_j$, and $b_j' \notin W_b'$. (We may think of b_j' as equivalent to "$b_j - a_i$"; in our mathematical treatment, the congruence postulates for angles will provide us with such a "subtraction", and this is why this simple procedure carries over successfully to the mathematical context.)



If W''_a is the formal sum obtained from W'_a by removing a_i, and W''_b the formal sum obtained from W'_b by either removing b_j (if $a_i \sim b_j$) or by replacing b_j with b'_j (if $a_i < b_j$), then the number of terms in W''_a is one less than the number of terms in W'_a, and the number of terms in W''_b is either the same, or one less than, the number of terms in W'_b. Thus at least one formal sum has been reduced (in its number of terms) and the other has not increased. Hence, if we continue drawing, comparing, and "subtracting" if necessary, the process must terminate. If W_a and W_b each reduce eventually to a single wedge, and these are equivalent as wedges, then W_a and W_b will be equivalent as formal sums.

An ordering ($\ll$) between formal sums can be introduced in the usual way ($W_a \ll W_b$ if there is a W'_b, such that $W_b \approx W'_b$ and W_a is equivalent to a proper subset of W'_b) and the same empirical procedure can be used to establish this relation. We assume that empirical evidence justifies the assumption that $\ll$, yields an order relation on equivalence classes of formal sums.

The definition of a join operation for equivalence classes of formal sums, is carried out in the obvious way. If $\tilde{\tilde{W}}_1$, $\tilde{\tilde{W}}_2$, are such equivalence classes, select W'_1, W'_2, from $\tilde{\tilde{W}}_1$, $\tilde{\tilde{W}}_2$, respectively, with $W'_1 \cap W'_2 = \emptyset$, then define $\tilde{\tilde{W}}_1 * \tilde{\tilde{W}}_2$ to be the equivalence class of $W'_1 \cup W'_2$. (This should remind you of the way in which we used "disjoint unions", when discussing numerosity measurement.)

If we consider the set of wedges to be a subset of the set of formal joins, then, on this subset, the equivalence relations $\sim$, and $\approx$ agree, and so do the order relations $<$, and $\ll$, and the join operations (where defined for wedges). The set of equivalence classes of formal sums is (empirically) an ordered abelian semigroup, in which the set of equivalence classes of wedges is isomorphically imbedded. This is what we set out to achieve: the establishment of suitable angular measure functions can now be carried out as before.

This is as far as we need to go in this direction, as our main objective was to describe (in informal language of the empirical situation) a process which is virtually identical (but more awkward to describe) in the formal mathematical domain. We return now to our general discussion of angular measure on wedges.

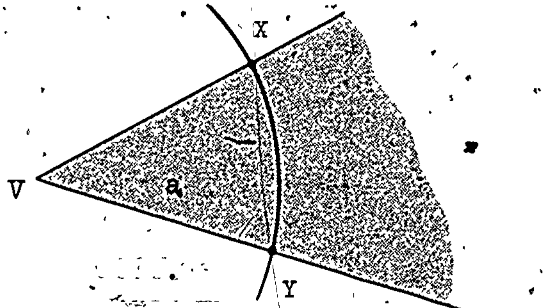
One interesting property of angularity measurement concerns the question of units: Because any open interval $(0,p)$ of positive reals is a satisfactory value space, if $p < 1$ then there will not be a "unit"; i.e., no element of the domain of wedges will have value 1 under the corresponding angular-measure function.

Except for the necessary modifications resulting from the restriction of the range to an open interval, all of the results for length measurement have their analogs for angularity measurement. The most important of these is the relationship involved in the question of "scale change". If η_1 and η_2 are two angularity-measure functions, then for some $k > 0$, $\eta_2 = k\eta_1$, and for every $a \in A$,

$$\frac{\eta_2(a)}{\eta_1(a)} = k .$$

This means that the relationship between two such functions is fully determined if we know the value of each function on one element $a_0 \in A$. In this case, for any $a \in A$, we have $\eta_2(a) = k\eta_1(a) = \frac{\eta_2(a_0)}{\eta_1(a_0)} \cdot \eta_1(a)$.

As you are aware, there are important connections between linear measures and angular measures. If we take any circle with center at the vertex of a wedge a, and use any (empirical) length function λ, then the circle meets the boundary of the wedge a at points X and Y,



and the ratio $\frac{\lambda(\widehat{XY})}{\lambda(\overline{VX})}$ ($\widehat{XY}$ denotes the arc of the circle) is independent of our choice of circle and of our choice of linear measure function. Moreover the function

$$\mu_\pi : A \to R^+$$

defined by

$$\mu_\pi(a) = \frac{\lambda(\widehat{XY})}{\lambda(\overline{VX})}$$

coincides with one of our earlier established angular measure functions, and has range $(0,\pi)$. In other words this function μ_π is itself a suitable angular measure function, meeting our requirement of preserving the structure of A. (This is the well-known radian measure function.) It follows that any angular measure function differs from μ_π by a positive similarity.

We emphasize that all of the above statements concerning the radian measure function have to be considered (at this stage) as having only empirical justification. In synthetic geometry there is a corresponding situation for angular measure functions, but, in that context, the corresponding assertions must be deduced from the axioms of synthetic geometry.

There is a connection between the relationship of angular and linear measures, and the subject of "scale models". In a scale model, the linear measures of an object and its scale model differ by a fixed "scale factor". Consequently the radian measures (and hence the angular measures under any fixed angular measure function) of corresponding angles on the object, and on a scale model, are the same.

3-3 Angular Measure In Mathematics

You will recall that, in the discussion of linear measure in formal mathematical systems, we had two questions to consider at the outset: What system or systems should we consider; and what should be our initial domain? The first question involved the choice of a geometry, and we considered three possibilities: classical synthetic geometry, metric geometry (as proposed by Birkhoff, and used by SMSG and others), and cartesian (coordinate) geometry. In metric geometry the distance/length functions were postulated, and in cartesian geometry they were easily established.

But in synthetic geometry we had to do quite a bit of work in order to establish the existence of suitable length/distance functions, and we discovered that if the corresponding postulates (the distance and coordinate postulates) in metric geometry were to be established as theorems in synthetic geometry, then we needed to strengthen the synthetic geometry by the addition of the archimedean postulate, and the Cantor-Dedekind (completeness) postulate. For angle measurement the situation is roughly the same, but perhaps the need to establish the existence of angular measure functions (suitably behaved with respect to congruence and "angle addition") in synthetic geometry is not quite so important, for the following reason: Using linear measures only, we can start with (augmented) synthetic geometry, prove the existence of distance and coordinate functions for lines, and go on to establish coordinate functions for planes and space, using only the synthetic congruence postulates for angles (see below). In the resulting cartesian geometry, we can develop a theory of curve length as discussed in the last chapter, and the existence of suitably behaved angular measure functions can then be established by using the theory of curve length, as applied to simple circular arcs. (This is, of course, the way in which radian measure functions are formally defined; see later.) We must, of course, be careful that, in such a development of linear measures for curves, we have not used the existence of angular measures, but this does not present any serious difficulties.

One advantage of the use of the arc-length approach in the establishment of the existence of suitable angular measures, is that it is a natural vehicle to use in discussions of the extension of the domain to include angular measures for rotations of various kinds. We shall have more to say about this below. But there is something not quite satisfactory about this approach: we feel, intuitively, that the empirical procedures for angular measurement should have their mathematical counterparts, and that we should be able to establish, directly, the existence of suitably behaved angular measures in synthetic geometry. This can, in fact, be done, but (as you might expect on the basis of our experience with the corresponding problem for linear measure) it is by no means simple. We shall give an outline only of such a treatment.

Angular Measurement In Synthetic Geometry. For linear measurement we decided that our initial domain should consist of segments. In the same way, for angular measurement our domain will consist of angles. This is not as simple a statement as it seems, because, although there is universal agreement on a definition for "segment" there is no such agreement for "angle". In synthetic geometry, the elementary idea of "angle" can be conveyed in many ways. For example:

(i) a pair of rays with a common endpoint, excluding collinear rays;

(ii) the point set union of such a pair of rays;

(iii) the union of ~~two~~ rays as in (i), together with all points "between" them;

(iv) as in (i), but with rays replaced by segments;

(v) an ordered triple of non-collinear points;

(vi) a set of rays, consisting of a pair of non-collinear rays, as in (i), together with all rays (with the same endpoint) which lie "between" them;

(vii) the union of all rays as in (vi);

(viii) a suitably defined 1-1 bicontinuous function from an interval of real numbers to an open semi-circle.

Remarks:

1. From the point of view of angular measurement, there would be many advantages in adopting (vi) as our definition. This would (in effect) make an angle a "segment of rays". (The restriction of non-collinearity must be made, because betweenness is not defined for collinear non-coincident rays; see below.) With (vi) as definition, we could parallel our language and treatment for length measurement on segments, except where forced to diverge by the really essential differences. Moreover it would provide a simple and direct way to generalize the notion of angle, by analogy with the situation for curves. (We return later to this idea.)

2. Possibilities (iii) and (vii) describe the same point set. This point set corresponds to our vague word "wedge", and seems to describe most closely the simple physical idea of an angle.

3. The exclusion of coincident rays is in keeping with our original definition of a segment. There is no difficulty in later extending the definition of angle to include such "zero angles", or "null rotations". Zero angles are not needed in elementary geometry, and excluding them from the first stage of a discussion of angular measure permits the

use of a suitably modified archimedean property for pairs of angles, and allows us to state results about angles without the need to consider special cases.

4. The exclusion of collinear non-coincident rays, permits us to introduce a suitable notion of betweenness for rays with a common end point, and to talk unambiguously of the interior of an angle.

5. Possibility (v) corresponds to the way that the angles of a triangle are determined, and to the usual method for naming angles; e.g., /ABC.

6. Possibility (viii) is close to the one which we will use (and generalize) for a discussion of rotation, but it requires that we find some way of making precise the notion of continuity which is involved. This cannot be done (in synthetic geometry) in an elementary way.

Notwithstanding the possible advantages of (vi), (ii) seems to be the most common definition in current use. We therefore define an angle to be the union of two different non-collinear rays. This is consistent with the SMSG approach, and with the treatment in the important reference [14]. The main point to remember is that each of the other suggested possibilities determines a unique angle, so that if we use any of these alternate descriptions, there will be no doubt which angle we are talking about.

If our treatment of angular measure is to parallel that for linear measure, we should first look at those postulates and properties which were particularly important for the theory of linear measure, and see what their counterparts for angular measure would be. Among these counterparts are the following:

There is a postulated relation of betweenness for (some) ordered triples of rays; if $\overrightarrow{OB}$ is between $\overrightarrow{OA}$ and $\overrightarrow{OC}$, we write $\overrightarrow{OA} - \overrightarrow{OB} - \overrightarrow{OC}$. Before giving the properties of betweenness we introduce a useful abbreviation: rays with a common end point are said to be co-halfplanar, if they are coplanar, and if there is a line through the common endpoint, and in the common plane, such that the interiors of all of the rays (i.e., the rays except for their endpoints) are contained in (the same) one of the halfplanes determined by the line. (Any two rays whose union is an angle, are co-halfplanar; a fact which you may prove from the axioms below.)

The required properties for betweenness of rays are (cf. 2-5 and note how closely these correspond to the congruence/betweenness postulates for segments and points):

BR-1. If $\overrightarrow{C_1X} - \overrightarrow{C_2Y} - \overrightarrow{C_3Z}$, then $\overrightarrow{C_1X}$, $\overrightarrow{C_2Y}$ and $\overrightarrow{C_3Z}$ are co-halfplanar. (In particular, $C_1 = C_2 = C_3$.)

BR-2. Given three different co-halfplanar rays, exactly one is between the other two.

BR-3. The relation is symmetric in the sense that $\overrightarrow{CX} - \overrightarrow{CY} - \overrightarrow{CZ}$ if and only if $\overrightarrow{CZ} - \overrightarrow{CY} - \overrightarrow{CX}$.

BR-4. If $\overrightarrow{CX}$, $\overrightarrow{CY}$ are two rays, then there are rays $\overrightarrow{CZ}$, $\overrightarrow{CW}$, such that $\overrightarrow{CX} - \overrightarrow{CZ} - \overrightarrow{CY}$, and $\overrightarrow{CW} - \overrightarrow{CX} - \overrightarrow{CY}$.

BR-5. Any four co-halfplanar rays can be named in an order $\overrightarrow{CX_1}$, $\overrightarrow{CX_2}$, $\overrightarrow{CX_3}$, $\overrightarrow{CX_4}$ such that $\overrightarrow{CX_1} - \overrightarrow{CX_2} - \overrightarrow{CX_3} - \overrightarrow{CX_4}$, where the notation means that all of the implied betweenness relations for triples are true.

There is a relation of congruence on the set of all angles. It has the properties (cf. Section 2-5):

CA-1. Congruence is an equivalence relation on the set of all angles. (We denote angle congruence by the symbol "$\cong$".)

CA-2. Given an angle $\angle AOB$, and a halfplane bounded by a line $\overleftrightarrow{CX}$, there is a unique ray $\overrightarrow{CZ}$ whose interior lies in the given halfplane, and such that $\angle XCZ \cong \angle AOB$.

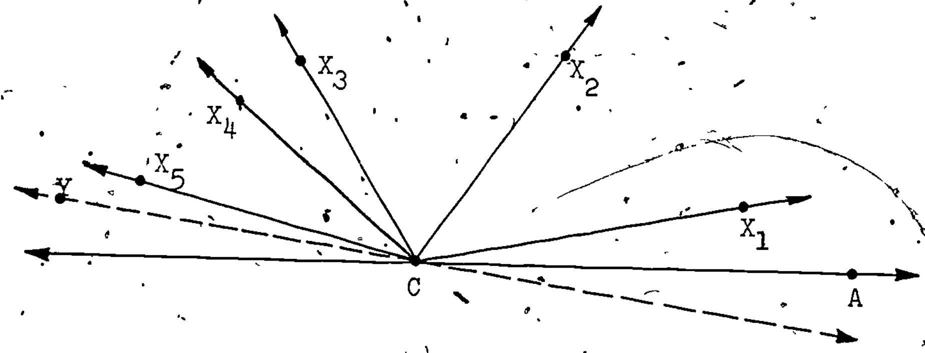
Congruence of angles and betweenness for rays are related by the properties (cf. Section 2-5):

CB-1. If $\overrightarrow{C_1X_1} - \overrightarrow{C_1Y_1} - \overrightarrow{C_1Z_1}$ and $\overrightarrow{C_2X_2} - \overrightarrow{C_2Y_2} - \overrightarrow{C_2Z_2}$, and if $\angle X_1C_1Y_1 \cong \angle X_2C_2Y_2$, then $\angle Y_1C_1Z_1 \cong \angle Y_2C_2Z_2$ if and only if $\angle X_1C_1Z_1 \cong \angle X_2C_2Z_2$.

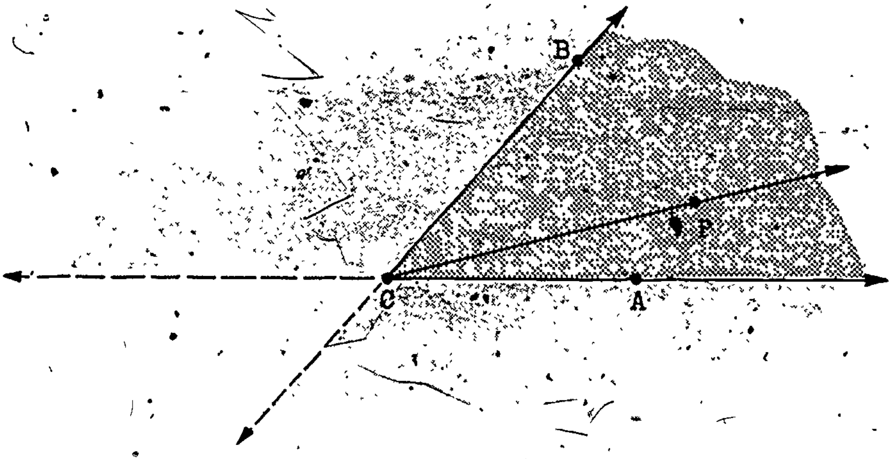
These properties are all valid in synthetic geometry, but it is not necessary to take all of them as postulates. For example, the relation of betweenness for rays can be defined by: $\overrightarrow{CX} - \overrightarrow{CY} - \overrightarrow{CZ}$ if and only if for every pair of interior points X_1, Z_1 of $\overrightarrow{CX}$, $\overrightarrow{CZ}$, respectively, $\overrightarrow{CY} \cap \overline{X_1Z_1} \neq \emptyset$; and its properties can be deduced. The congruence properties, and property CB-1 , are taken as postulates. We have listed together all of these betweenness properties for co-halfplanar rays, and congruence properties for angles, in order to show how close they are to the corresponding betweenness properties for collinear points, and the congruence properties for segments.

Remarks.

1. One useful consequence of properties BR-4 and BR-5 , is that if we add one of the "bounding" rays to a finite set of co-halfplanar rays, then the augmented set is still co-halfplanar (but not, of course, with respect to the same half-plane: see diagram). In the diagram, $\overrightarrow{CA}$ is the initial half-plane boundary, and $\overrightarrow{CY}$ is the new boundary when $\overrightarrow{CA}$ is added to the finite set of rays, $\{\overrightarrow{CX_1}, \ldots, \overrightarrow{CX_5}\}$.



2. In geometry, the notion of the interior of an angle is useful: the interior of the $\angle ACB$ is the intersection of the two half-planes which are determined by $\overrightarrow{CA}$, $\overrightarrow{CB}$, and which contain B , A , respectively. It is easy to show that P is in the interior of $\angle ACB$, if and only if $\overrightarrow{CA} - \overrightarrow{CP} - \overrightarrow{CB}$, and hence that the interior of the angle is the union of the interiors of all such rays (i.e., the open rays, with the common end point, C , removed).



3. We know that there must be some differences between the establishment of length-measurement functions and the establishment of angle measurement functions, so we look for their origin in the essential differences between the betweenness/congruence properties for points and segments, and the corresponding properties for rays and angles. The key point of difference lies in the restriction of the betweenness properties of rays to those which are co-halfplanar.

<u>An Angularity Structure for the Set of All Angles</u>. We have no intention of giving a fully detailed treatment of angular measure in synthetic geometry, but we can indicate where the treatment must differ from that for linear measure, and where the additional difficulties lie.

First of all, we define an <u>angular measure function</u> to be a function $\gamma : A \to R^+$ (A denotes the set of all angles) such that

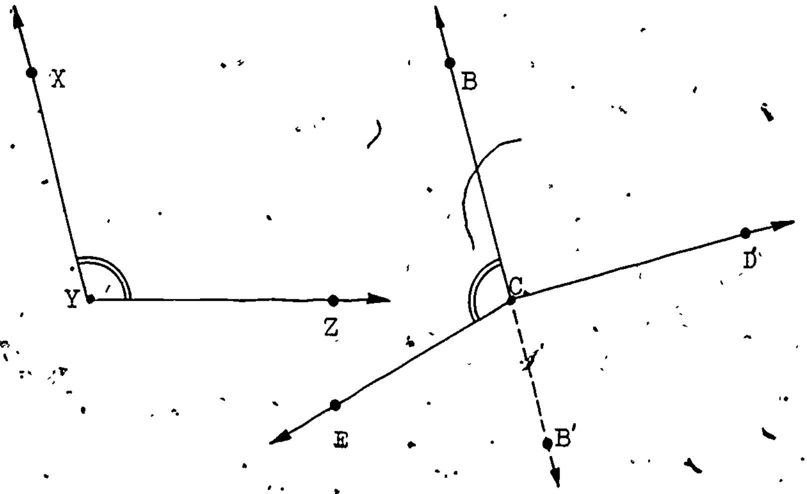
(i) congruent angles have the same value under γ ;

(ii) if $\vec{CX} - \vec{CY} - \vec{CZ}$, then

$$\gamma(\angle XCZ) = \gamma(\angle XCY) + \gamma(\angle YCZ) .$$

Let $\bar{A}$ denote the set of all congruence classes of angles. Exactly as for linear measure, we use the congruence and betweenness postulates to define a relation $<$ on $\bar{A}$, and to show that this is a strict total order relation on $\bar{A}$. In addition, it is dense, and, with respect to the relation $<$, we can show that $\bar{A}$ has no least element and no greatest element.

We next look for a join operation for equivalence classes, and immediately we run into a difficulty. Given any angles $\angle BCD$, $\angle XYZ$, we can follow our length-join procedure, and show that there exists an angle



$\angle BCE$ (with E on the opposite side of $\overleftrightarrow{BC}$ to D) which is congruent to $\angle XYZ$. But (contrary to the situation for linear measure) it is not necessary that $\vec{CD} - \vec{CB} - \vec{CE}$ (see diagram), so we cannot conclude, from the postulates, that the congruence class of $\angle DCE$ is uniquely determined by the given angles. And even if it were, we can see some other problems. If we were to try to define the join of the equivalence classes of the given angles to be the equivalence class of $\angle DCE$, then the angle classes which are determined by $\angle DCB'$ and $\angle ECB'$ would have the same join; and this could hardly be considered satisfactory. You might think that we could resolve the difficulty by making some minor variation in the definition of angle, but this is not so: the difficulty is quite real.

In order to avoid getting bogged down in notation, let us denote congruence classes of angles by letters such as a , b , c , ... , and define two congruence classes a , b , of angles, to be co-halfplanar, if there exist angles, $\angle ECB$, $\angle BCD$ in a , b , respectively, such that E and D are on opposite sides of $\overrightarrow{CB}$, and such that the rays $\vec{CE}$, $\vec{CB}$, $\vec{CD}$ are co-halfplanar as previously defined. This idea can be extended to a finite number of congruence classes, without difficulty. We can then define (as for linear measure) a join operation (*) for those pairs of angle classes which are co-halfplanar. This leads to a binary operation (join) in $\tilde{A}$, which, where defined, is commutative and associative, and preserves order. In other words $(\tilde{A}, *, <)$ has all of the properties of a densely ordered abelian semigroup, except that it is not closed under the join operation.

We now turn to the question of whether there exist functions from $\tilde{A}$ to R^+ which preserve this (limited) structure. It is easy to see that any function which preserves this structure is an angular measure function; and, conversely, that any angular measure function will preserve this structure. The situation is quite similar, in this respect, to that for length functions. If we try to parallel our procedures for the establishment of linear measure functions, we find that we need some sort of archimedean property. It can be shown (but the proof is not simple) that if we assume the archimedean postulate for segments, then there is an "archimedean" property for angles in the sense that if $a \in \tilde{A}$, then there exists a positive integer n $(n > 1)$ such that na (iterated join) is not defined (i.e., the construction for the join leads to non-co-halfplanar rays). We assume such an archimedean property.

We can now proceed with the construction of an angular measure function, γ , for $\tilde{A}$, using our first method as for segments. That is, we first take some angle class $a_0 \in \tilde{A}$, as unit. Then, for any $a \in \tilde{A}$, we compare a with the successive joins $2a_0$, $3a_0$, ... , (if they exist). If there is a least integer n_0' such that $(n_0' + 1)a_0 > a$, let $n_0 = n_0'$. Otherwise, let n_0 be the least integer such that $(n_0 + 1)a_0$ does not exist, and $n_0(a_0) \leq a$. If the equality holds, we define $\gamma(a) = n_0$. Otherwise (using the postulates) we can form the unique "difference class" $a_1 = a - n_0(a_0)$. (Notice that, while postulate CB-1 sets limitations on the possibility of "adding" angles, it indicates that non-congruent angles can always be "subtracted" in one or the other order.)

If we continue to copy our earlier procedure as for segments, we take $n > 1$, and look for a' such that $n(a') = a_0$. But, unless we assume the Cantor-Dedekind postulate, we cannot prove that such submultiples will always exist (and, in fact, they need not. For $n = 3$, that is related to the existence of angle trisectors. In [14] , Chapter 19, it is shown that, in surd cartesian geometry -- which satisfies the axioms of classical synthetic geometry -- such trisectors do not always exist). Fortunately, we can show from the classical postulates that angle bisectors always exist, so we can limit ourselves to binary expansions, and proceed as for length (cf. Section 2-5 for details) to establish a function

$$\gamma : \mathcal{A} \to R^{+}$$

which is based on a_0 as unit. We should then prove that γ preserves the structure of $\mathcal{A}$, and that, if we assume the Cantor-Dedekind postulate, γ maps $\mathcal{A}$ onto an open interval $(0,p)$ of positive real numbers, where $p = \gamma(a') + \gamma(a'')$ for every linear pair (a', a''). This is quite a formidable task and we have no intention of attempting it! (A linear pair of congruence classes of angles (a',a'') has the property that if $\angle A'CB \in a'$, then $\angle A''CB \in a''$, where $\vec{CA''}$ is the opposite ray to $\vec{CA'}$; that this definition gives a satisfactory notion of "supplementary" congruence classes, is shown in Chapter 8 of [14].)

If we look at the alternative procedure, which we carried out in considerable detail for linear measure on segments, the position looks quite difficult: for $a_1, a_2 \in \mathcal{A}$ we need to compare arbitrary positive integral multiples (iterated joins) $m(a_1)$ and $n(a_2)$, and these will not generally exist. We might attempt to get around this difficulty, by re-examining our discussion of the corresponding situation in the establishment of empirical angular measure functions, where we considered generalized joins of wedges, and attempt to extend $\mathcal{A}$ to a larger set $\underline{\mathcal{A}}$ on which an extended join operation, and an extended order relation, would be defined, so as to make $(\underline{\mathcal{A}},*,<)$ an ordered abelian semigroup, in which $(\mathcal{A},*,<)$ is isomorphically imbedded. If we could do this, then we could carry through a ratio procedure as for linear measurement, and set up angular measure functions (on $\underline{\mathcal{A}}$) whose range would be R^{+} if the Cantor-Dedekind postulate is assumed. These functions, restricted to $\mathcal{A}$, would give the desired angular measure functions.

Such a program can, in fact, be carried out, using the "formal sum" procedure rather than some mathematical equivalent of the "generalized join" process. (The difficulty with the latter, is that it takes a lot of mathematical development to cope with the empirically simple idea of counting the number of rotations in a "spiral" of joined wedges.)

We shall not give a detailed description of the procedure for setting up a "formal sum" structure on the set of all angles, but the discussion (in the preceding section) of the procedure for handling formal sums of wedges, contains all of the main ideas. Formal sums of angles are simply finite aggregates of angles. A notion of (finite) decomposition for angles, is defined using the restricted join operation, and this is extended to decompositions of formal sums. Equivalence of formal sums is defined as piecewise congruence under decomposition, and an inequality relation (on equivalence classes) is defined in the natural way. (Roughly speaking, for equivalence classes a , b , of formal sums, a < b if b contains a formal sum such that some proper subset is in a .) Trichotonomy can be shown by using the comparison/subtraction procedure as for formal sums of wedges. A join operation for equivalence classes of formal sums of angles, is defined as for wedges. And so on.

There is a mass of detail to cope with, but in a certain sense, the ideas involved are elementary. With enough patience, and careful attention to detail, we feel that it should be possible to carry through the treatment, and obtain a densely ordered, archimedean semigroup $(\underline{\tilde{A}},*,<)$, in which the restricted system $(\tilde{A}, *, <)$ is isomorphically imbedded. (I.e., if angles are considered as one-element formal sums, equivalence corresponds to congruence, and the join operations and the order relations correspond.) Among the properties which we would need to prove for this imbedding, is that a formal sum is equivalent to an angle if, and only if, it is less than the particular equivalence class of sums which contains all linear pairs.

Such a program can, in fact, be carried through. For the extended structure $(\underline{\tilde{A}}, *, <)$, we can establish a theory of ratios, and hence angular measure functions

$$\gamma : (\underline{\tilde{A}}, *, <) \to (R^+, +, <)$$

whose range includes both arbitrarily large and arbitrarily small numbers. If the Cantor-Dedekind postulate is assumed, it can be shown that γ is onto (this point is discussed more fully below), and that the restriction of γ to the set of congruence classes of angles yields a suitable angular measure function

$$\gamma_p : \tilde{A} \to R^+$$

(and a corresponding function, γ_p , for A) whose range is an open interval (0,p) of positive real numbers, and which has the property that for any linear pair (x_1 , x_2) of angles, $\gamma_p(x_1) + \gamma_p(x_2) = p$. As we would expect, different angular measure functions are related by positive similarities of their ranges. (Cf. Exercise 2-2.5, whose result can be generalized in the manner of Exercise 2-2.19, and the subsequent remark.) Questions of units, and scale change, are just as before, so we do not repeat them. [The fact that (in augmented synthetic geometry) each suitable angular measure function (for simple angles) is onto an open interval, is proved below in the discussion of radian measure.]

Radian Measure. Another procedure for establishing the existence of suitable angular measure functions in synthetic geometry, leads to the familiar radian measure. Because of its importance we describe this briefly, outlining the main steps necessary to show the existence of such a function, and how we would go about proving its main properties. Most details are omitted, but we give enough to indicate that the usual elementary treatment makes some fairly substantial assumptions.

You will recall that radian measure involves the question of the length of a circular arc. Our earlier treatment of curve length was so general that we cannot apply it directly to circular arcs. (It is not obvious that a circular arc is a curve as previously defined!) Hence, in our treatment of radian measure, we must first say something about arc length. Let $\angle PZQ$ be any angle, and let C be any circle with center Z, and in the plane of $\angle PZQ$. Let λ be any length function for space, and let α be the related distance function. Let $C \cap \overrightarrow{ZP} = X$ and $C \cap \overrightarrow{ZQ} = Y$ (see diagram below) and let $\widehat{XY}$ denote the angular arc consisting of X, Y, and those points of the circle which are also in the interior of the angle $\angle PZQ$. If A denotes the set of all angles, we may define a function

$$\gamma : A \to R^{+}$$

by

$$\gamma(\angle PZQ) = \frac{\lambda(\widehat{XY})}{\lambda(\overline{ZX})}$$

where $\lambda(\widehat{XY})$ denotes the arc length of $\widehat{XY}$ as defined below.

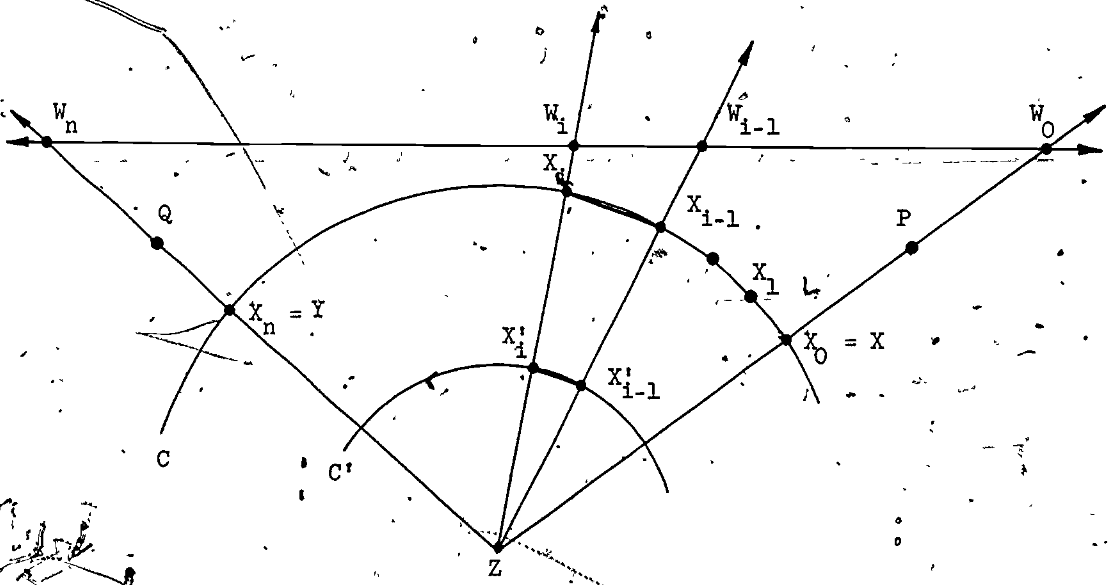
In order that γ be a suitable angular measure function, we must prove that it gives the same value to congruent angles, and that it carries joins into sums. We can also show, with very little extra effort, that γ does not depend on the choice of λ, or on the choice of the circle C; in other words, we obtain only one angular measure function by this process, no matter which circle we use, or which linear measure function we use.

Arc Length. We now give an elementary treatment for arc length. The main advantage of the function definition for a simple curve (or any curve) was that it gave us a natural way to order the "points" of the curve. For an angular arc of a circle, we can order the points by using the betweenness notion for co-halfplanar rays. As mentioned earlier, betweenness for co-halfplanar rays can be defined in terms of betweenness of points of a line, and it can be shown that any finite set of co-halfplanar rays can be ordered by the notion of betweenness. Each ray from the center of a circle (and in the plane of the circle) contains exactly one point of the circle, hence any finite set of points on any open (i.e., without endpoints) semi-circular arc can be ordered. It follows that a finite set of points on any angular arc can be ordered.

Let λ be a length function for space S. We take any $n + 1$ points $X_0 = X$, X_1, X_2, ..., $X_n = Y$ in order on $\widehat{XY}$, (see diagram) and define

$$\lambda(\widehat{XY}) = \sup\left(\sum_{i=1}^{n} \lambda(\overline{X_{i-1}X_i})\right)$$

where the least upper bound is taken over all "inscribed" elementary broken segments, $\overline{X_0X_1\ldots X_n}$.



In order to show that this least upper bound exists, we must show that the set of the "lengths" of all broken segments $\overline{X_0X_1\ldots X_n}$, is bounded. The idea of such a proof is contained in the diagram: there are halfplanes which contain (except for Z) $\overrightarrow{ZP}$ and $\overrightarrow{ZQ}$, and a line may be taken in such a halfplane, parallel to its determining line, and far enough away so that it does not meet the circle C. ~~Such a line must meet~~ $\overrightarrow{ZP}$ and $\overrightarrow{ZQ}$. Let $\overleftrightarrow{W_0W_n}$ be such a line, meeting $\overrightarrow{ZP}$, $\overrightarrow{ZQ}$ in W_0, W_n, respectively, and meeting $\overrightarrow{ZX_i}$ in W_i. Then a straightforward geometric proof shows that, for each segment $\overline{X_{i-1}X_i}$, $\lambda(\overline{X_{i-1}X_i}) < \lambda(\overline{W_{i-1}W_i})$, so that the length of the broken segment $\overline{X_0X_1\ldots X_n}$ is bounded by the length of $\overline{W_0W_n}$. Hence the set of the lengths of all such inscribed broken segments is bounded, and the least upper bound exists. Thus the function $\gamma : A \to R^+$ is properly defined.

The fact that γ is independent of the choice of λ is trivial: any other length function λ_1 is related to λ (on segments) by $\lambda_1 = k\lambda$. It is a simple property of the least upper bound that if G is a bounded set of real numbers, $k > 0$, and $kG = \{kx : x \in G\}$, then $\sup \{kG\} = k \sup G$. Hence $\lambda_1 = k\lambda$ on the set of "angular arcs", and therefore

$$\frac{\lambda_1(\widehat{XY})}{\lambda_1(\overline{ZX})} = \frac{k\lambda(\widehat{XY})}{k\lambda(\overline{ZX})} = \frac{\lambda(\widehat{XY})}{\lambda(\overline{ZX})} .$$

The fact that γ does not depend on the particular circle C, is an exercise in the use of similar triangles, proportionality, and the properties of the least upper bound. [See figure above:

$$\frac{\lambda(\overline{X_{i-1}X_i})}{\lambda(\overline{ZX_i})} = \frac{\lambda(\overline{X'_{i-1}X'_i})}{\lambda(\overline{ZX'_i})} .]$$

We can now show that γ gives the same value on congruent angles (a straightforward exercise in congruence, using the invariance of linear measures under congruence) and hence γ gives a function (which we still call γ)

$$\gamma : \tilde{A} \to R^+ .$$

In order to complete the proof that γ is an angular measure function (and hence that γ preserves the structure of $\tilde{A}$) it is sufficient to show that if a_1, a_2 are two angles, such that $a_1 * a_2$ is an angle, then

$$\gamma(a_1 * a_2) = \gamma(a_1) + \gamma(a_2) .$$

This will follow from the related result for arc length. Clearly any broken segment inscribed in the angular arc $\widehat{a_1}$ which is determined on a given circle by the angle a_1, together with any broken segment inscribed in the arc $\widehat{a_2}$, will yield a broken segment for the arc $\widehat{a_1 * a_2}$. Hence, from properties of the least upper bound, $\gamma(a_1) + \gamma(a_2) \leq \gamma(a_1 * a_2)$. The proof is completed by showing the reverse inequality. We can show

(i) that any broken segment for $\widehat{a_1 * a_2}$ has a "refinement", which gives broken segments for $\widehat{a_1}$ and $\widehat{a_2}$ (i.e., which includes as an endpoint the point where the common ray of a_1, a_2 meets the circle: see dotted segments on diagram); and

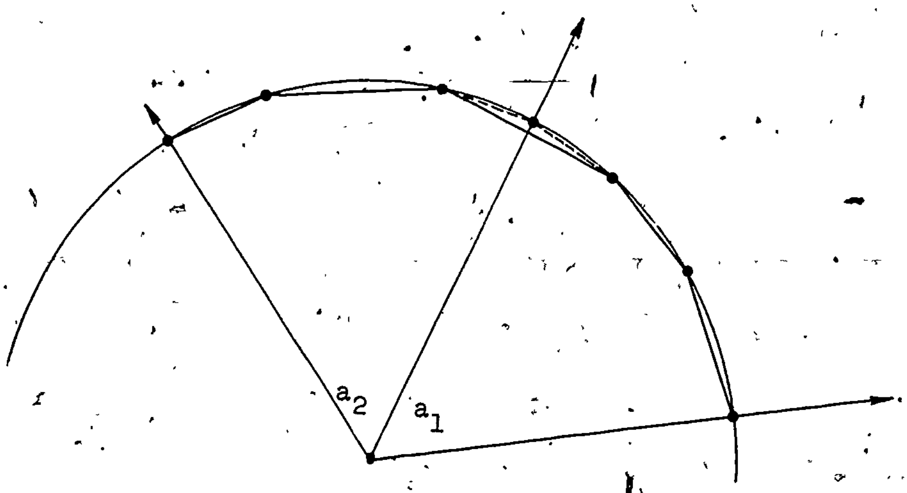
(ii) the length of any refinement of a broken segment is greater than the length of the broken segment. (This is a direct consequence of the "triangle inequality", that the sum of the lengths of two sides of a triangle exceeds the length of the third side.)

Hence, from properties of the least upper bound, we get

$$\gamma(a_1 * a_2) \leq \gamma(a_1) + \gamma(a_2),$$

and therefore

$$\gamma(\widehat{a_1 * a_2}) = \gamma(a_1) + \gamma(a_2).$$



Thus γ carries joins into sums, and hence it is an angular measure function. This important function is called the radian measure function for angular measurement. It is easy to prove that γ is 1-1 on $\mathcal{A}$, and that it preserves order.

We are interested in identifying the range of γ. First of all we should show that the length of a semi-circular arc exists: the proof is a minor modification of our argument concerning angular arcs (our earlier argument concerning the existence of an upper bound has to be modified). This is not difficult, and, as a result, we can show that if π denotes, as usual, the ratio of the length of a semi-circular arc to the length of the radius (this ratio is easily proved (as above) to be independent of both the circle and the length function used) then it follows that the range of the angular measure function γ is contained in the open interval $(0,\pi)$.

In order to obtain an angular measure function as postulated in the metric treatment of geometry, we still need to show two things:

(i) If (a,a') is a linear pair, then $\gamma(a) + \gamma(a') = \pi$.

(ii) γ is onto $(0,\pi)$.

The proof of (i) is a trivial extension of the argument above which showed that $\gamma(a_1 * a_2) = \gamma(a_1) + \gamma(a_2)$; (in effect, we need to extend the "additivity" theorem for arc length).

The proof of (ii) is much more interesting. You might think that we could prove it by noting that it is equivalent to showing the existence (on a circle of radius 1) of angular arcs of any length r, where $0 < r < \pi$. But we do not know that such arcs exist, and if we try to prove that they exist, we will find that the proof is no easier than it is to prove that γ is onto $(0,\pi)$. The truth of the matter is that neither property is necessarily true unless we make some assumptions beyond those of classical synthetic geometry. The situation is quite similar to that which concerned us when we were discussing the "ontoness" of length functions on segments, and the same assumption (the Cantor-Dedekind completeness postulate) is sufficient to enable us to complete the proof that γ is onto $(0,\pi)$. We outline this proof in some detail, as it is hard to find in the literature, and because, in addition to the existence question for arcs of specified length and angles of specified radian measure in the interval $(0,\pi)$, it enables us to show that, if the Cantor-Dedekind postulate is assumed, then "trisectors" exist for any angle. (For a proof that trisectors do not necessarily exist in a geometry which satisfies only the classical postulates of synthetic geometry, see Chapter 19 of [14] .) Note carefully that our proof of the existence of trisectors does not imply constructability in the classical sense: the proof of non-constructability (which you can also find in [14]) is still valid in the augmented geometry which includes the Cantor-Dedekind postulate.

Let r be any real number between 0 and π. We want to show that there exists an angle whose radian measure is r. You will recall that, in considering the corresponding "ontoness" question for length functions, we used the fact that we could show the existence of (actually construct) all segments whose lengths were rational multiples of the length of a given segment. We can't expect to do this for angles -- after all $\frac{1}{3}$ is a rational number, and we know that we can't even show (constructively) the existence of trisectors. In fact all that we can prove in this direction, is that every angle has a unique bisector. (We leave to you this well-known and straightforward proof, involving the existence of midpoints for segments, and congruence of triangles.) Fortunately this is sufficient: if each angle has a bisector, we can show the existence of 2^n-sectors for each positive integer n. (This merely involves repeated bisection.) Hence, using our join operation, given any angle a, we can show the existence of angles with radian measures $\frac{m}{2^n} \cdot \gamma(a)$, for all of those positive integers m and n for which this number is less than π. Since a may be a right angle, whose radian measure is easily proved to be $\frac{\pi}{2}$, we can thus find angles whose radian measures are all numbers $\frac{m\pi}{2^n}$, $(n \geq 1, m < 2^n)$. [We point out the interesting fact that the numbers $\frac{m}{2^n}$, with $m < 2^n$, are just those rational numbers between 0 and 1 which have finite binary expansions. This set of numbers is dense in the sense of order (i.e., between any two there is another) but of course this set does not include all rational numbers. It is also a dense subset of $(0,1)$ in the topological sense, and this is used indirectly in the following argument.]

We return to the problem of showing that there exists an angle whose radian measure is r, for r between 0 and π. Let

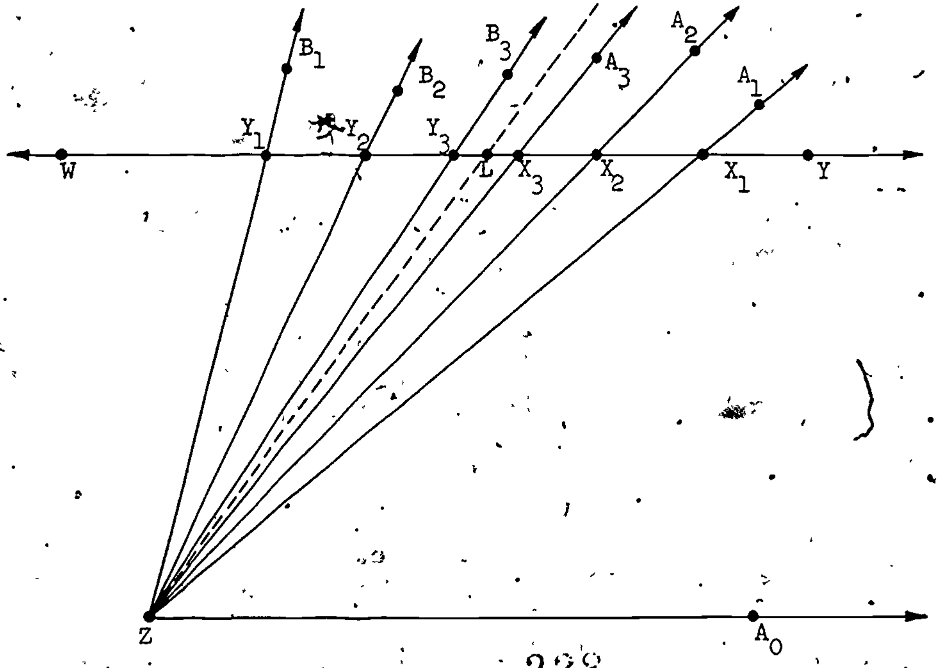
$$K = \{\frac{m}{2^n}\pi : m < 2^n\}$$

$$K_1 = \{x : x \in K , x < r\}$$

$$K_2 = \{x : x \in K , x \geq r\} .$$

It is not hard to show (from real number properties) that neither K_1 nor K_2 is empty, and that there are sequences of numbers (a_i), (b_i) from K_1, K_2 respectively, such that $a_i < a_{i+1} < b_{i+1} \leq b_i$, for each i, and such that $|a_i - b_i| < \frac{1}{2^i}$. (In other words, a_i is strictly monotone increasing, and b_i is strictly monotone decreasing; a_i is always less than b_i; and the difference $b_i - a_i$ is arbitrarily small, for sufficiently large i. The proof that such sequences exist involves only properties of real numbers, and the proof is quite similar to the proofs of similar theorems about cuts which we have used earlier.)

Now let $i = 1, 2, 3, \ldots$, and let $\overrightarrow{ZA_i}$, $\overrightarrow{ZB_i}$ be rays which (except for Z) lie in the same halfplane, and such that $\gamma(\angle A_iZA_0) = a_i$, $\gamma(\angle B_iZA_0) = b_i$. (See diagram.)



Let $\overleftrightarrow{WY}$ be any line parallel to $\overrightarrow{ZA_0}$, and in the same halfplane as each B_i . Then each ray $\overrightarrow{ZA_i}$, and each ray $\overrightarrow{ZB_i}$ must meet $\overleftrightarrow{WY}$. Let X_i , Y_i be the respective points of intersection. It now follows (from the definition of γ , and from the betweenness properties of rays and points, and the relationship between them) that, for each i ,

$$\overrightarrow{ZA_1} - \overrightarrow{ZA_2} - \dots - \overrightarrow{ZA_i} - \overrightarrow{ZB_i} - \overrightarrow{ZB_{i-1}} - \dots - \overrightarrow{ZB_1}$$

and

$$X_1 - X_2 - \dots - X_i - Y_i - Y_{i-1} - \dots - Y_1$$

Hence we have a sequence of segments, $\sigma_i = (\overline{X_iY_i})$, such that $\sigma_{i+1} \subset \sigma_i$ for each i . It follows from the Cantor-Dedekind postulate, that $\bigcap_i \sigma_i \neq \emptyset$. If this intersection were to contain more than one point, we would obtain two different angles, each greater than or equal to every $\angle A_iZA_0$, and each less than or equal to every angle $\angle B_iZA_0$. This would imply that the (different) radian measures of these angles were similarly related, and hence that the intersection of the real number segments $[a_i,b_i]$ contained at least two different numbers. But it is a straightforward real number argument to prove that $\bigcap_i [a_i,b_i]$ is the single real number r . Hence there is a single point L in $\bigcap_i \sigma_i$. The ray $\overrightarrow{ZL}$ determines an angle $\angle A_0ZL$, and, because the radian measure function is order preserving, we have for each i ,

$$a_1 \leq \gamma(\angle AZL) \leq b_i \; .$$

Hence $\gamma(\angle AZL) \in \bigcap_i [a_i,b_i] = \{r\}$, and hence γ is onto $(0,\pi)$, as we set out to show.

Remarks:

1. Henceforth we denote the radian measure function by γ_π . Observe that the definition of γ_π did not depend on the Cantor-Dedekind postulate, so the concept of radian measure is defined in any synthetic geometry; but we needed the Cantor-Dedekind postulate in order to prove that γ_π was onto $(0,\pi)$.

It is easy to see that, for any $k > 0$, the function $k\gamma_\pi$ is also an angular measure function. Moreover, it can be shown that if γ is any angular measure function, then there exists $k > 0$ such that $\gamma = k\gamma_\pi$. If the Cantor-Dedekind completeness property is assumed, then the proof of the latter fact is relatively straightforward (corresponding to Exercise 2-2.5); but otherwise the proof uses the monotone property of angular measure functions and also uses (in essence) the (topological) denseness of the set of "binary" rational numbers $\{\frac{m}{2^n} : m, n, \text{ positive integers}, m \leq 2^n\}$ in the interval $[0,1]$. This is very similar to the corresponding question for length functions (see the remark which follows Exercise 2-2.19). We do not wish to go into detail here.

It follows that, in any geometry which satisfies the postulates of classical synthetic geometry, any two angular measure functions are similar, and the set of all angular measure functions is a ratio scale. If the Cantor-Dedekind postulate is assumed, each angular measure function $\gamma = k\gamma_\pi$ is onto an open interval $(0,k\pi)$ of real numbers. For the degree measure function (which is used in the SMSG treatment of metric geometry) $k = \frac{180}{\pi}$.

2. The proof that γ_π is onto $(0,\pi)$, implies that, for any angle, a, and any real number r such that $r[\gamma_\pi(a)] < \pi$, there exist angles whose radian measures are $r[\gamma_\pi(a)]$. In particular, this is true for $r = \frac{1}{q}$, where q is a positive integer. Hence (with the assumption of the Cantor-Dedekind postulate) we can show the existence of q-sectors for every angle and every positive integer q. Thus trisectors certainly exist in augmented classical geometry, even though they cannot be constructed by "classical" methods.

3. It is another by-product of the fact that γ_π is onto $(0,\pi)$, that a circle of radius c has angular arcs of lengths rc, for every real $r \in (0,\pi)$, and that any angular arc of length s has "sub-arcs" of length s', where s' is any real number between 0 and s.

4. We are accustomed to using all of the above properties. This means that we are, in effect, assuming (usually implicitly) a geometry which satisfies the Cantor-Dedekind completeness property. In the metric treatment of geometry, and in the cartesian treatment with real coordinates, this property follows from the postulated properties of the distance and the coordinate functions.

The Relationship of Linear and Angular Measures. One of the most important facts about measure functions (both empirical and mathematical) is that there are relationships between them. Cartesian products and commutative diagrams are a natural vehicle for illustrating these relationships.

From what we have proved above, it follows that the radian measure function, whose definition involved the use of a length function, coincides with one of the angular measure functions which were defined (and whose existence could be established) quite independently of length. Whenever this sort of situation occurs, (equality, or relationship, of two functions which are defined, or arrived at, quite differently) you should suspect that there might be an underlying commutative diagram, representing the fact that, starting from any domain element (in this case an angle) you could reach the same point (it's angular measure) by different "paths". We illustrate this "commutativity" situation in the diagram below. D is a domain for length functions. We assume that D includes at least all segments and all angular arcs. λ is any length function on D ; $(\lambda : D \to R^+)$. $\lambda \times \lambda$ is the natural function determined by λ on the cartesian product $D \times D$. $(\lambda \times \lambda : D \times D \to R^+ \times R^+)$. A denotes the set of all angles. The function $f : A \to D \times D$, may be defined as follows: Select any fixed plane and any fixed circle C (with center Z) in that plane. If $a \in A$, select any angle $\angle XZY$ congruent to a , where X , Y are points of C (see diagram used in definition of radian measure). Then f is defined by $f(a) = (\widehat{XY}, \overline{XY})$. The mapping $\delta : R^+ \times R^+ \to R^+$ is the division function, $\delta(x,y) = \frac{x}{y}$. Then the fact that the radian measure function is equal to one of the independently defined angular measure functions, is equivalent to the statement that there is one of these functions, (γ_0 say) which (for every choice of λ , C , and $\angle XZY$) satisfies : $\gamma_0 = \gamma_\pi$, where γ_π is the composite function $\gamma_\pi = \delta(\lambda \times \lambda)f$. That is, there is a γ_0 which makes the following diagram commutative:

$$\begin{tikzcd}
D \times D \arrow[r, "\lambda \times \lambda"] & R^+ \times R^+ \arrow[d, "\delta"] \\
A \arrow[u, "f"] \arrow[r, "\gamma_0"] & R^+
\end{tikzcd}$$

It is useful to show what happens to a particular element of A , using a "parallel" diagram:

$$\begin{tikzcd}
(\widehat{XY}, \overline{XY}) \arrow[r, "\lambda \times \lambda"] & (\lambda(\widehat{XY}) , \lambda(\overline{XY})) \arrow[d, "\delta"] \\
a \arrow[u, "f"] \arrow[r, "\gamma_0"] & \gamma_0(a) = \lambda(\widehat{XY}) / \lambda(\overline{XY})
\end{tikzcd}$$

Because $\gamma_0 = \gamma_\pi$, there is no need to distinguish between them because of the way in which they were defined: they are the same function.

Arc Lengths as Angle Measures. Perhaps you are accustomed to a definition of the radian measure of an angle, as the length of an arc of a "unit" circle. There is nothing wrong with such a procedure: it leads to exactly the same function. We have used the alternate procedure (defining radian measure in terms of the quotient of two related length measures) because it makes more apparent the important fact that, although we use length measures in defining radian measure for angles, the radian measure function is completely independent of any particular length function. Of course this is still true if we use the arc length definition; in that case a change of length function from λ_1 to $\lambda_2 = k\lambda_1$, changes a unit circle by the similarity "point transformation" $[\frac{1}{k}]$. Thus, if an angle a determines an arc $x_1 = (g_1(a))$ on a unit circle when λ_1 is used, and an arc $x_2 = (g_2(a))$ on a unit circle when λ_2 is used, then $x_2 = [\frac{1}{k}] x_1$, and we have (in essence) "commutativity", as represented by the diagram:

$$\begin{tikzcd}
& a \arrow[dl, "g_1"'] \arrow[dr, "g_2"] & \\
x_1 \arrow[rr, leftrightarrow, "{[\frac{1}{k}]}"'] \arrow[dr, "\lambda_1"'] & & x_2 \arrow[dl, "\lambda_2 = k\lambda_1"] \\
& \lambda_1(x_1) = \lambda_2(x_2) &
\end{tikzcd}$$

That is (cf. the earlier discussion of point transformations and similarities) $\lambda_1 g_1(a) = k\lambda_1[\frac{1}{k}]g_1(a) = \lambda_2 g_2(a) = \gamma_\pi(a)$.

A second reason for our choice of treatment for radian measure, was that it exhibited the relationship of linear and angular measures in a way which we can follow closely when exhibiting the relationship of length and area, and the relationship of length and volume.

It is also possible to define the radian measure function in terms of area. As is well known (we discuss it later) area functions are related to length functions, so this procedure relates length functions and angular measure functions indirectly.

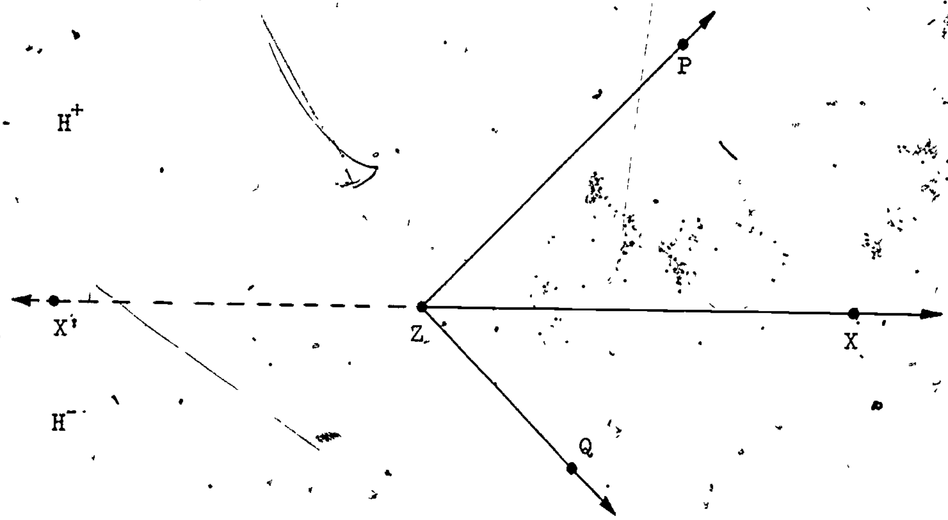
Angular Distance and Angular Coordinates. It is now a simple matter to introduce the concepts of angular distance and angular coordinates, and the related notion of polar coordinates. Let $\underline{Z}$ denote the set of all rays (in space) with a common endpoint Z . If z_1 , z_2 belong to $\underline{Z}$, and if γ is an angular measure function with range $(0,p)$, we can define an angular distance function

$$\alpha_\gamma : \underline{Z} \times \underline{Z} \to [0,p]$$

by:

$$\alpha_\gamma(z_1,z_2) = \begin{cases} 0 \text{ if } z_1 = z_2 . \\ p \text{ if } z_1 \text{ is opposite } z_2 . \\ \gamma(z_1 \cup z_2) \text{ otherwise. (Remember that} \\ \qquad z_1 \cup z_2 \text{ is an angle.)} \end{cases}$$

If we now restrict attention to the subset $\underline{Z}_H$ of $\underline{Z}$, consisting of those rays of $\underline{Z}$ which lie in a given plane H, then, we can set up an angular coordinate system for rays. Take any ray $\overrightarrow{ZX}$ in the plane as "initial ray" or "origin ray". Then the line $\overleftrightarrow{ZX}$ determines two half-planes, which we designate arbitrarily as H^+ and H^-. (See diagram.)



If $P \in H^+$, we define the angular coordinate, $f_\gamma(\overrightarrow{ZP})$, of the ray $\overrightarrow{ZP}$ (with respect to α_γ) to be the angular distance $\alpha_\gamma(\overrightarrow{ZX},\overrightarrow{ZP})$. If $Q \in H^-$, we define the angular coordinate $f_\gamma(\overrightarrow{ZQ})$ of $\overrightarrow{ZQ}$ to be $-\alpha_\gamma(\overrightarrow{ZX},\overrightarrow{ZQ})$. The ray $\overrightarrow{ZX'}$ opposite to $\overrightarrow{ZX}$ is given the angular coordinate p. (Clearly -p would be equally satisfactory: we choose one of the possibilities so as to make the resulting coordinate function

$$f_\gamma : \underline{Z}_H \to (-p,p]$$

1-1 and onto. Eventually, of course, it is convenient to consider the range of f_γ as the set of equivalence classes of real numbers mod 2p, and -p, p, then belong to the same equivalence class.)

We have drawn the diagram in the conventional way, but of course notions like left and right, and up and down, have no objective mathematical meaning. Any ray could be taken as initial ray, and either halfplane (with respect to the line containing the initial ray) can be designated as the positive halfplane.

An angular distance function, and a corresponding angular coordinate function for concurrent rays in a plane, have a relationship somewhat like that for a distance function, and a corresponding coordinate function for a line. For any two rays z_1, $z_2 \in \underline{Z}_H$, we have

$$\alpha_\gamma(z_1,z_2) = \begin{cases} |f_\gamma(z_1) - f_\gamma(z_2)| \text{ ; if } |f_\gamma(z_1) - f_\gamma(z_2)| \leq p \\ \\ 2p - |f_\gamma(z_1) - f_\gamma(z_2)| \text{ , if } |f_\gamma(z_1) - f_\gamma(z_2)| \geq p \text{ .} \end{cases}$$

You should draw appropriate diagrams and verify this for yourself.

Polar Coordinates. A polar coordinate system (for a plane) is a combination of an angular coordinate system with either a system of concentric circles, or with a coordinate system for a line.

In order to determine a polar coordinate system for a plane, we first need to have an angular coordinate system. This means that we have an initial ray, a designated positive halfplane, and an angular measure function or an angular distance function. The initial ray $\overrightarrow{ZX}$, is called the polar axis, and its endpoint, Z, is called the pole. We also need to be given a length function or a distance function.

The simplest way of looking at a polar coordinate system, combines the angular coordinate system with the family of concentric circles which are centered at the pole. (Think of the way latitude circles and longitude "rays" look at the poles, and you will see the origin of the terms "pole", and "polar coordinates".) A circle in this family is fully determined by its radius (i.e., the positive number which measures the length of its radius under the given length function). The polar coordinates of a point P (not the pole) are a pair of numbers (r,θ). The first number, r, is the radius of the unique circle (with center Z) on which P lies, and the second number, θ, is the angular coordinate (in the given angular coordinate system) of the unique ray on which P lies. If $(0,p)$ is the range of the given angular measure function, and $\tilde{\theta}$ is an equivalence class mod $2p$, then there is a 1-1 correspondence of points P (not the pole) and pairs $(r,\tilde{\theta})$, $r > 0$. The pole is given the coordinates $(0,\tilde{\theta})$, where $\tilde{\theta}$ is any equivalence class.

There is a second way of looking at a polar coordinate system. This is consistent with, but extends the earlier procedure, in permitting the first coordinate, r , to take negative values. This second approach combines a given angular coordinate system (as before) with a coordinate system for a line. Observe that if any ray $\overrightarrow{AB}$ is given, and if we have a fixed distance function α , then there is a unique coordinate system f , for the line $\overleftrightarrow{AB}$, which is compatible with α , and which has $f(A) = 0$, $f(B) > 0$. We call this the coordinate system for $\overrightarrow{AB}$ which is determined by $\overrightarrow{AB}$ and α . This idea is used in the second type of polar coordinate system. Again, polar coordinates for a point P are a pair of numbers (r,θ) . The second number is the angular coordinate of either the ray $\overrightarrow{ZP}$, or the opposite ray $\overrightarrow{ZP'}$. The first number, r , is then the line coordinate of P in the coordinate system for $\overleftrightarrow{PP'}$ determined by $\overrightarrow{ZP}$, or $\overrightarrow{ZP'}$, respectively. That is, the second number , θ , determines a unique ray, and hence it determines a line and a coordinate system for that line; the first number, r , is then the coordinate of a point on that line. Thus the pair (r,θ) determines a unique point. In this type of polar coordinate system r may be any real number. Moreover, if $(0,p)$ is the range of the given simple angular measure function, then the coordinates (r,θ) , $(-r,\theta + p)$ determine the same point, so that each point (except the pole) is represented by two essentially different coordinate pairs. It is easily verified that if $r \geq 0$, and if fixed underlying distance and angular distance functions are assumed, then the pair (r,θ) determines the same point in either of the two types of polar coordinate systems which we have described.

3-4. Extension of the Domain for Angular Measure: Directed Angles and Rotations

In this section we consider some extensions of the notions of angle and angular measure. The first of these (involving the notion of some sort of "generalized angle", whose measure could be any positive real number) was implicit in the work of the earlier sections; i.e., in the extended joins for wedges, and in the formal sums of angles. In the (empirical) extended join situation, we could have defined a generalized angle to be an equivalence class of extended joins. But the corresponding situation for formal sums (in the purely mathematical context) is not very satisfactory, because we have lost completely the idea of "joining the angles together in a continuing order". We can approach this latter idea in two ways: either by using directed angles, or by using rotations. We shall explain each of these procedures.

Directed Angles. The notion of directed angle is closely related to the notion of directed segment, and to vector ideas. A simple directed angle is defined to be an ordered pair of non-collinear rays with a common endpoint. Clearly, each angle determines two simple directed angles, and each simple directed angle determines a unique angle. If we restrict our attention to the set of all simple directed angles in a fixed plane, then it is possible to define an equivalence relation of "same orientation" on this set, and there are exactly two equivalence classes, which we can think of as corresponding to "clockwise" and "counterclockwise", or "positive" and "negative". As you might expect, it is a little more complicated to formally define such an equivalence relation for coplanar simple directed angles, than it was for collinear directed segments, but it can be done by filling in the details of the following scheme.

First of all, for two coplanar simple directed angles (a_1,a_2), (b_1,b_2), we define a relation, $\uparrow$ (read as "has the same orientation as") by: $(a_1,a_2) \uparrow (b_1,b_2)$ if (but not of course, only if) the pairs of rays (a_1,b_1) and (a_2,b_2) are parallel and similarly "directed". (This can easily be made precise.) Hence (since we will want $\uparrow$ to be transitive) we can reduce consideration to those coplanar simple directed angles which have a common vertex. Let (a_1,a_2), (b_1,b_2) be such simple directed angles. We need to find some device for conveying the idea of "rigidly rotating" (b_1,b_2) until its "initial ray" b_1, coincides with a_1, and then comparing the relative locations of the new "terminal ray" b_3, and the terminal ray a_2. We then define $(a_1,a_2) \uparrow (b_1,b_2)$ if a_2 and b_3 are on the same side of a_1. (By this we mean, of course, the same side of the unique line which a_1 determines.) In order to handle the "rigid rotation" idea, using only congruence and betweenness properties, the following scheme may be used: (You should draw diagrams and, if possible, prove that the scheme really works.)

(a) If $a_1 = b_1$, take $b_3 = b_2$.

(b) If $a_1 = b_1'$ (we use primes to denote opposite rays), take $b_3 = b_2'$.

(c) Otherwise take b_3 so that $\angle(a_1 \cup b_1) \cong \angle(b_2 \cup b_3)$, and

(i) if $a_1 = b_2$, take b_3 on the b_1' side of b_2;

(ii) if $a_1 = b_2'$, take b_3 on the b_1 side of b_2;

(iii) if $b_1 - a_1 - b_2$, or $b_1 - b_2 - a_1$ take b_3 on the b_1' side of b_2;

(iv) if $b_1 - a_1 - b_2'$ or $b_1 - b_2' - a_1$, take b_3 on the b_1 side of b_2.

It can be shown that the relation ↑ so defined is an equivalence relation, and that if (a_1,a_2), (b_1,b_2) are two directed angles, then either $(a_1,a_2) \uparrow (b_1,b_2)$, or $(a_2,a_1) \uparrow (b_1,b_2)$. It follows that there are exactly two equivalence classes of simple directed angles in a plane. Each class is called an <u>orientation of the plane</u>. Two different simple directed angles which determine the same angle (e.g., (a_1,a_2) and (a_2,a_1)) always belong in "opposite" classes. We "orient" the plane by selecting one of these equivalence classes. Thus each simple directed angle determines an orientation of its plane. An orientation may be given in other ways. For example, by three non-collinear points, named in order; by an ordered pair of different half-planes, whose bounding lines are not parallel; by a ray and a designated halfplane bounded by the line which the ray determines; (hence, by an angular coordinate system). You should draw sketches and see how each of these situations can be used, in a natural way, to determine an orientation. You should also verify that if A, B, C, are the vertices of a triangle, then opposite orientations are determined by each of the two "cyclic orders" in which the vertices can be named; i.e., the simple directed angles $(\overrightarrow{BA}, \overrightarrow{BC})$, $(\overrightarrow{CB}, \overrightarrow{CA})$, and $(\overrightarrow{AC}, \overrightarrow{AB})$ each determine the same orientation.

The notion of directed angle, like that of directed segment, is essentially a vector notion. The set of all simple directed angles in a plane can be classified by combining the relations of congruence and orientation. If δ is a simple directed angle, denote the corresponding (undirected) angle by $\overline{\delta}$. We then define an equivalence relation, $\sim$, on the set of all simple directed angles in a plane, by:

$$\delta_1 \sim \delta_2 \text{ if and only if } \overline{\delta}_1 \cong \overline{\delta}_2 \text{ and } \delta_1 \uparrow \delta_2 .$$

We could now define a join operation for equivalence classes of simple directed angles, using the same approach as for angles; but, as for angles, we ~~would~~ find that the set is not closed with respect to the join operation. We therefore first seek to extend the notion of directed angle so as to create a larger system, which includes the simple directed angles and which is closed under an appropriate join operation. There are two closely related ways of doing this. One is to extend the notion of simple directed angle by defining a directed angle to be an ordered n-tuple $(n \geq 1)$ of similarly oriented directed angles, with the property that the terminal ray of each is the initial ray of the next. (Cf. the definition of broken segment.) The other procedure, which is easily shown to be equivalent to the first, is to define a directed angle to be an ordered n-tuple $(n \geq 2)$ of coplanar rays $(a_1 , a_2, \ldots , a_n)$ which have a common endpoint, and which have the property that all of the ordered pairs (a_i , a_{i+1}) determine similarly oriented simple directed angles. (This common orientation is called the orientation of the directed angle.) With this definition, a simple directed angle is a directed angle with $n = 2$.

Notice that what each of these equivalent definitions does, is to overcome the limitations of our elementary idea of angle by using the notions of simple directed angle and orientation to convey the intuitive idea of "rotation in the same direction". We could have done this without the explicit introduction of orientation, but there would not be any saving in the work involved; and, in any case, orientation is an important idea in its own right.

We still have not quite reached an "angle" notion which corresponds directly to the idea of a (sensed) rotation of a ray: for this notion the important ideas seem to be the "initial" ray, the "direction" of rotation, and the "amount" of rotation (which determines the "final" ray). In other words, as far as the idea of a sensed rotation is concerned, we should somehow classify our directed angles into those which have the same initial and terminal rays, the same orientation, and the same angular measure. In order to do this, we have to define angular measure functions for directed angles. This is quite straightforward.

We first define a generalized angle to be any directed angle, without regard to its orientation.

[Remark. We could have defined generalized angles without first defining directed angles, but we would have had to find some way of making precise the same idea of "fitting simple angles together in the same direction". In the long run it is more economical to define the notion of directed angle first.]

We consider next the relationship of our generalized angles to the finite "formal sums" of angles which we used in setting up angular measure functions for angles. Clearly every generalized angle determines a "formal sum": the "formal sum" of its "component" angles. Moreover, as you may show, in every equivalence class of "formal sums" there are many "formal sums" which are determined in this way by generalized angles. Therefore we can use the equivalence relation for "formal sums" to obtain an equivalence relation for generalized angles, and our resulting set of equivalence classes of generalized angles will be in 1-1 correspondence with the set of equivalence classes of "formal sums". It is natural to use this situation to give an "additive" structure to the set of equivalence classes of generalized angles, and to obtain angular measure functions for generalized angles. In other words, if $A = (a_1, a_2, \ldots, a_n)$ is any generalized angle (each a_i is a ray), and γ is a measure function for "formal sums" (and hence, of course, for simple angles) we thus obtain

$$\gamma(A) = \sum_{i=1}^{n-1} \gamma(a_i \cup a_{i+1})$$

With the archimedean and Cantor-Dedekind postulates assumed, the range of γ on generalized angles will be the same as the range on "formal sums"; i.e., R^+.

It is not too difficult (using mathematical induction) to show that if two directed angles are coplanar, have the same initial ray, the same orientation, and the same angular measure (when considered as generalized angles), then they have the same terminal ray. The equivalence relation for generalized angles, taken together with orientation, gives an equivalence relation for coplanar directed angles, and it is a straightforward matter to define an associative, commutative join operation for these equivalence classes.

At this point we are very close to the idea of rotation. We can define a geometric plane rotation to be an equivalence class of coplanar, similarly oriented, directed angles, with a common vertex, and with the same angular measure. If we add the additional restriction that all of the directed angles should have a common initial ray, then the resulting equivalence classes are called geometric ray rotations. Clearly, both geometric plane rotations and geometric ray rotations may be oriented. It is not difficult to show that, for a fixed plane and vertex, each geometric ray rotation is a subset of a geometric plane rotation, and that the correspondence of plane rotations, and ray rotations with a fixed initial ray, which is thus established, is 1-1 and preserves orientation.

From this point it is only a short step to the establishment of a vector space structure on the set Δ_p of geometric plane rotations (and, consequently an isomorphic structure on the set Δ_r of geometric ray rotations with a fixed initial ray). Addition is defined in the natural way: if two elements of Δ_p are similarly oriented, we add them by using the join operation for equivalence classes of directed angles. If not similarly oriented, then we have to introduce a "subtraction" procedure; this does not present any real difficulty. Scalar multiplication by real numbers is introduced in the usual way (in particular, negative multiplication reverses orientation) and we obtain a 1-dimensional vector space over the reals. An isomorphism of this space to the reals can be established by selecting one of the orientations as "positive", and one as "negative", and using any angular measure function γ to define a vector measure function (which is an isomorphism), $\underline{\gamma} : \Delta_p \to R$ as follows. Let $\tilde{\delta} \in \Delta_p$, and let $\delta \in \tilde{\delta}$ define

$$\underline{\gamma}(\tilde{\delta}) = \begin{cases} \gamma(\overline{\delta}) & \text{if } \tilde{\delta} \text{ has positive orientation,} \\ -\gamma(\overline{\delta}) & \text{if } \tilde{\delta} \text{ has negative orientation.} \end{cases}$$

An isomorphic vector space Δ_r is obtained if we restrict ourselves to the equivalence classes of coplanar directed angles at a point and with a specified initial ray. In this case, we can use any linear isomorphism from Δ_r to R (established from a particular angular measure function) to extend the angular coordinate system earlier established. In this extended system, the angular coordinate of a ray, ρ, may be any one of the real numbers which correspond to a directed angle which has the "origin ray" as initial ray, and the ray ρ as terminal ray. It can be shown that these numbers are congruent mod $2p$, where $(0,p)$ is the range of the underlying simple angular measure function.

Rotation. In the above discussion of directed angles we reached the point where we were able to give a geometrical definition of a simple rotation of a ray or of a plane. In view of the intuitive simplicity of the notion of "rotation", you probably found the formal introduction surprisingly awkward. You might find it useful to try to recall how you normally handle questions of rotation, and to try to devise a simpler way of introducing the notion into the formal structure of geometry. We now look at an analytical approach to this question, making use of the notion of continuous function. The treatment has much in common with that of curve length.

The first kind of rotation which we look at is that which, intuitively, corresponds to a simple motion of a "ray" in a plane, with its endpoint held fixed as a "center of rotation". (I.e., something like a single sweep of a simple pendulum.) We formalize this idea as follows: Let Γ_H be the set of all rays in a fixed plane H, and with a common endpoint Z, and let $[a,b]$ be an interval of real numbers. (You may think of $[a,b]$ as a time interval.) Then a simple ray rotation, with center Z, is a 1-1 continuous function from $[a,b]$ to Γ. (The notion of continuity may be defined by using any angular distance function for rays, and the usual distance structure of R: we do not want to go into detail here.) The combination of continuity with the 1-1 property, ensures that the rotation is "simple" in that it reflects the intuitive idea of "continuously turning in the same direction". The 1-1 condition also ensures, of course, that such a simple ray rotation is less than a "full revolution".

Given any measure function γ for angles, we can easily define a corresponding measure function for simple ray rotations; but we do not stop to do this, because it will be just as easy to define angular measure for the more general type of ray rotation which we consider next, and which includes simple ray rotations. We do however remark here that, if the image set (of rays) of a simple ray rotation $f : [a,b] \rightarrow \Gamma_H$ lies in a half-plane, then, corresponding to each measure function γ for angles, there will be an angular measure function ($\bar{\gamma}$, say) for rotations, such that

$$\bar{\gamma}(f) = \gamma(f(a) \cup f(b)).$$

You should observe that if we had originally defined an angle to be a "segment of rays", then for a simple ray rotation whose image lies in a half-plane, the image would be an angle, and the angular measures of the rotation and its image angle would be the same.

As a next step, we might consider how to extend our simple ray rotations so as to obtain more general rotations which correspond to the earlier-defined geometric rotations. To do this we will have to find some way of conveying the "monotone" idea of "continuously turning in the same direction". But before we do this, let us first look at a more general notion of ray rotation, and define suitable angular measure functions. These angular measure functions will then apply automatically to the more restricted ray rotations.

Let Z be any point of space, let Γ be the set of all rays with end-point Z, and let $[a,b]$ be an interval of real numbers. We define a ray rotation (at Z) to be a continuous function

$$f : [a,b] \to \Gamma .$$

You should observe the parallelism between this definition and that for a space curve: it has just the same degree of generality. For the same reasons as for curves, it would not be feasible (in general) to define the rotation to be the range of the function, although in this case you probably don't have quite as strong an urge to do this as you might have had in the case of curves. As for curves, we may identify the function with its graph in $R \times \Gamma$, and (with a suitably-defined notion of continuity) this graph will be the 1-1 continuous image of the naturally related function F, defined on $[a,b]$ by

$$F : x \to (x, f(x)) .$$

The definition of the angular measure of a rotation f is defined exactly as for the length of a curve: it simply makes precise the intuitive idea of "total angular distance traveled".

If γ is an angular measure function with range $(0,p)$ and if α_γ is the related angular distance function for rays, then a corresponding angular measure function $\bar{\gamma}$ for (some) ray rotations is defined as follows:

Assume $a < b$, and let

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

be any finite ordered set of points on $[a,b]$. Then we define

$$\overline{\gamma}(f) = \sup \left(\sum_{i=1}^{n} \alpha_\gamma \left(f(x_{i-1}) , f(x_i)\right)\right)$$

provided that the least upper bound exists. Here we understand that the least upper bound is taken over all finite "partitions" of the interval $[a,b]$, and that $\alpha_\gamma(\rho_1 , \rho_2) = p$ if the ray ρ_2 is opposite to the ray ρ_1.

With this definition it can be proved that:

(i) If γ_1 & γ_2 are angular measures with $\gamma_2 = k\gamma_1$, then $\overline{\gamma}_2 = k\overline{\gamma}_1$.

(ii) If f is a simple co-halfplanar ray rotation, (i.e., the image rays are co-halfplanar) then $\overline{\gamma}(f) = \gamma(f(a) \cup f(b))$.

(iii) If f is a plane ray rotation (i.e., all image rays are coplanar) which is "piecewise simple" and "directed", in the sense that there exists a "partition" $a = y_0 < y_1 < y_2 < \ldots < y_m = b$ of $[a,b]$, such that

(a) f is a simple co-halfplanar ray rotation on each $[y_i , y_{i+1}]$;

(b) each directed angle $(f(y_i) , f(y_{i+1}))$ has the same orientation;

then $\delta = (f(y_0) , f(y_1) , \ldots , f(y_m))$ is a directed angle, and $\overline{\gamma}(f) = \gamma(\overline{\delta})$, where $\overline{\delta}$ denotes the corresponding generalized angle.

That is, if a ray rotation corresponds to a generalized angle in this way, then its angular measure is the same as that of the corresponding generalized angle. This suggests that we might drop the bar over the γ, and regard $\overline{\gamma}$ simply as an "extension" of γ. With certain reasonable assumptions, it can be shown that each γ has a unique "extension".

Remark: A ray rotation can be determined in many ways. In order to determine a ray, we need only one point in addition to the endpoint Z. It can be shown that a ray rotation is thus determined by any continuous function defined on an interval $[a,b]$, and with values in $S - Z$. (I.e., by any curve whose range does not include the fixed endpoint Z.) In particular, if the range of such a curve is on a sphere with center Z (or a circle in the case of a plane ray rotation) then the length of the curve may be used as an angular measure of the rotation. If the sphere has radius "1", then this measure will be the same as the radian measure of the rotation.

We do not wish to discuss the problem of actually calculating angular measures for particular rotations. As in the case of curve length, the method of calculation depends largely on how the rotation function is specified. In the case of a simple repetitive situation, such as a rotating flywheel or a pendulum, the angular measure of a rotation can usually be calculated (by elementary means) directly from the definition; but in the case of more complex rotations, calculus methods are often used.

Plane Ray Rotations. Those ray rotations whose images lie in a fixed plane, are of special interest. We call these plane ray rotations. As mentioned above, some of these correspond to directed angles, and these rotations have the same angular measures as the angles to which they correspond. It seems reasonable to call such rotations "sensed", or "monotone", and we could, if necessary, distinguish between those which are "strictly monotone" (i.e., "locally" 1-1) and those which are not. You might find it interesting to try to formalize this monotone idea in different ways, remembering that it should be a "local" property only, because we want to express the idea of "keeping rotating in the same direction", and, at the same time, allow a "return to the same place".

One way of achieving this objective is to define a plane ray rotation f to be monotone if, for all similarly directed sub-intervals of $[a, b]$ on which f is a simple rotation, the directed "image angles" have the same orientation. A monotone rotation is strictly monotone if it is also locally 1-1. (I.e., each point x of the domain $[a,b]$ is contained in some open (or semi-open if $x = a$ or b) interval, on which f is 1-1.) It is not too difficult to show that the strictly monotone plane ray rotations are those which correspond to directed angles, as described above. (More precisely, a strictly monotone plane ray rotation corresponds to a geometric ray rotation.)

Monotone ray rotations can be oriented in the obvious way, and they separate into two classes, which we might designate as "clockwise" and "counter-clockwise".

Rotations of the Plane. Although it is only marginally connected with the subject of measurement, it seems advisable to tie together the notion of rotation that we have discussed, and the familiar notion of a "rotation of the plane". A rotation of the plane is a congruence (or rigid motion) of the plane which preserves orientation, and which is either the identity function, or leaves exactly one point (the "center of rotation") fixed. In other words, a rotation of the plane is a point transformation $r : H \to H$ of a plane H, which takes angles into congruent and similarly oriented angles, and which leaves either the whole plane, or a single point (Z say) fixed. Let Π_Z denote the group (with respect to composition) of those rotations which leave Z fixed. Each transformation $r \in \Pi_Z$ takes a ray with endpoint Z into a ray with endpoint Z, and hence r determines a 1-1 transformation (ρ say) of Γ_H, the set of all rays in H at Z. The set of those transformations, ρ, which are obtained in this way, is a group, $\vec{\Pi}_Z$; this group is, of course, isomorphic to Π_Z. If

$$f : [a,b] \to \Gamma_H$$

is a ray rotation, we can compose f and ρ to get another ray rotation

$$\rho f : [a,b] \to \Gamma_H .$$

All such ray rotations ρf (for fixed f; and ρ derived from any rotation $r \in \Pi_Z$) have the same angular measure. Moreover if f is monotone (and hence has an orientation) all related transformations ρf are monotone and have the same orientation.

We define ray rotations in a plane H to be "equivalent", if they have the same center and if they differ by composition with an element ρ of $\vec{\Pi}_Z$. This relation is easily shown to be an equivalence relation. The resulting equivalence classes are called plane rotations. (Notice that these are not the same as "rotations of the plane": the two ideas are closely related, but we have used different forms of expression in order to distinguish them.) Thus we can use the familiar rotations of the plane (point transformations) to make precise the intuitive idea that a rigid rotation (as a continuous motion) of the plane is completely determined by the motion of any ray in the plane which has its endpoint at the center of rotation. The angular measure of such a plane rotation is simply the (common) angular measure of the ray rotations which it contains. It is not too difficult to see that the correspondence of strictly monotone ray rotations with geometric ray rotations, induces a similar correspondence of strictly monotone plane rotations and geometric plane rotations.

Every plane rotation determines a unique rotation of the plane; (intuitively, that rotation of the plane which takes the initial position of each ray into the final position) but the converse is by no means true: infinitely many plane rotations lead to the same point transformation on the plane. We can assign a "measure" to a rotation of the plane by specifying an orientation, and by using the greatest lower bound of the measures of all of the monotone plane rotations which have the specified orientation, and which determine the given rotation of the plane; (intuitively, this is the measure of the shortest "angular path" in the specified "angular direction"). Thus we typically speak of rotations of 75° clockwise, 210° counterclockwise, and so on. When we do so we do not generally distinguish in our minds whether we are thinking of a rotation of the plane as a point transformation, or of the more complicated idea of a plane rotation, which is concerned not only with "where we are going", but also with "how we get there". If degree measure is used, this measure function establishes an isomorphism of the group (under composition) Π_Z, with the additive group of real numbers modulo 360. If radian measure is used, the isomorphism is with the additive group of reals, modulo 2π.

Finally, we remind you of the "dual" relationship between point transformations and coordinate transformations. Under this relationship a rotation of the plane corresponds to a rotation of coordinates. This correspondence may be used to "measure", or specify, a rotation of coordinates.

3-5. The Elementary Theory of Area

Area is a concept for which we have a strong intuitive feeling. There is an empirical aspect of area measurement, and there is a mathematical aspect. Our discussion will be almost entirely confined to the latter, but this does not imply that we are ignoring the empirical question: the mathematical theory enters into virtually every empirical area measurement.

It is convenient to break down the mathematical treatment into two parts. The first of these (the so-called elementary theory) deals with the areas of polygonal regions. This part of area theory may be thought of as corresponding roughly to the theory of length for broken segments, and you can detect many similarities in the treatment. However, area is a little more complicated than length, (both empirically and mathematically it is harder to compare regions with respect to their "areas", than it is to compare segments with respect to their "lengths") and we should not be surprised that this additional complication shows up in the development. On the other hand, as we shall see, we can make use of the theory of length which we have already developed, in order to simplify some parts of the treatment of area.

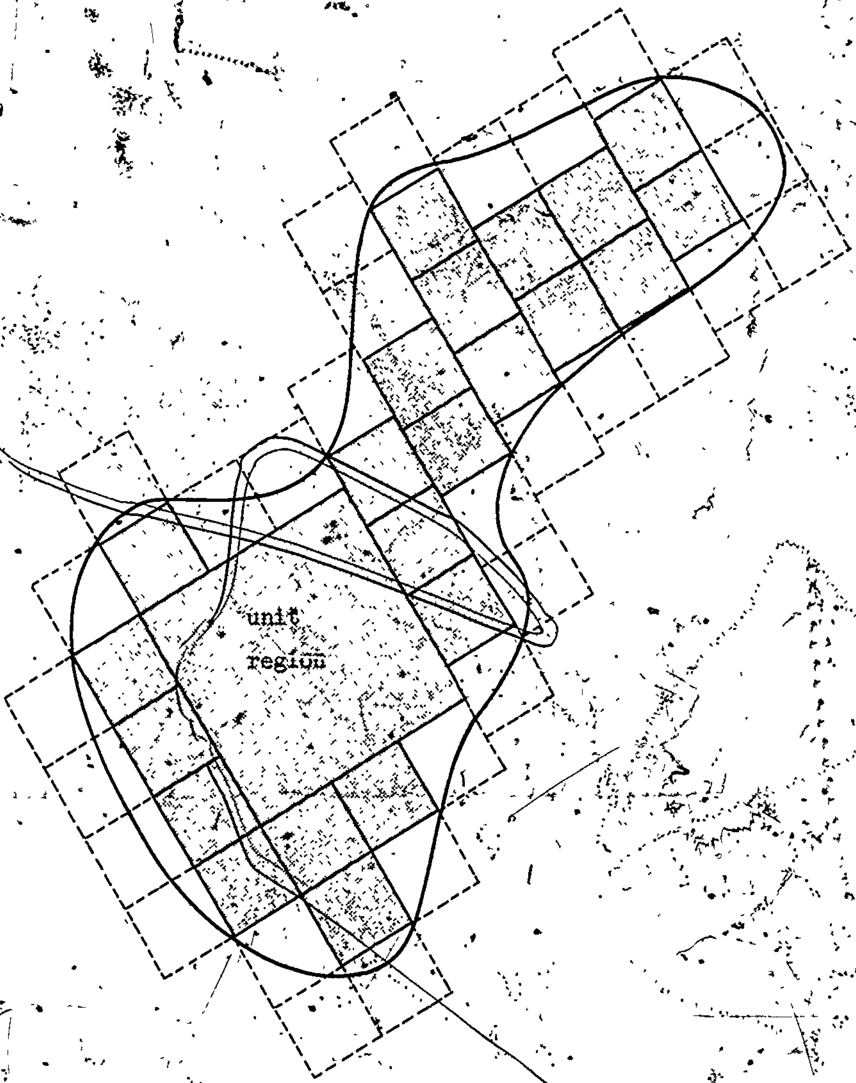
The secona part of our treatment will be concerned with extensions of the domain for area functions.

Area From an Empirical Standpoint. As with length and angularity, we first need to recognize a domain of objects which possess the (undefined) attribute which we think of as "area". At first we are likely to restrict our attention to "plane" objects (where "plane" has to be given empirical meaning in terms of some empirical test). For this domain of objects we devise some empirical procedures for "comparing" objects with respect to their "areas". Our procedures will almost certainly involve the assumption that objects with the same "size and shape" (a notion which corresponds to the mathematical notion of congruence, but which has to be established empirically) have equal "areas", and that two objects also have equal "areas" if they can be "decomposed" into pieces in such a way that there is a 1-1 correspondence of the pieces, with corresponding pieces having the same size and shape. By this means we might hope to separate the objects in the domain into "equal area" equivalence classes, and to "order" these classes. Difficulties arise because of problems of "fitting" (i.e., testing for "same size and shape"). Undoubtedly, with a sufficiently restricted domain, and with enough patience in "cutting and fitting", (probably using, implicitly, an assumed transitivity, so as not to destroy all objects in the process!) we could establish an approximate order structure on the domain and on the "equal area" equivalence classes.

An operation of "adding" or joining classes of objects is more straightforward--one may simply use the procedure of "disjoint unions"--and with a lot of work, and some imagination, we can envisage that the set of equivalence classes might have the structure of an ordered (even densely ordered) abelian semigroup. We would then look for structure-preserving functions which are defined on this domain, and which have positive real values, and we might seek to study the relationship between such functions (if any).

As for other measure situations, existence would be established by selecting any particular class as "unit", and by devising an appropriate procedure for "comparing" each object with this unit. We could, for example, take the class containing a particular (empirical) "square" region as unit.

and devise a procedure for comparing this with any other region by "fitting" as many "congruent" copies as we could, then taking an appropriate "sub-unit", fitting congruent copies of these to the "remainder", and so on. If we wished our "sub-units" to be "squares" also, we would find that each "unit square" could be most easily partitioned into n^2 congruent "squares", so that we might as well work with a "perfect square" (n^2, $n \geq 2$) as base, to find (empirically) an appropriate real number value (to base n^2) for the area of the region under consideration. We would also find, of course, that our "residues" at each fitting stage would not necessarily have less than unit area (cf. the situation for length), so there would be problems in establishing, with certainty, any position in the expansion to base n^2. (See diagram below, where we have taken $n = 3$, and where we have shown one stage only of a decomposition/fitting process.) It is clear that the fitting could generally be done in infinitely many ways, and we would hope to have empirical justification for the assumption that (with respect to a fixed "unit area") the value found for the "area" of a given region did not depend on how we went about the decomposition/fitting process. [Incidentally, if you check the statement below the diagram, you will be reminded of the importance of numerosity measurement (in this case, counting) in the establishment of more complicated measure functions.]



$1 + \frac{22}{9} \geq$ Area of region $\geq 1 + \frac{22}{9} + \dots$

If one object in the domain is a subset of another, clearly the area of the subset should be less than or equal to that of the whole set. By using such a monotonicity property we could find upper bounds for the value of an area function by considering those "disjoint" unions of unit regions and sub-units, which "cover" the region under investigation. (See unshaded square regions in figure, with partially dotted boundaries.) For successive stages of our process, the "inner approximation" yields (in general) an increasing number, while the "outer approximation" yields a diminishing upper bound. Taken together, these figures give some indication of the accuracy of our incomplete measurement, should we decide to stop after some finite number of steps.

If all of this were carried through carefully, we would find that the area function so established appeared to preserve the ordered semigroup structure of the domain, that the different area functions derived from different units were similar, that length measurement and area measurement were related in the well-known way, and so on. But long before we had done all this experimental work, we would probably have decided that we would switch to a mathematical "model", and devise a mathematical theory of area in the context of the model. When this is done the mathematical theory becomes a component in the process of empirical measurement, but it does not replace the need for making some physical measurements (such as length measurements and angle measurements) in order to find the areas of physical regions. Moreover, as mentioned earlier, the assumption of a suitable model situation (e.g., using euclidean geometry) is actually a hypothesis concerning the connection between empirical objects, operations, relations, and functions, and their "model" counterparts. Such a hypothesis can frequently be expressed in terms of the equality of certain functions; if these functions are represented on diagrams, then the hypothesis that certain composite functions are equal, is equivalent to the hypothesis that the diagrams are (at least in part) commutative. We may gain confidence in the usefulness of the model by testing commutativity for many domain elements and for various parts of the diagram, but our check involves physical measurement, so that agreement can never be established with certainty, even for those domain elements which we actually test. In the case of area, there is enough cumulative evidence to indicate that the mathematical theory which we will discuss provides a most useful model for the empirical situation, and that we are justified in using it as a component in the empirical measurement of area.

Area Functions for Rectangular Regions. The so-called "elementary" theory of area deals with the existence and properties of area measure functions for polygonal regions. Within that theory, we can distinguish an even simpler theory, which deals with the existence and properties of area functions for rectangular regions. Our reasons for dealing first with rectangular regions are twofold: this is the way in which most of us first encounter some semblance of a mathematical treatment of area; and there are significant omissions in most such elementary treatments.

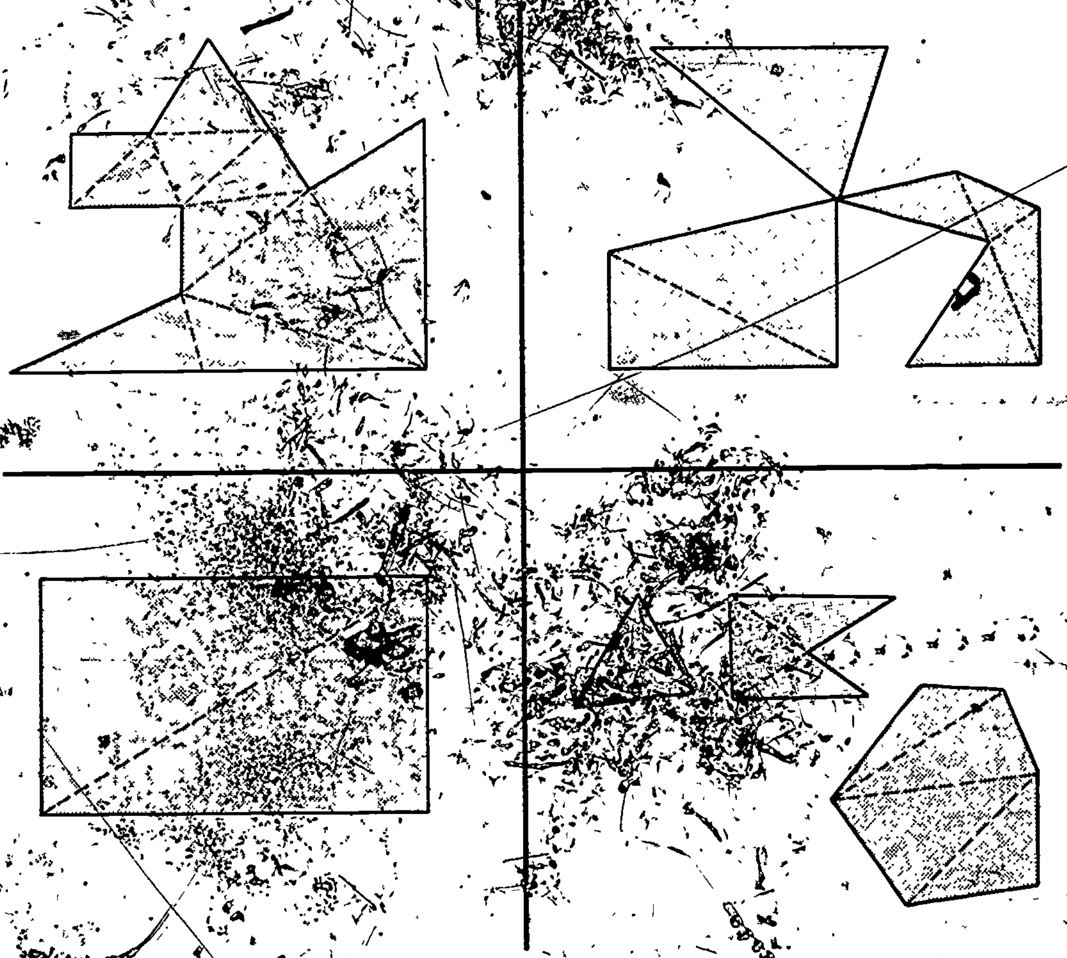
First of all, we need to make clear the context in which we shall discuss area theory: our treatment is limited to a discussion of area theory in euclidean geometry. We assume the archimedean and Cantor-Dedekind postulates throughout, so it does not matter whether we think of euclidean geometry as augmented synthetic geometry, metric geometry, or real cartesian geometry. We shall make considerable use of our previous work concerning the existence (in euclidean geometry) and properties of length functions and coordinate functions.

We comment that it is possible to develop an area theory purely within the framework of classical synthetic geometry (i.e., without the Cantor-Dedekind postulate), but by now you should be able to tell what the main differences (particularly concerning range questions) would be between this theory, and that which we develop below; having discussed this point at length in the treatment of linear and angular measures, we do not propose to deal with it further. We also remark that there are area theories for other than euclidean geometries, but we do not propose to deal with them here. For a treatment of area for hyperbolic geometry, see Chapter 24 of [14].

The domain for the elementary theory of area is the set of all polygonal regions. A polygonal region is a plane figure (i.e., set of points) which can be expressed as the union of a finite number of triangular regions, no two of which have any common interior point. (That is, the intersection of each two of the triangular regions is either empty, or consists of points of the triangles only.) A triangle is, of course, the union of three segments in the usual way, the interior of a triangle is the intersection of the interiors of the three angles which the triangle determines, and a triangular region is the union of a triangle and its interior. It follows that triangles and triangular regions are plane figures. Each of the four diagrams below illustrates a polygonal region, with the dotted lines indicating one way of expressing the region as a union of suitable triangular regions, as required by the definition. We point out two things:

(i) the expression is never unique: every polygonal region may be expressed as a finite union of suitable triangular regions in infinitely many ways; each such expression is called a triangulation of the polygonal region;

(ii) there is no "connectedness" condition (we haven't even defined the term), so that any finite union of disjoint coplanar polygonal regions is again a polygonal region. (If we wished we could define the topological notion of connectedness, and restrict attention to connected regions, but this is not necessary; in fact, if we did so restrict, we would probably end up by defining "formal sums", or "disjoint unions", as part of the development of area theory, so as to get a domain which is closed under a suitable "join" operation.



A rectangle is defined as usual, as the union of four coplanar segments $\overline{AB}$, $\overline{BC}$, $\overline{CD}$, $\overline{DA}$, such that the angles at A, B, C, D, are all right angles. The interior of the rectangle is the intersection of the interiors of its angles. A rectangular region is the union of a rectangle with its interior. The rectangle itself is called the boundary of the rectangular region. It is a simple matter to prove that a rectangular region is a polygonal region; (if you refer to the definitions, you will see that there actually is something to prove). Thus the set P_r of all rectangular regions is a subset of the set P of all polygonal regions.

We shall be interested first in area functions for rectangular regions. A function

$$f : P_r \to R^+$$

is called an area function for the set P_r of rectangular regions, if it satisfies the following conditions:

AR-1. (Congruence Condition). If r_1, r_2, are two congruent rectangular regions, (i.e., there is a rigid motion of space which maps one onto the other) then $f(r_1) = f(r_2)$.

AR-2. (Finite Additivity Condition). If $\{r_1, r_2, \dots, r_n\}$ is a finite set of rectangular regions such that

(a) $\bigcup_{i=1}^{n} r_i$ is a rectangular region;

(b) the intersection of any two different regions r_i, r_j, is either empty, or consists of boundary points (of r_i and r_j) only;

then $f(\bigcup_{i=1}^{n} r_i) = \sum_{i=1}^{n} f(r_i)$.

Remark: We shall be interested in area functions for rectangular regions, for polygonal regions, for triangular regions, and so on. Where there can be no confusion, we often refer to the area of a triangle, a rectangle, or a square, when we mean the area of the corresponding region. We do not do this for polygonal regions generally, because it is not necessarily true that the boundary (undefined as yet) of a polygonal region is a polygon, unless we extend the usual definition of "polygon". Moreover, for arbitrary polygons the notion of "interior" is more complicated than it is for simple convex polygons such as triangles and rectangles.

You will probably agree that the above requirements (AR-1 and AR-2) for an area function for rectangular regions, could hardly be reduced further. The important question which we must answer is whether or not there are any functions which satisfy these conditions. The existence of such functions and some of their most important properties, are proved in Theorems 3-5.1 to 3-5.3 below.

Theorem 3-5.1. Let r_0 be any square region, and let the segment σ_0 be one of its sides. Let λ_0 denote the length function (for segments) based on σ_0 as unit. (I.e., $\lambda_0(\sigma_0) = 1$.) Let r be any rectangular region, with sides σ_1 , σ_2 , and let $\lambda_0(\sigma_1) = a$, $\lambda_0(\sigma_2) = b$. Then, if f is any area function for P_r , f satisfies

$$f(r) = ab\, f(r_0).$$

Corollary 1. If f is an area function for P_r , then so is kf for every $k \in R^+$. Hence, with the same notation and assumptions as above, there is a unique area function f_0 for which $f_0(r_0) = 1$.

Corollary 2. Each area function for rectangles maps P_r onto R^+ , and it also maps the subset of all square regions onto R^+ .

Corollary 3. The function

$$\eta : \lambda_0 \to f_0$$

determined as in the theorem and Corollary 1, is a 1-1 correspondence of the set of length functions for segments and the set of area functions for rectangles.

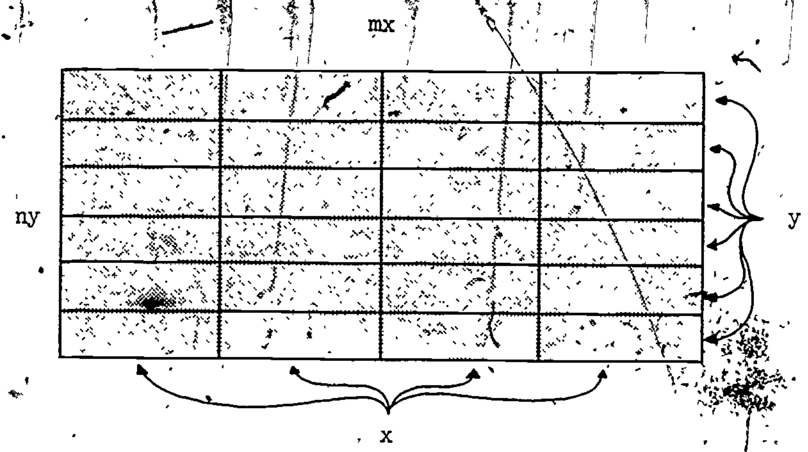
Corollary 4. If f_1 , f_2 , are two area functions for P_r , then there exists $k \in R^+$, such that $f_2 = kf_1$. (I.e., every two area functions are similar, and the set of all area functions for P_r is a ratio scale.)

Remarks:

1. Notice carefully that the existence of area functions on P_r is not asserted. We simply assert that, if there are any area functions, then they behave as stated in the theorem. The existence of area functions will be proved in Theorem 3-5.3. You will find that most elementary treatments of area theory for rectangles either postulate the existence of area functions, or they overlook the need to verify that the functions found from Corollary 1 above actually satisfy the requirements of the definition of an area function.

2. It is important to note that the particular length function λ_0 is not essential to the statement and proof of the theorem. The numbers a and b which appear are actually the ratios of the lengths of σ_1, and σ_2, to σ_0; and, as we saw in the discussion of length, these ratios can be defined without the use of any particular length function, and they are unit-free. (I.e., $\lambda(\sigma_1)/\lambda(\sigma_0) = a$ for every λ.)

Proof of Theorem. In this proof, all lengths are with respect to the function λ_0. Let r_1 be a rectangle whose sides have lengths x, y, and let r_2 be a rectangle whose sides have lengths mx, ny, where m, n, are positive integers. Then it is possible to express r_2 as the union of mn rectangles, each congruent to r_1, whose only pairwise intersections are boundary points. (See diagram: we shall not verify the above intuitively obvious assertion, but if you wish to fill in the details of this, and corresponding later statements concerning "decompositions", you will find it simplest to work in terms of cartesian coordinates. If you fill in the details you will find that we make considerable use of the additive property of length functions on "partitioned" segments.)



It follows from AR-1 and AR-2, that

$$(*) \qquad f(r_2) = mn\, f(r_1).$$

[Notice that this result involves "counting".] Hence if $x = y = 1$ (so that r_1 is congruent to r_0) and if r_2 has sides of positive integral lengths m, n, then, from $(*)$, $f(r_2) = mn\, f(r_0)$. On the other hand, if $x = \frac{1}{m}$, $y = \frac{1}{n}$, then r_2 is congruent to r_0, and we have, from $(*)$, $f(r_0) = f(r_2) = mn\, f(r_1)$, so that $f(r_1) = \frac{1}{mn} f(r_0) = \frac{1}{m}\frac{1}{n} f(r_0)$. (I.e., the rectangle whose sides have lengths $\frac{1}{m}$, $\frac{1}{n}$, has area $\frac{1}{m}\frac{1}{n} f(r_0)$.) Finally, if r_2 has sides of rational lengths $p = \frac{m}{t}$, $q = \frac{n}{s}$ (m, n, t, s, positive integers) and r_1 has sides of lengths $\frac{1}{t}$ and $\frac{1}{s}$, then a further application of $(*)$ gives

$$f(r_2) = mn\, f(r_1) = mn \cdot \frac{1}{t} \cdot \frac{1}{s} f(r_0)$$

$$= pq\, f(r_0).$$

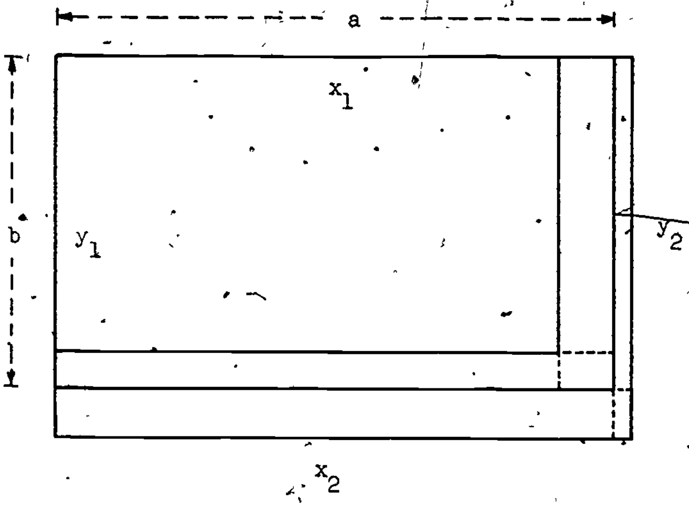
Thus we have proved the theorem if a and b are both rational.

Now let r be any rectangle, and let a, b, be the lengths of the sides of r. Then a and b are positive real numbers. Let k_1, k_2, be rational numbers such that $k_1 < ab$, $k_2 \geq ab$. Then it follows from the definitions of the real numbers and of multiplication for real numbers, that there exist rational numbers x_1, y_1, x_2, y_2, such that

(i) $x_1 < a$, $y_1 < b$, and $x_1y_1 = k_1 < ab$;

(ii) $x_2 \geq a$, $y_2 \geq b$, and $x_2y_2 = k_2 \geq ab$.

Now let r_1 , r_2 , be rectangles with sides of rational lengths x_1 , y_1 , and x_2 , y_2 , respectively, each rectangle having two adjacent sides collinear with (the same) two adjacent sides of r , as indicated in the diagram:



It is easy to show that (see dotted lines in diagram)

(i) there is a "decomposition" of r , such that r_1 is one of the rectangles in the "decomposition";

(ii) there is a "decomposition" of r_2 , such that r is one of the rectangles in the decomposition. (Of course it is possible that $r = r_2$.)

Moreover the necessary intersection conditions are satisfied, and hence from AR-2 ,

$$f(r_1) < f(r) \leq f(r_2) .$$

But r_1 , r_2 have sides of rational lengths so that

$$f(r_1) = x_1y_1f(r_0) = k_1f(r_0) < f(r) \leq f(r_2) = x_2y_2f(r_0) = k_2f(r_0);$$

$$\text{i.e., } k_1 < f(r)/f(r_0) \leq k_2 .$$

In other words, every positive rational less than ab is less than $f(r)/f(r_0)$, and every positive rational less than $f(r)/f(r_0)$ is less than ab . It follows from our definition of real numbers in terms of cuts, that

$$f(r)/f(r_0) = ab ;$$

i.e.,
$$\therefore f(r) = ab\, f(r_0) ,$$

which is what we set out to prove.

Proofs of Corollaries.

1. It is trivial to verify that if f satisfies AR-1 and AR-2 , then so then so does kf . Hence if f is an area function as in the theorem, so is $f_0 = (1/f(r_0))f$. Clearly $f_0(r_0) = 1$, and $f_0(r) = ab$; and any area function f_1 for which $f_1(r_0) = 1$, must agree with f_0 on all of P_r .

2. Let f be any area function, let r_0 be any square region, and let λ_0 be the length function (for segments) based on a side of r_0 as unit. Then, from the theorem, for any rectangle r , whose sides have lengths a , b , we have
$$f(r) = ab\, f(r_0) .$$
If z is any positive real number, choose positive real numbers x , y , such that $xy = z/f(r_0)$. (Clearly this can be done in infinitely many ways.) Then, if r is a rectangle whose sides have lengths x , y , (in terms of λ_0) we have $f(r) = xy\, f(r_0) = z$, and hence f is onto R^+ . Moreover, if we choose $x = y = \sqrt{z/f(r_0)}$, then r will be a square. Hence f maps the subset of square regions onto R^+.

3. If λ_0 is any length function, then the theorem and Corollary 1 show that there is a unique area function f_0 $(= \eta(\lambda_0)$, say) for which $f_0(r_0) = 1$, where r_0 is any square region whose side has length "1" , under λ_0 . If $\lambda_1 (\neq \lambda_0)$ is another length function, let r_1 be a square whose side has length "1" under λ_1 and length $a (\neq 1)$ under λ_0 . Then if $\eta(\lambda_1) = f_1$, we have $f_1(r_1) = 1$, and $f_0(r_1) = a^2 \neq 1$. Hence $f_0 \neq f_1$, and therefore η is 1-1 . We still have to show that η is onto. Let f_2 be any area function; then, from Corollary 2, f_2 is onto R^+ and there exists a square region r_2 such that $f_2(r_2) = 1$. If λ_2 is the unique length

function determined by the side of r_2, then $\eta(\lambda_2) = f_2$. Hence η is 1-1 and onto; i.e., η is a 1-1 correspondence.

4. If f_1, f_2, are area functions for P_r, let $r \in P_r$, with sides σ_1, σ_2. Then, from the theorem,

$$f_1(r) = [\lambda_0(\sigma_1) \cdot \lambda_0(\sigma_2)]\ f_1(r_0)\ ; \text{ and}$$

$$f_2(r) = [\lambda_0(\sigma_1) \cdot \lambda_0(\sigma_2)]\ f_2(r_0)$$

$$= f_1(r) \cdot k\ ,$$

where $k = f_2(r_0)/f_1(r_0)$. Hence $f_2 = kf_1$.

Remarks:

1. Chapter 13 of [14] contains an interesting variation on part of the above proof of Theorem 3-5.1 and its Corollaries.

2. Because of the 1-1 correspondence η of length functions and area functions, it is very convenient (but by no means necessary) to identify and name area functions in terms of the length functions to which they correspond under η. Thus, corresponding to the inch function λ_{in} we have a unique area function $\eta(\lambda_{in})$ which we usually denote as $f_{sq.in}$, (or f_{in^2}), and which we normally call the "square inch" area function; and when we say (for example) that a region r has an area of 27 square inches, we mean that $f_{sq.in}(r) = 27$. This relationship of length and area functions is very important, and we shall refer to it again in relation to "change of unit" questions, and in connection with the notion of "dimension".

From a practical standpoint, the 1-1 correspondence of length and area functions, coupled with the hypothesis that this mathematical discussion is relevant to empirical measurement (i.e., that this is a satisfactory "model" for part of the empirical situation) implies that it is not necessary to adopt separate empirical units for area. In other words, the "standards" for area (and, later, volume) are determined by the "standards" adopted for length.

3. We have already discussed the question of extending the domain for length functions, and we discovered that, for each extension situation, there was a 1-1 correspondence between the simple length functions for segments and their extensions to more complicated domains. We shall see later that area functions for rectangles can also be extended to larger domains, and that each area function for rectangles has its own unique extension. This uniqueness of extension for the length and area functions is most important, because it enables us to use (without ambiguity) the same names (inch, centimeter, square inch, square centimeter, etc.) for the extended functions. Moreover, we will be able to prove that the relationship of length and area functions established in Theorem 3-5.2 below, still holds for the extended functions.

4. In most elementary treatments, area functions are based (as suggested in our discussion of empirical area) on an assumed "unit" of area, which is usually a square region whose side is the unit of length. If this length unit is the "inch", the square region is called a "square inch" or an "inch square", and the corresponding area function is the "square inch function". This is another example of the correspondence of measure functions and units, with a "duality" of domain structure and measure scale structure. From a mathematical point of view it is usually simpler to deal with the functions than with "units".

Although we have not yet shown the existence of area functions for rectangular regions, it is convenient to continue first with the above train of thought, and prove a very important theorem relating area functions and the corresponding length functions by means of which the area functions are normally identified:

Theorem 3-5.2. If λ_1 and λ_2 are any two length functions for segments (so that, for some $k \in R^+$, $\lambda_2 = k\lambda_1$) and if $f_1 = \eta(\lambda_1)$, $f_2 = \eta(\lambda_2)$ are the corresponding area functions for rectangles, determined uniquely as above, then $f_2 = k^2 f_1$.

Proof. From Theorem 3-5.1, Corollary 1, if r is any rectangular region with segments σ_1, σ_2 for its adjacent sides, then

$$f_1(r) = \lambda_1(\sigma_1) \cdot \lambda_1(\sigma_2) \text{, and}$$

$$\begin{aligned} f_2(r) &= \lambda_2(\sigma_2) \cdot \lambda_2(\sigma_2) \\ &= (k\lambda_1)(\sigma_1) \cdot (k\lambda_1)(\sigma_2) \\ &= k(\lambda_1(\sigma_1)) \cdot k(\lambda_1(\sigma_2)) \\ &= k^2 f_1(r) \ . \end{aligned}$$

Hence $$f_2 = k^2 f_1 \ .$$

Remarks:

1. This simple theorem is the basis for all of the "formulas" relating to the conversion of "units" for area functions when (as is normally done) the area functions are identified by the length functions to which they correspond under η. It is important to remember that the formulas are normally given in terms of "units", rather than in terms of the corresponding functions. Thus the everyday statement

   12 inches equals 1 foot ,

   is equivalent, in functional language, to

   $$\lambda_{in} = 12\lambda_{ft} \ .$$

   Hence for the corresponding area functions we have

   $$f_{sq.in} = 12^2 f_{sq.ft} = 144 f_{sq.ft} \ ,$$

   or

   $$f_{(in)^2} = 144 f_{(ft)^2} \ ,$$

   which, in the everyday language of units, becomes the well-known, but usually rather obscure statement:

   144 square inches equals 1 square foot.

   [The obscurity relates to the usually ill-defined "scalar multiplication" of domain elements by numbers, and the ill-defined use of the word "equals"; as in the case of length, we can make such a statement clear by introducing a domain structure. If we introduce an equivalence relation on the domain of the area functions ($r_1 \sim r_2$ if and only if $f_1(r_1) = f(r_2)$ for any--and hence every--area function f) then it is a trivial matter to introduce (as we did for length) an "addition" (join) and a scalar multiplication by positive real numbers, for equivalence classes (or "units") of the domain, such that this set becomes an R^+-semimodule. If "square inch" denotes the relevant equivalence class, then "144 square inches" can be interpreted in terms of this multiplication. However it is generally simpler (mathematically) to use the language of functions, in which equality, and multiplication by real numbers, have well-established meanings, rather than to use the dual structure of the domain.]

The "reciprocal" relationship of functions and units is undoubtedly responsible for much of the confusion concerning "change of units", and "change of scale", but there is not much that can be done about it. It is a simple and obvious fact, that the larger the "unit" the smaller the functional values, whether for length, area, or any other class of similarity-related measure functions, with values in R^+, and with units.

2. The relationship between different length functions for segments and the corresponding area functions for rectangular regions can be conveniently illustrated by means of a commutative diagram. First of all, observe that if σ_1 and σ_2 are segments, then the pair (σ_1, σ_2) determines a unique congruence class of rectangular regions (regions with sides congruent to σ_1, σ_2, respectively). Hence, if D denotes the set of all segments, $\tilde{D}$ the set of congruence classes of segments, and $\tilde{P}_r$ the set of congruence classes of rectangular regions, then there is a unique function $\varphi : \tilde{D} \times \tilde{D} \rightarrow \tilde{P}_r$. (It is of interest to note that for each $(d_1, d_2) \in \tilde{D} \times \tilde{D}$, there are $\sigma_1 \in d_1$, $\sigma_2 \in d_2$; with σ_1 perpendicular to σ_2, and with a common endpoint, such that $\varphi(d_1, d_2)$ is the equivalence class of the rectangle which may be "identified" with the cartesian product $\sigma_1 \times \sigma_2$ of the segments σ_1, σ_2. Cf. the corresponding situation for numerosity measurement, where the cartesian product of domain elements (finite sets) was again a domain element.) The function φ is clearly onto, and it satisfies $\varphi(d_1, d_2) = \varphi(d_2, d_1)$.

Let λ_1, λ_2, (with $\lambda_2 = k\lambda_1$) be any two length functions, and let

f_1 , f_2 , be the corresponding area functions under η . Then we may regard f_1 , f_2 , as defined on $\tilde{P}_r$, since congruent regions have the same area. Moreover, as shown above, $f_2 = k^2 f_1$. The various results proved above show that the following diagram is commutative:

$$\begin{tikzcd}
(R^+ \times R^+) \arrow[rrr, "\text{multiplication}"] \arrow[dd, leftrightarrow, "\bar{k} \times \bar{k}"'] & & & R^+ \arrow[dd, leftrightarrow, "\bar{k}^2"] \\
& (\tilde{D} \times \tilde{D}) \arrow[ul, "\lambda_1 \times \lambda_1"'] \arrow[dl, "\lambda_2 \times \lambda_2"] \arrow[r, "\varphi"] & \tilde{P}_r \arrow[ur, "f_1"] \arrow[dr, "f_2"'] & \\
(R^+ \times R^+) \arrow[rrr, "\text{multiplication}"'] & & & R^+
\end{tikzcd}$$

Notice that the commutativity of the upper and lower "trapezoids" follows from the definition of η , which was equivalent to

$$\eta(\lambda_1) = f_1 = (\cdot)(\lambda \times \lambda)\varphi^{-1}$$

where $(\cdot)$ denotes the multiplication operation for R^+ .

You might find it instructive to start with any $(d_1, d_2) \in \tilde{D} \times \tilde{D}$, (or with any $\tilde{r} \in \tilde{P}_r$, using the inverse relation to φ ; this inverse is double-valued in general, but the diagram is commutative under each value) and trace the various "function-paths" for this element, to see just what the commutativity of the diagram implies. The relationships exhibited in this diagram are important in the understanding of such common statements as: "area is a two dimensional measure with respect to length"; and "area has the dimension of $(\text{length})^2$". We shall return later to this matter of "dimension", but meanwhile we observe that the "dimensionality" depends on the function η from the set of length functions to the set of area functions.

3. You might have wondered why we did not treat the area problem for rectangular regions as we did the length problem for segments, and the angular measure problem, and first try to establish an "area structure"

on P_r. The reason for this is that, while P_r certainly has (or may be given: see Remark 1 above) such a structure (i.e., an "area equivalence" relation under which the set of equivalence classes is an ordered R^+-semimodule) there is no straightforward way of establishing this structure directly, within the domain of rectangular regions. It is possible (as we shall see) to establish such a structure directly for polygonal regions, and the treatment will show the existence of such a structure for rectangular regions. Alternatively, we can show the existence as in Remark 1 above, by using any area function and "working backwards" from the structure of the positive reals.

We now come to the long-delayed proof of the existence of area functions for rectangular regions. In view of the fact that all of the results from Theorem 3-5:1 onwards were proved on the basis of "if there are any area functions...", if we can show the existence of at least one such function, then it will follow that there are infinitely many, related by positive-similarity transformations to each other, and that the set of all area functions is related by the 1-1 correspondence η, to the set of length functions for segments.

Theorem 3-5.3. There exists an area function for P_r; i.e., there exists a function $f : P_r \to R^+$ which satisfies the conditions AR-1 and AR-2.

Proof. Let r_0 be any square region, and let λ_0 be the length function for which a side of r_0 is the unit of length. Let r be any rectangular region, and let a, b, be the length of its sides, under λ_0. Then Theorem 3-5.1 tells us that if there is any area function for P_r, then there is an area function f_0 such that $f_0(r_0) = 1$, and such that $f_0(r) = ab$. This suggests that we take any length function λ_0, define a function

$$f: P_r \to R^+$$

by $f(r) = ab$, (where a, b, are the side lengths of r under λ_0) and test f to see whether or not it satisfies AR-1 and AR-2.

The verification of AR-1 is immediate: if $r_1 \cong r_2$ then the sides of r_1, r_2, have the same lengths (under every length function) and hence $f(r_1) = f(r_2)$.

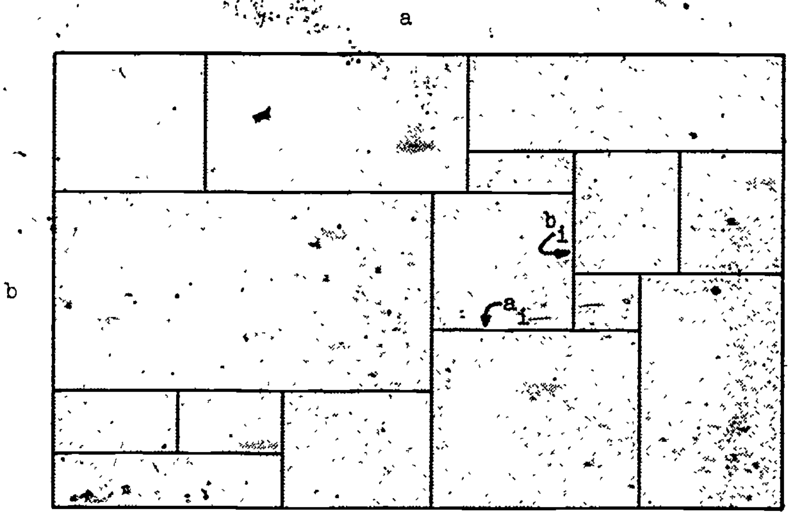
In order to verify AR-2 , we need to show that if a rectangular region r is the union of a finite number of rectangular regions r_i , $(i = 1 , 2 , \ldots , n)$ which satisfy the condition that the intersection of each pair r_i , r_j is either empty or consists only of boundary points of each, then

$$f(r) = \sum_{i=1}^{n} f(r_i) .$$

That is, if a , b , denote the lengths (under λ_0) of the sides of r , and if a_i , b_i , denote the lengths of the sides of r_i , we need to show that

$$ab = \sum_{i=1}^{n} a_i b_i .$$

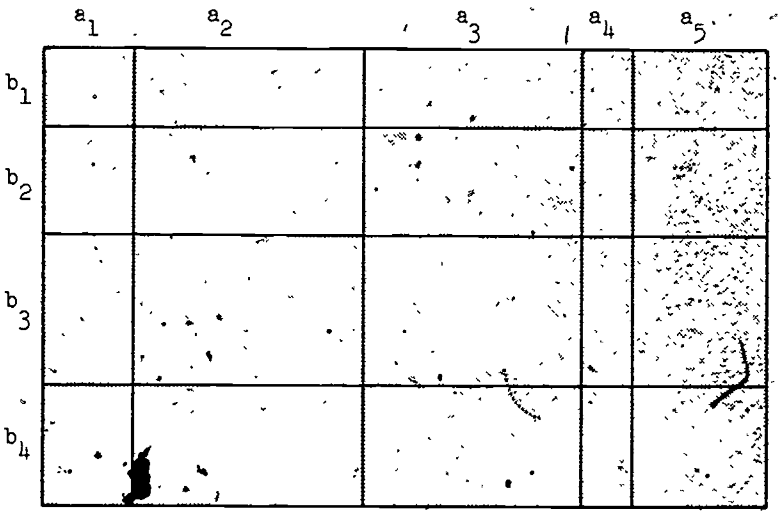
The following diagram of a typical decomposition which satisfies the requirements of AR-2, shows that this result is not completely obvious:



It is convenient to break down the proof into parts by means of lemmas. We first define a <u>simple decomposition</u> of a rectangular region r , to be a decomposition into rectangular regions by means of segments which are parallel and congruent to the sides of r , as indicated in the diagram below.

Lemma 1. If r is a rectangular region with sides of length a, b, and with a simple decomposition into rectangular regions, then

$$ab = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j$$



(We are not really dependent on the diagram, but it will save a great deal of work to assume that what the diagram suggests is in fact the case.)

Proof. The proof is a simple application (twice) of the distributive property of multiplication over addition, for the real numbers; and it also uses the known fact that length functions are finitely additive for a "partition" of a segment. (I.e., $\sum_{i=1}^{n} a_i = a$, and $\sum_{j=1}^{m} b_j = b$.)

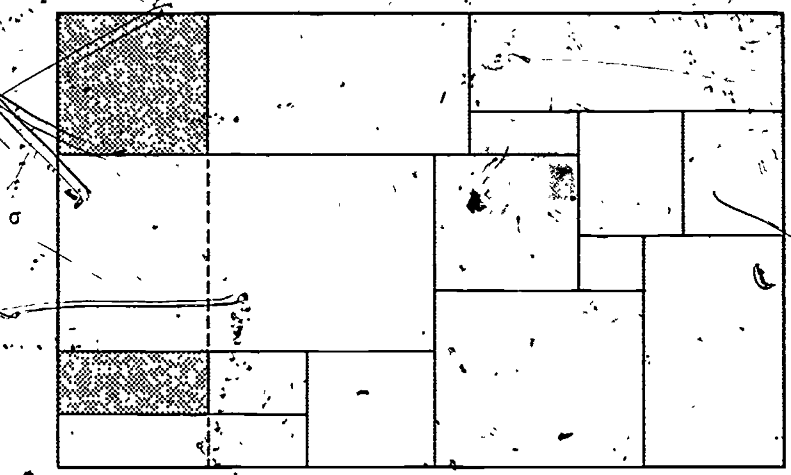
$$\begin{aligned}\sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j &= \sum_{i=1}^{n} [a_i(\sum_{j=1}^{m} b_j)] \\ &= \sum_{i=1}^{n} [a_i b] \\ &= b \sum_{i=1}^{n} a_i \\ &= ab .\end{aligned}$$

(If you have trouble with the summation notation, try writing out the sum "in full". All that we have done is to add up the function values down each "column" of rectangles, and then add the resulting sums for the large "column rectangles" using the constant factor b .)

We can complete the proof by reducing the general case (i.e., as for the first figure above), to the special case represented by a simple decomposition. We intend to do this by "completing" all of the segments which appear as sides. But here we encounter a minor difficulty: how can we be sure that, for any decomposition, the sides of the component rectangles, r_i , are parallel to the sides of the union, r ? This brings us to the second lemma:

Lemma 2. If r is a rectangle with a decomposition (as in AR-2) into a finite number of rectangles $r_i (i = 1 , 2 , \ldots , n)$, then the sides of each rectangle r_i are parallel to the sides of r .

Proof. We shall sketch a proof of this strongly intuitive property, which can be proved in various ways. Our proof uses an induction on the number (n) of regions. If $n = 1$, $r_1 = r$ and the result is trivially true for all r. Assume that the result holds (for all rectangles r) for all decompositions which have less than n rectangles $(n > 1)$ and let r be any rectangular region, with a decomposition (as in AR-2) into n rectangular regions $r_i , (i = 1 , 2 , \ldots , n)$. Then (see diagram below) all regions which "adjoin" a fixed side σ of r (the left side in the diagram) must have their sides parallel to the sides of r . (This follows directly from the well-known transitivity of the relation "parallel or collinear" on line segments, and from the fact that perpendiculars, in a fixed plane, to parallel or collinear line segments, are themselves parallel or collinear.)

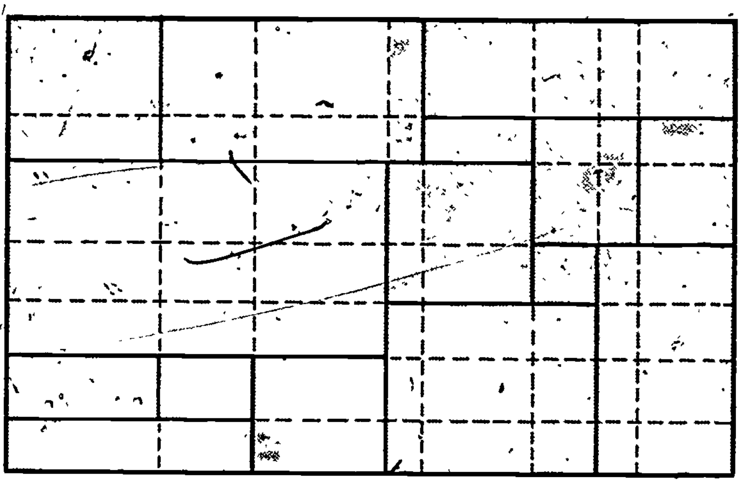


Because n is finite, there is one (or more) of these rectangles, "adjacent" to σ, which has least "width". (The "width" is, of course, the length of the side perpendicular to σ.) "Complete" the corresponding segment parallel to σ (shown dotted to avoid confusion), and let r' denote the "residual" rectangle to the right of, and including this "new" segment. Then the original decomposition of r, and the "new" segment, yield a decomposition of r'. Moreover at least one rectangular region (two, shown with darker shading, in the case represented by the diagram) of r which is adjacent to σ, does not appear in r'; and the remaining regions of r which are adjacent to σ, correspond 1-1 with "reduced" regions of r'. That is, r' has less than n regions. Hence, from the "inductive hypothesis", all sides of rectangles in the decomposition of r' are parallel to the sides of r'. It follows that all sides of rectangles in the original decomposition of r, are parallel to the sides of r.

We now complete the proof of the theorem by a double application of Lemma 1. Given any rectangle r, and any decomposition (as in AR-2)

$$r = \bigcup_{i=1}^{n} r_i ,$$

"complete" this decomposition to a simple decomposition of r by "completing" all of the sides of all rectangles r_i. (See diagram on the following page.)



For this simple decomposition of r, let r_{jk} be a single rectangular region with sides of length x_j, y_k. Then, from Lemma 1,

$$ab = \Sigma\,\Sigma\, x_j y_k .$$

(To avoid making a simple idea seem complicated, we do not develop a detailed notation. The double summation sign simply means that we add up the products $x_j y_k$ for each of the rectangles in the simple decomposition of r which results when all segments are "completed".)

We now observe that the "completion" of the original decomposition of r also yields a simple decomposition for each of the rectangles r_i in the original decomposition. Moreover each rectangle r_{jk} appears in the resulting simple decomposition of exactly one of the original r_i. Hence the sum $\Sigma\,\Sigma\, x_j y_k$ can be broken down: we can first add separately for all rectangles r_{jk} in each r_i, and then add for all r_i. Thus we get

$$\Sigma\,\Sigma\, x_j y_k = \sum_{i=1}^{n} \Big(\sum_{r_{jk} \subset r_i} x_j y_k\Big)$$

$$= \sum_{i=1}^{n} a_i b_i \quad \text{(from Lemma 1.)}$$

But we have already seen that $\Sigma\Sigma\, x_j y_k = ab$ (Lemma 1 again) and hence we have: $f(r) = ab = \sum_{i=1}^{n} a_i b_i = \sum_{i=1}^{n} f(r_i)$. In other words, f satisfies AR-2, and hence f is an area function. That is, there exist area functions for rectangular regions.

Remarks:

1. The above proof is much harder to write down than it is to work out. Don't be deceived into thinking that it is difficult!
2. If (as is common in elementary work) the area of a rectangle is simply defined to be the product of the lengths (under some length function) of adjacent sides, then the above theorem proves that such a definition is consistent with the additivity property AR-2. This property is generally used in elementary work without being explicitly stated.

Area Functions For Polygonal Regions. We return now to what is usually called the "elementary theory of area": the study of the existence and relationships of area functions for polygonal regions. This theory, which is strongly geometrical in character, has an extensive history, partly related to the fact that one of its most famous theorems (Bolyai's Theorem: that polygonal regions have equal areas only if they are "piecewise congruent under decomposition"--see below) has no counterpart in the theory of volume, a fact which was only proved (by M. Dehn) at the beginning of the present century.

Area theory for polygonal regions can be developed in different ways. One way is to proceed as we did in developing the notion of area for rectangular regions, by defining "area function" and then establishing the existence and properties of such functions. Relations of "equal area" and "greater area" can then be defined by using these functions.

Another way is to proceed (as we did for length) by defining a relation of "equal area" and developing a semi-group structure on the resulting set of equivalence classes before asking whether there are functions (with positive real values) which preserve all of this structure. Such functions are, of course, area functions. Each of these approaches requires about the same amount of work to carry through.

We adopt the first of these methods of approach, because we wish to emphasize the closeness of the method (in spite of the more complicated details) to that which we used for the discussion of area for rectangular regions. You should also refer back to the discussion (in Section 2-8) of length for broken segments: there is considerable similarity between that discussion and the one which we now give for the area of polygonal regions, although the corresponding area theorems are somewhat harder to prove.

As above, we denote by P the set of all polygonal regions, and by P_r the subset of rectangular regions. We define a function $f : P \to R^+$ to be an area function for P if it satisfies the following conditions.

AP-1. (Congruence condition). If p_1, p_2, are two congruent polygonal regions, then $f(p_1) = f(p_2)$.

AP-2. (Additivity condition). If two polygonal regions p_1, p_2 are such that their intersection is either empty, or consists of boundary points only, then

$$f(p_1 \cup p_2) = f(p_1) + f(p_2) .$$

Remarks:

1. In this definition of area function we have referred to "boundary points". The notion of boundary is a topological notion: we could define it as the union of those segments which (in some triangulation) are on the boundaries of only one region, but then we would need to show that this definition is independent of the particular triangulation. Alternatively, we may define an interior point of a polygonal region to be a point with the property that there exists a circular region (easily defined) with that point as center, and wholly contained in the polygonal region: boundary points are then those points which are not interior points. Clearly, this definition is independent of any triangulation. It is not hard to devise other definitions (e.g., a boundary point is one which has the property that there is some segment in the plane of the region, for which the boundary point is an interior point, and such that all points of the segment on one "side" of the boundary point lie outside of the region) but it is not particularly easy to work with any of them. In the subsequent discussion we will use the notion of boundary and interior intuitively, but it is important to know that these ideas can be handled precisely.

2. A simple proof by mathematical induction enables us to extend condition AP-2 to a polygonal region which is the union of a finite number of polygonal regions whose interiors are pairwise disjoint. We refer to the property proved by this extension as finite additivity, and we refer to the finite set of polygonal regions (with pairwise disjoint interiors) as a polygonal decomposition of their union.

In an effort to help you to follow the sequence of ideas involved in the development of area theory for polygonal regions, we summarize the main steps:

(i) We observe that conditions AP-1 and AP-2, applied to the subset P_r of rectangular regions, imply conditions AR-1 and AR-2. Hence, if there are any area functions for P, they must be extensions of area functions for rectangular regions.

(ii) We prove that if f is any area function for P (and hence, suitably restricted, for P_r, so that there is a "corresponding" length function λ, with $\eta(\lambda) = f$, as above) then the value of f on a triangular region must be $\frac{1}{2}bh$, where b and h are the lengths (under λ) of any base and corresponding altitude.

(iii) It follows from finite additivity, that the value of f on any polygonal region must be the sum of the values $\frac{1}{2}bh$ on the triangles of any triangulation. Thus if there is any area function, it must behave this way.

(iv) This suggests that, for a given polygonal region, and a fixed length function, we should try to show that the sum $\Sigma(\frac{1}{2}bh)$ (taken over all triangular regions in a triangulation) is independent of the particular triangulation, so that we can use it to define a function from P to R^+. If this is to be an area function, we must show that it satisfies AP-1 and AP-2.

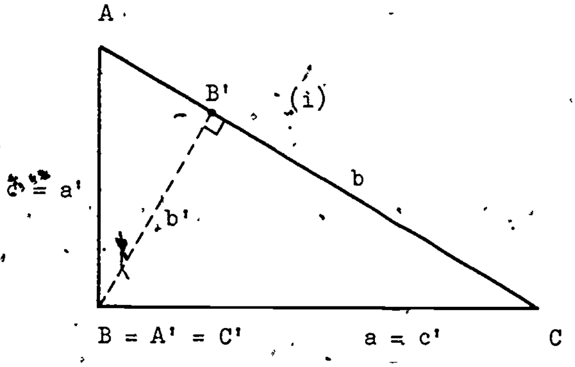
We now proceed to fill in the details. To keep the treatment within reasonable bounds we will give formal definitions and statements of theorems, but we will only sketch most of the proofs. Moreover, to some extent we will depend on diagrams, and assume that what appears to be the case actually is the case. To carry out all of this in complete detail is a lengthy, and rather tedious, undertaking.

Theorem 3-5.4. If $f : P \to R^+$ is an area function for polygonal regions, then $f|P_r$ (i.e., the restriction of f to the domain P_r) is an area function for rectangular regions.

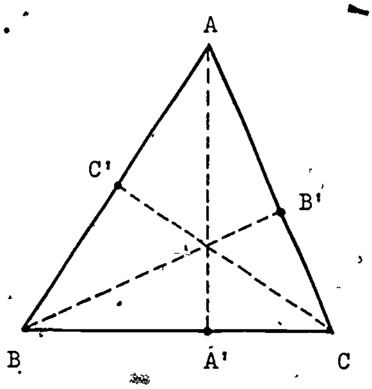
Proof. AP-1 implies AR-1, and AP-2 implies the finite additivity of $f|P_r$.

Theorem 3-5.5. If λ is any length function, and if a, b, c; a', b', c', denote the lengths (under λ) of the sides and corresponding altitudes of a triangle ABC, then $aa' = bb' = cc'$.

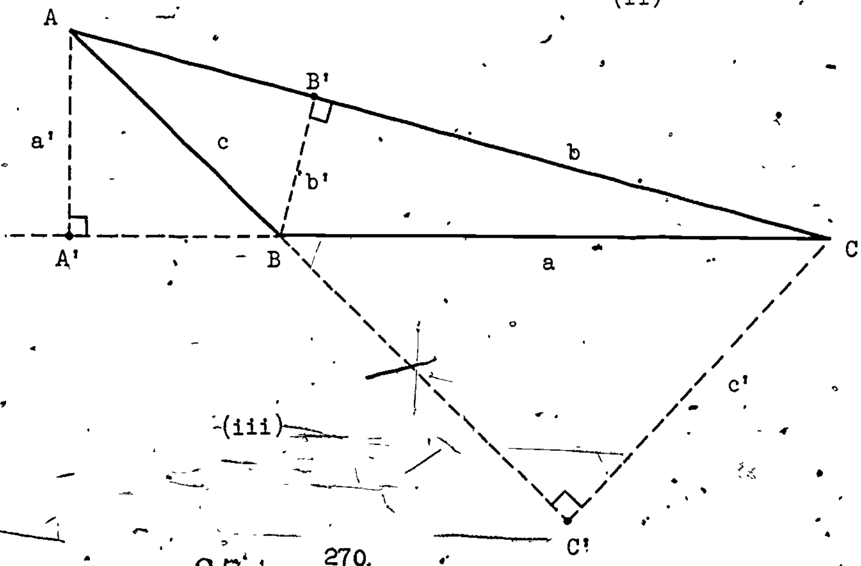
Proof. The diagrams below illustrate the situation (i) when one angle is a right angle; (ii) when all angles are acute; (iii) when one angle is obtuse. These are the only possibilities.



(i)



(ii)



(iii)

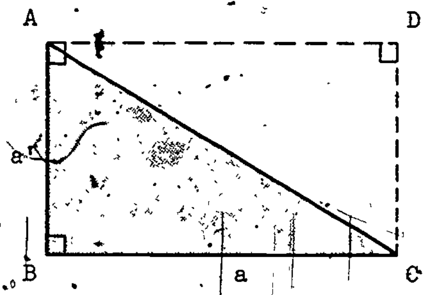
In case (i) $aa' = cc'$ is trivial, and $aa' = bb'$ follows from the similarity of ΔABC and $\Delta BB'C$, which gives $a'/b = b'/a$. In cases (ii) and (iii), we have $\Delta ABA'$ similar to $\Delta CBC'$ so that $a'/c = c'/a$, whence $aa' = cc'$; that each product is equal to bb' is shown similarly.

Theorem 3-5.6. If f is any area function for P, and if λ is the length function which corresponds to the area function $f|P_r$ (i.e., f has value "1" on the square region whose side length under λ is "1"), then, for a triangle ABC with sides and altitudes of lengths a, b, c; a', b', c', under λ, we have

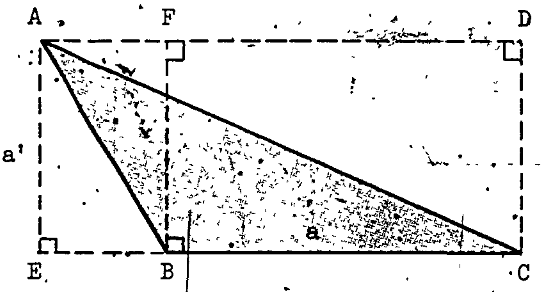
$$f(\blacktriangle ABC) = \frac{1}{2} aa' = \frac{1}{2} bb' = \frac{1}{2} cc' .$$

(The symbol $\blacktriangle ABC$ denotes the triangular region bounded by ΔABC.)

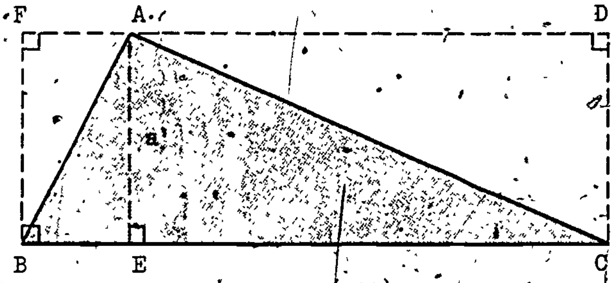
Proof. From Theorem 3-5.5, it is sufficient to prove that $f(\blacktriangle ABC) = \frac{1}{2} aa'$. Three cases need to be considered. These are represented by the diagrams (i), (ii), (iii).



(i)



(ii)



(iii)

In case (i), $\triangle ABC \cong \triangle CDA$; hence, from AP-1, $f(\blacktriangle ABC) = f(\blacktriangle CDA)$. But, from AP-2, $f(\text{rect. } ADCB) = f(\blacktriangle ABC \cup \blacktriangle CDA)$

$$= f(\blacktriangle ABC) + f(\blacktriangle CDA)$$

$$= 2f(\blacktriangle ABC)$$

Hence $f(\triangle ABC) = \frac{1}{2} f(\text{rect. } ADCB)$

$$= \frac{1}{2} aa' .$$

In case (ii), for similar reasons we have $f(\text{rect. } ADCE) = f(\text{rect. } AFBE) + f(\text{rect. } FDCB)$;

therefore $2f(\blacktriangle AEC) = 2f(\blacktriangle AEB) + aa'$;

therefore $2f(\blacktriangle AEB \cup \blacktriangle ABC) = 2f(\blacktriangle AEB) + 2f(\blacktriangle ABC) = 2f(\blacktriangle AEB) + aa'$;

therefore $2f(\blacktriangle ABC) = aa'$;

therefore $f(\blacktriangle ABC) = \frac{1}{2} aa'$.

Case (iii) is left for you to complete for yourself.

We have now shown that, if there are any area functions for polygonal regions, then each gives an area value to triangular regions which agrees with the familiar "formula". Because of the additivity condition AP-2, this means that, if there is to be an area function f for polygonal regions, then for a given polygonal region p , and any triangulation $p = \cup\delta_i$. (where each δ_i is a triangular region) the value of f on p must be $\Sigma f(\delta_i)$ or $\Sigma(\frac{1}{2} a_i a_i')$, where a_i , a_i' , are as above for δ_i . This suggests that we test this possibility to see if a function defined in this way will satisfy AP-1 and AP-2.

Let λ_0 be any length function. For any triangular region δ , let x , x' be the lengths (under λ_0) of a "base" and a corresponding "altitude". Define a function f_0 by

$$f_0(\delta) = \frac{1}{2} xx' .$$

(As we saw above, this value does not depend on the choice of base.) Let p be a polygonal region with a triangulation $\{\delta_i\}$, $i = 1, 2, \ldots, n$. We extend the domain of the function f_0 to include each <u>triangulated</u>

polygonal region $(p, \{\delta_i\})$, by defining

$$f_0(p, \{\delta_i\}) = \sum_{i=1}^{n} f_0(\delta_i) .$$

For each triangulated region this gives a unique value, and hence a function f_0 is defined for triangulated regions. We wish to prove that the value of f_0 on $(p, \{\delta_i\})$ depends only on p (and, of course, on λ_0), and not on the triangulation $\{\delta_i\}$. The proof of this is rather complicated in detail, but the idea behind it is similar to that which we used in proving the corresponding result for rectangular regions. (Theorem 3-5.3.)

Theorem 3-5.7. Let $\{\delta_i\}$, $\{\overline{\delta}_j\}$ be two triangulations of a polygonal region p, $(i = 1, 2, \ldots, m ; j = 1, 2, \ldots, n)$. Then

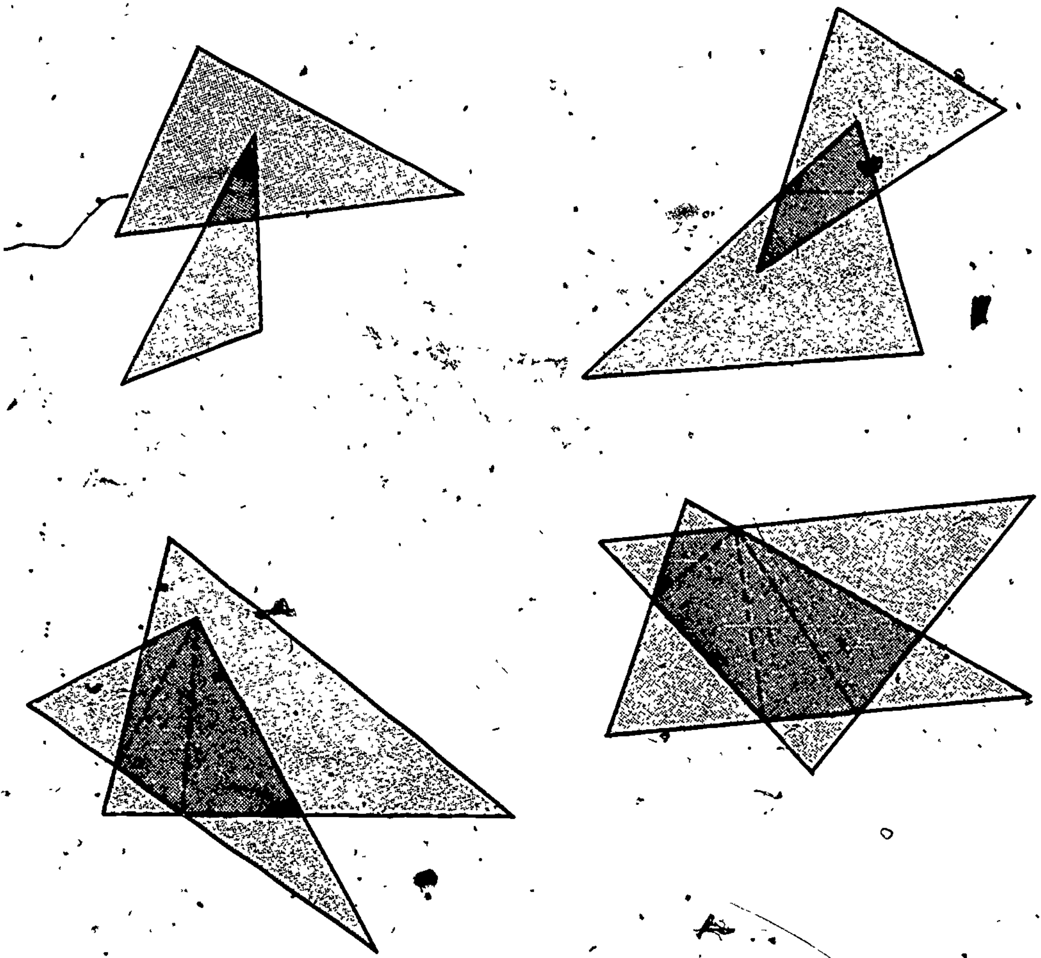
$$f_0(p, \{\delta_i\}) = f_0(p, \{\overline{\delta}_j\}) . \quad (\text{I.e., } \sum_{i=1}^{m} f_0(\delta_i) = \sum_{j=1}^{n} f_0(\overline{\delta}_j).)$$

In order to make the proof easier to follow, we prove two lemmas. We first define a subdivision of a triangulated polygonal region $(p, \{\delta_i\})$, to be a triangulated polygonal region $(p, \{\delta'_k\})$, such that each triangular region δ'_k is contained in some δ_i; and we refer to the triangulation $\{\delta'_k\}$ as a refinement of the triangulation $\{\delta_i\}$.

Lemma 1. Every two triangulations $\{\delta_i\}$, $\{\overline{\delta}_j\}$ of a polygonal region p, have a common refinement.

Proof. As above, let i run from 1 to m, and j run from 1 to n. Then we have (from the elementary algebra of sets)

$$p = \bigcup_{i=1}^{m} \delta_i = \bigcup_{j=1}^{n} \overline{\delta}_j = (\bigcup_{i=1}^{m} \delta_i) \cap (\bigcup_{j=1}^{n} \overline{\delta}_j) = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} (\delta_i \cap \overline{\delta}_j) .$$ The sets δ_i and $\overline{\delta}_j$ are triangular regions. Each is convex, hence their intersection must be a convex polygonal region with three, four, five, or six sides. Some examples are shown in the diagram below. (In case you are not familiar with the notion of convexity, a convex region is one with the property that it contains the segment joining any two of its points; you may easily prove that the intersection of convex regions is convex.) Any convex polygonal region can, of course, be triangulated: the dotted lines

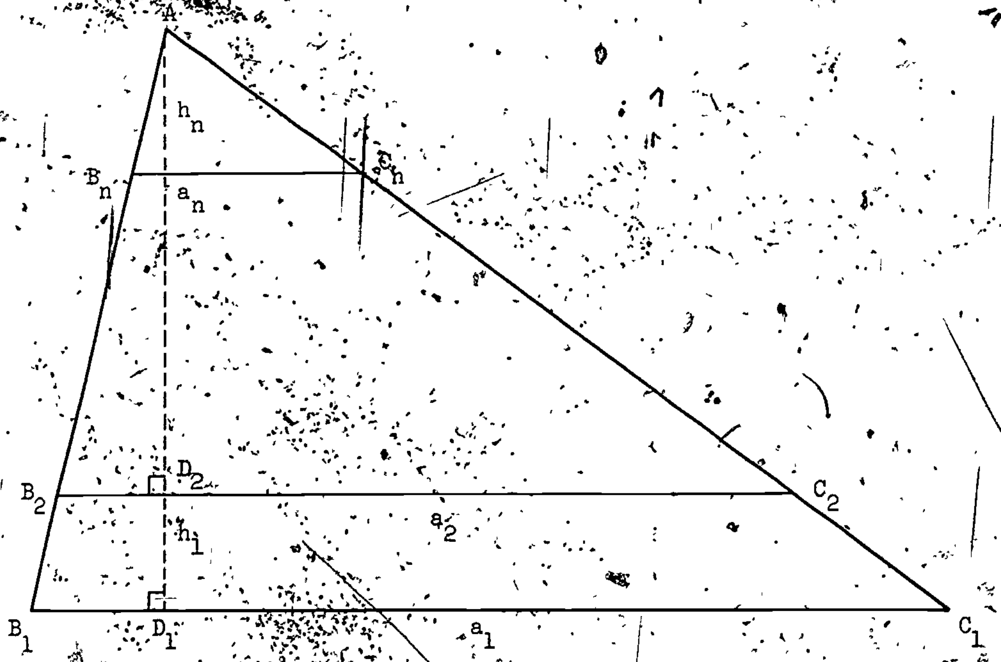


in the diagrams indicate one way of doing this for each intersection shown. To keep the diagrams simple, we do not attempt to show the resulting triangulation of p beyond a single intersection $\delta_i \cap \bar{\delta}_j$ (shown with darker shading). Moreover, we do not illustrate all of the various ways in which 3, 4, 5, or 6-sided convex regions can arise as intersections of two triangular regions: this does not affect the proof. By triangulating each $\delta_i \cap \bar{\delta}_j$, and by observing that two different convex regions $\delta_{i_1} \cap \bar{\delta}_{j_1}$, $\delta_{i_2} \cap \bar{\delta}_{j_2}$, intersect (if at all) only in their boundaries, we can obtain a triangulation of p which is a common refinement of $\{\delta_i\}$ and $\{\bar{\delta}_j\}$.

Remark: Now that we know that any two triangulations of a polygonal region p have a common refinement, Theorem 3-5.7, will be proved if we can prove that the value of f_0 on a triangulated polygonal region $(p, \{\delta_i\})$ is the same as the value of f_0 on any refinement of $(p, \{\delta_i\})$. Now $f_0(p, \{\delta_i\})$ is simply the sum of the values of f_0 on each triangular region δ_i, and the value of f_0 on a refinement is the sum of the values of f_0 on all triangular regions in the refinement. To find this latter sum, we may add first for all triangular regions in each δ_i separately, and then add these partial sums. Hence the theorem will be proved if we can prove that f_0 has the same value for a single triangular region and for any triangulation of that region. In order to show this we need another lemma.

Lemma 2. Let AB_1C_1 be a triangle, and let $\overline{B_2C_2}$, $\overline{B_3C_3}$, ..., $\overline{B_nC_n}$ be segments parallel to $\overline{B_1C_1}$, as shown. Let a_i be the length of $\overline{B_iC_i}$ (with respect to a fixed length function) and let h_i, be the distance between $\overline{B_iC_i}$ and $\overline{B_{i+1}C_{i+1}}$ as indicated $(i=1, 2, \dots, n-1)$, with h_n, the altitude of $\triangle AB_nC_n$. Then

$$a_1(AD_1) = (a_1 + a_2)h_1 + (a_2 + a_3)h_2 + \dots + a_nh_n .$$



Proof. We give an outline of the proof only; a detailed proof requires an induction on n. $\triangle AB_1C_1$ is similar to $\triangle AB_2C_2$. Hence (using the finite additivity of length functions)

$$\left(\sum_{i=1}^{n} h_i\right) / a_1 = \left(\sum_{i=2}^{n} h_i\right) / a_2$$

therefore

$$a_1 \sum_{i=2}^{n} h_i = a_2 \sum_{i=1}^{n} h_i$$

$$= a_2h_1 + a_2 \sum_{i=2}^{n} h_i .$$

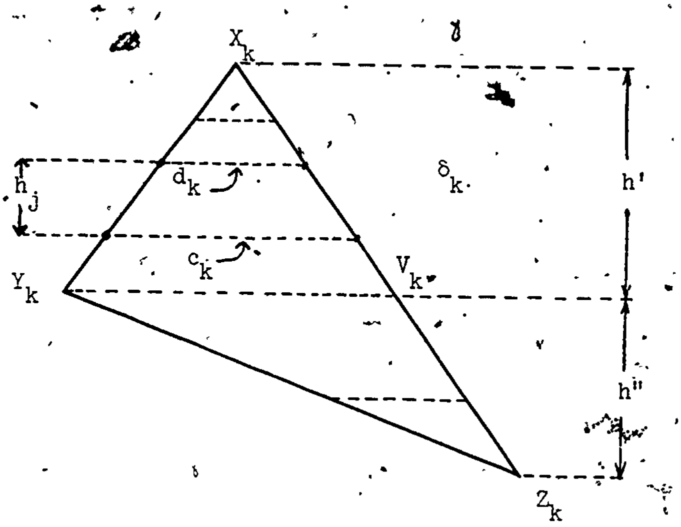
Hence, adding a_1h_1 to each side,

$$a_1(AD_1) = a_1 \sum_{i=1}^{n} h_i = a_1h_1 + a_2h_1 + a_2 \sum_{i=2}^{n} h_i$$

$$= (a_1 + a_2)h_1 + a_2 \sum_{i=2}^{n} h_i .$$

Repeating this process, we obtain the required result: it is not hard to give a complete proof by induction. This rather simple lemma is the key to the proof of the theorem.

Proof of Theorem 3-5.7 As observed earlier (see remark following the proof of Lemma 1) it is sufficient to prove the theorem for a single triangular region, δ, and a triangulation $\{\delta_k\}$ of δ. Let $\delta = \triangle ABC$, and let x, x' be the lengths (with respect to a fixed length function) of $\overline{BC}$, and the corresponding altitude. Through every vertex of each triangle δ_k in the triangulation, draw the segments which are parallel to $\overline{BC}$ and whose end points are on $\overline{AB}$, $\overline{AC}$, respectively. These segments will decompose a particular region δ_k as indicated on the following page.

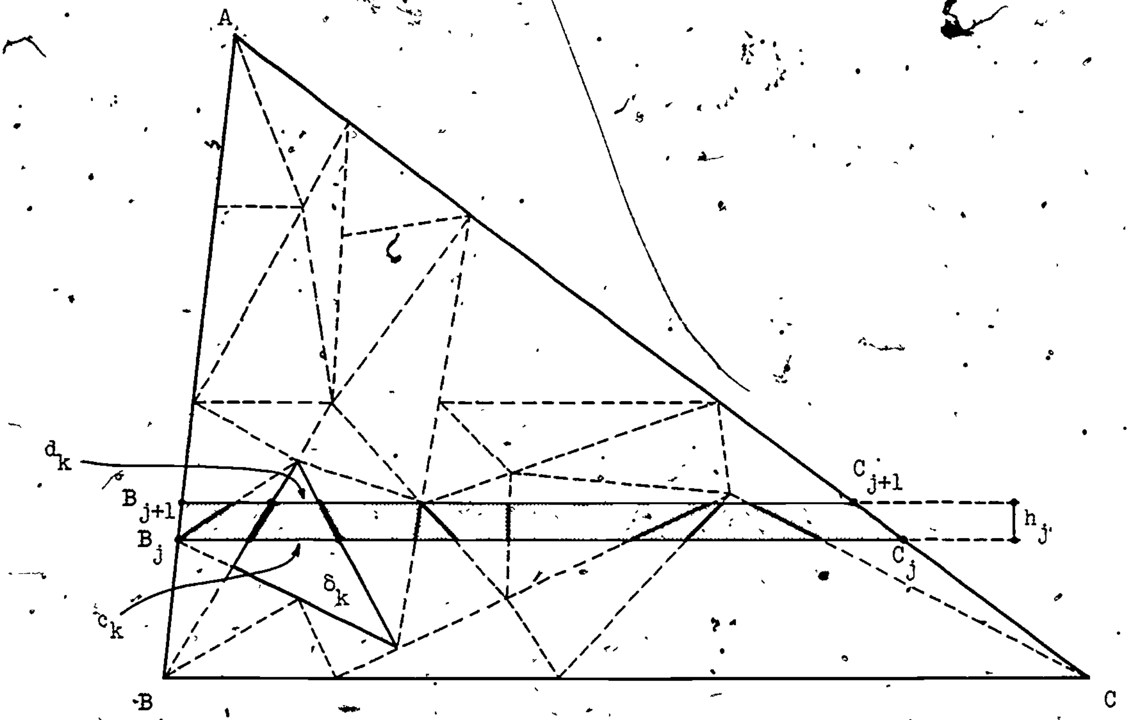


For δ_k, if q denotes the length of the altitude from Y_k to $\overleftrightarrow{X_kZ_k}$, we have (using mainly Theorem 3-5.5)

$$\begin{aligned} x_kx_k' = q(X_kZ_k) &= q(X_kV_k + V_kZ_k) \\ &= q(X_kV_k) + q(V_kZ_k) \\ &= h_k'(Y_kV_k) + h_k''(Y_kV_k). \end{aligned}$$

Hence, for the region δ_k, the product x_kx_k' can be broken down (as in Lemma 2) as $\Sigma h_j(c_k + d_k)$. (We do not attempt to develop a detailed notation, and we recall that this sum has to be properly interpreted for the "extreme" triangular regions at X_k and Z_k.)

We point out that what we have achieved so far, through the use of Lemma 2, is to break down the number x_kx_k' into a sum of separate products of numbers corresponding to certain lengths obtained from the regions into which δ_k is divided by the segments parallel to $\overline{AB}$. We now complete the proof by "adding" these subproducts across each full "parallel strip" of δ, and then we use Lemma 2 again on this "parallel strip" decomposition of δ. At no stage are we adding "areas"--whenever numbers are added they are obtained from length functions, and we make strong use of the additive property of length functions on "decomposed" segments. As in the case of rectangular regions, we also use the distributive property of multiplication over addition. The diagram below indicates the regions δ, δ_k, and one full "parallel strip" which includes part of δ_k. To avoid cluttering



up the diagram, only one "parallel strip" is shown (shaded) and the boundaries of all triangular regions (except δ_k) are shown dotted, except where they intersect the strip shown.

We now have

$$\sum_k x_k x_k' \doteq \sum_k (\Sigma\, h_j\, (c_k + d_k)) .$$

For fixed h_j, the terms $\Sigma\, h_j c_k$ and $\Sigma\, h_j d_k$ in this total sum can be collected, and the common factor h_j can be taken out. The resulting $\Sigma\, c_k$ and $\Sigma\, d_k$ simply represent the sum of the lengths of segments whose interiors are disjoint and whose unions are $\overline{B_jC_j}$, $\overline{B_{j+1}C_{j+1}}$, respectively. Hence, if the lengths of these segments are p_j, p_{j+1}, respectively, it follows from the properties of length functions that these sums (for fixed h_j) are $h_j\, p_j$ and $h_j\, p_{j+1}$ respectively. We thus obtain

$$\sum_k x_k x_k' = \sum_j h_j\, (p_j + p_{j+1}) .$$

A further application of Lemma 2 (applied to δ) shows that the sum on the right is xx' . Hence we have

$$xx' = \sum_k x_k x_k' .$$

In other words, the sum $\sum_k x_k x_k'$ has the same value (xx') for every triangulation of δ . It follows that $\sum_k \frac{1}{2} x_k x_k' = \frac{1}{2} xx'$, and that the sum on the left is the same for any triangulation. Hence the value of f_0 on a triangulated polygonal region is unchanged under "refinement" of the triangulation; and because any two triangulations have a common refinement, f_0 is a function of the polygonal region only, and is independent of the triangulation. This completes the proof of the theorem. Admittedly the details are fairly complicated, and it is difficult to devise a simple notation. But the underlying ideas are quite similar to those used for rectangular regions: the essential idea of the proof (as was the case for rectangular regions) is to reduce things to the point where we can use our knowledge of the additive properties of length functions on segments.

Let us now review what we have done. We have shown that if there exist any area functions for polygonal regions, these functions must have certain properties in relation to triangular regions and triangulations. We have used these properties to motivate the definition of a function for triangulated polygonal regions, and we have shown that this function does not depend on the particular triangulation. We thus obtain a function (for which we use the same symbol) $f_0 : P \to R^+$, where $f_0(p) = f_0(p, \{\delta_i\}) = \sum \frac{1}{2} x_i x_i'$, for any triangulation $\{\delta_i\}$. We wish to prove that f_0 is an area function for P . The fact that f_0 satisfies the congruence property AP-1 is immediate: under congruence all of the length values x_i , x_i' are unchanged. This leaves the additive property, AP-2 , still to be proved.

Theorem 3-5.8.

(i). The function $f_0 : P \to R^+$ defined above from a length function λ_0 , satisfies AP-2, and hence is an area function for polygonal regions.

(ii) If r is a rectangular region with sides of lengths a and b under the length function λ_0 , then $f_0(r) = ab$.

That is, the restriction $f_0|P_r$ of f_0 to rectangular regions, is the function $\eta(\lambda_0)$.

(iii) f_0 is the only area function for P which extends the area function $f_0|P_r$, and hence the 1-1 correspondence, η, of length functions and area functions for rectangular regions, extends to a 1-1 correspondence of length functions and area functions for polygonal regions.

(iv) If f is any area function for P, then so is the function kf, for each $k \in R^+$.

(v) f_0 is onto R^+.

(vi) If f_1, f_2 are any two area functions for P, then for some $k > 0$, $f_2 = kf_1$.

Proof.

(i) The proof of AP-2 is now quite simple. If p, p', are two polygonal regions with disjoint interiors, then two triangular regions δ_i, δ'_j in any triangulations $\{\delta_i\}$, $\{\delta'_j\}$, of p, p', respectively, also have disjoint interiors. Hence $\{\delta_i\} \cup \{\delta'_j\}$ is a triangulation of the polygonal region $p \cup p'$. Hence

$$f_0(p \cup p') = \sum_i f_0(\delta_i) + \sum_j f_0(\delta'_j)$$
$$= f_0(p) + f_0(p') .$$

(ii) This is immediate: the rectangular region r can be given a triangulation (by completing one diagonal) as the union of two congruent triangles δ_1, δ_2; and $f_0(\delta_1) = f_0(\delta_2) = \frac{1}{2}ab$; hence, from (i), $f_0(r) = f_0(\delta_1) + f_0(\delta_2) = ab$. Hence $f_0|P_r$ is the unique area function which corresponds to λ_0 under the particular 1-1 correspondence η described earlier. (I.e., $f_0|P_r$ is the area function for which the square with side "1" with respect to λ_0, has area "1".)

(iii) If f_0' is any other area function which extends $f_0|P_r$, then

$$f_0'|P_r = f_0|P_r .$$

It follows from Theorem 3-5.6 that f_0, f_0' agree on all triangular regions. Hence for any polygonal region p, with any triangulation $\{\delta_i\}$, $f_0(p) = \sum_i f_0(\delta_i) = \sum_i f_0'(\delta_i)$ $= f_0'(p)$, so that $f_0 = f_0'$. In other words each area function for rectangular regions can be extended to polygonal regions in exactly one way. Moreover, as we saw earlier, every area function for polygonal regions is an extension of an area function for rectangular regions, hence the 1-1 correspondence η of length functions for segments and area functions for rectangular regions, carries over to a 1-1 correspondence (for which we use the same symbol, η) of length functions and area functions for polygonal regions.

(iv) This property, which is quite similar to the corresponding property for area functions for rectangular regions, is easily verified.

(v) $f_0|P_r$ is already onto R^+.

(vi) This result can be established directly, as for the corresponding result for length functions, or it can be derived (see below) from the corresponding result for rectangular regions.

Remarks:

1. The importance of (iii) above is that we can use the names of length functions (inch function, foot function, etc.) to name (unambiguously) the area functions for polygonal regions, giving them the same names ("square inch function", "square foot function", etc.) that we used in the discussion of area for rectangular regions.

2. The results (iv) - (vi) show that all area functions for polygonal regions are similar; (i.e., the set Δ of such functions is a ratio scale). This situation is quite similar to that for length functions, and it is a simple matter to show that $(\Delta, +)$ is an R^+-semimodule. This situation will continue to hold when the domain is further extended.

Change of Scale for Area Functions on Polygonal Regions. If λ_1 and $\lambda_2 = k\lambda_1$ are two length functions, then we have seen that the corresponding area functions $f_1 = \eta(\lambda_1)$, $f_2 = \eta(\lambda_2)$, for polygonal regions, have the property that, for any rectangular region r, $f_2(r) = k^2 f_1(r)$. For any triangular region δ, with base and altitude x_1, x'_1, under λ_1, and x_2, x'_2, under λ_2, we have

$$\begin{aligned} f_2(\delta) = \frac{1}{2} x_2 x'_2 &= \frac{1}{2}(kx_1)(kx'_1) \\ &= k^2 \frac{1}{2} x_1 x'_1 \\ &= k^2 f_1(\delta) . \end{aligned}$$

Hence for any polygonal region p, and any triangulation $\{\delta_i\}$,

$$\begin{aligned} f_2(p) = \sum_i f_2(\delta_i) &= \Sigma\, k^2 f_1(\delta_i) \\ &= k^2\, \Sigma\, f_1(\delta_i) \\ &= k^2 f_1(p) . \end{aligned}$$

Hence we have proved the following

Theorem 3-5.9. If f_1, f_2, are the area functions for P which correspond as above to the length functions λ_1, λ_2 $(=k\lambda_1)$ respectively, then $f_2 = k^2 f_1$.

Remarks:

1. The preservation of this relationship in the extension of area functions from rectangular to polygonal regions is very important. It means that, in considering the areas of polygonal regions, we can continue to use the familiar "formulas" for changing area units, and for relating area functions. This property will continue to hold when we further extend the domain of area functions.

2. If p_1 and p_2 are polygonal regions, with $f_1(p_1) = f_1(p_2)$ for some area function f_1, then $f_2(p_1) = f_2(p_2)$ for any other area function f_2. When this situation holds, we say that p_1 and p_2 have the

same area. It is trivial to prove that "same area" is an equivalence relation on the set P of polygonal regions.

Monotonicity of Area Functions on Polygonal Regions. We say that p is a sub-region of the polygonal region p' , if there is a triangulation $\{\delta_i\}$ of p , and a triangulation $\{\delta'_j\}$ of p' , such that every δ_i is a δ'_j. If f is any area function, we then have

$$f(p) = \sum_i f(\delta_i) \leq \sum_j f(\delta'_j) = f(p')$$

and the inequality is strict if p is a proper sub-region. (I.e., if there is some δ'_j which is not a δ_i.) Clearly this monotonicity property holds for all area functions, so we can make the unit-free statement "the area of p is less than or equal to the area of p' ".

Piecewise Congruence Under Decomposition: Bolyai's Theorem. We introduce a relation ($\sim$) into the set of all polygonal regions, by defining $p_1 \sim p_2$ if there are polygonal decompositions of p_1 and p_2 which correspond 1-1, with corresponding regions congruent. Hence, in particular, if $p_1 \cong p_2$ then $p_1 \sim p_2$. It can be shown that congruent regions have congruent triangulations (we cannot go into detail, as we have not developed the general concept of congruence) hence $p_1 \sim p_2$ if and only if p_1 and p_2 have piecewise congruent triangulations; i.e., triangulations such that there is a 1-1 correspondence of triangular regions, with corresponding regions congruent.

It follows from the properties which we have developed for area functions, that if $p_1 \sim p_2$, then $f(p_1) = f(p_2)$ for every area function f . The converse result, which we give below, was proved about 100 years ago by Bolyai.

Theorem 3-5.10. If p_1 and p_2 are polygonal regions with the same area, then $p_1 \sim p_2$.

We break down the proof with two lemmas:

Lemma 1. The relation $\sim$ is an equivalence relation.

Proof. The proof of symmetry and reflexivity is immediate. To prove transitivity we note that if $p_1 \sim p_2$ and $p_2 \sim p_3$, then p_2 has triangulations $\{\delta'_i\}$, $\{\delta''_j\}$, which are piecewise congruent to triangulations of p_1 and p_3 respectively. We know that there is a common refinement $\{\delta_k\}$, of $\{\delta'_i\}$ and $\{\delta''_j\}$, and this refinement can be "copied" using the original piecewise congruences, to give piecewise congruent triangulations of p_1, p_2, and p_3. (We are using here the fact that if two triangular regions are congruent, and one has a triangulation, then the other has a piecewise congruent triangulation.) Hence $p_1 \sim p_3$.

For the remainder of the proof of this theorem we use the word "equivalent" to replace "piecewise congruent under decomposition". If σ_0 is any fixed segment, we define a normal rectangular region (relative to σ_0) to be a rectangular region with one side congruent to σ_0.

Lemma 2.

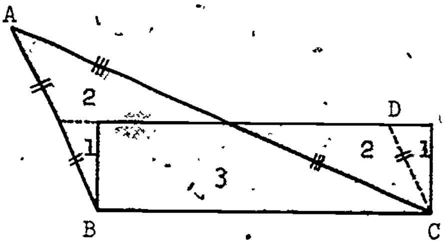
(i) Every triangular region is equivalent to a rectangular region.

(ii) Triangular regions with congruent bases and congruent corresponding altitudes, are equivalent.

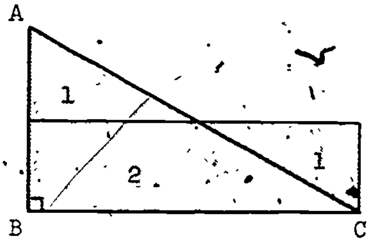
(iii) Every rectangular region (and hence, from (i), every triangular region) is equivalent to a normal rectangular region.

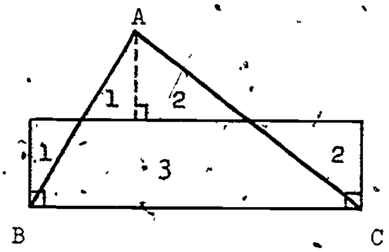
Proof.

(i) We show first that every triangular region, with any side as base, is equivalent to a rectangular region with the same base and half the altitude. The figure shows the various cases that can arise, and the proof can be completed by showing that the different regions which are similarly numbered, are congruent.



[$\overleftrightarrow{CD}$ is parallel to $\overleftrightarrow{AB}$]

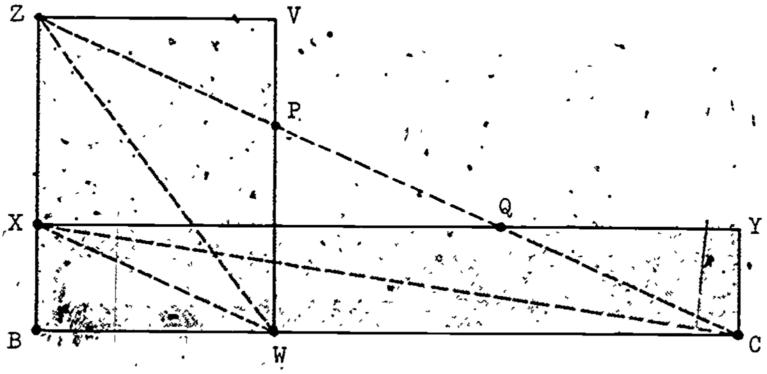




(ii) It follows, by the transitivity of equivalence, that triangular regions with congruent bases and congruent corresponding altitudes are equivalent to congruent rectangular regions, and hence to each other.

(iii) We show next that any rectangular region is equivalent to a normal rectangular region:

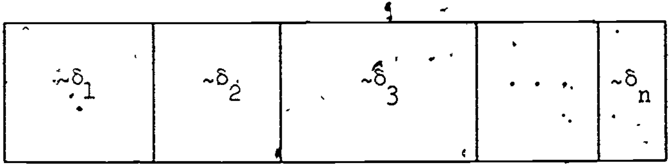
Let BXYC be the given rectangular region, and let $\overline{BZ}$ be congruent to σ_0, with X, Z, on the same side of $\overleftrightarrow{BC}$ as shown. We complete the proof



for the case X between Z and B. (If Z is between X and B the proof is similar; if Z = X, there is nothing further to prove.)

Let $\overleftrightarrow{XW}$ be parallel to $\overleftrightarrow{CZ}$, meeting $\overleftrightarrow{BC}$ at W, and complete the normal rectangle $BZVW$. To show that the rectangular regions $BXYC$, $BZVW$ are equivalent, it is clearly sufficient to show that the "half-regions", ▲BZW and ▲BXC, are equivalent. The latter equivalence follows from the facts that ▲BXW is common, and that ▲XWZ and ▲XWC have the same base and congruent altitudes. It follows from the transitivity of equivalence, that every triangular region is equivalent to a normal rectangular region.

Proof of Theorem 3-5.10. Since every triangular region is equivalent to a normal rectangular region, a polygonal region p, with a triangulation $\{\delta_i\}$ $(i=1, 2, \ldots, n)$ is equivalent to a normal rectangular region obtained by "joining" side-by-side, normal rectangular regions which are equivalent to the separate δ_i. The diagram illustrates the idea: this "join" property is one of the reasons for using rectangular regions with one side congruent to the fixed segment σ_0.



We have now shown that each polygonal region is equivalent to a normal rectangular region. We assert that this normal rectangular region is uniquely determined, up to congruence:

Let λ_0 be the length function which corresponds to σ_0 as unit, and let $f_0(=\eta(\lambda_0))$ be the area function which corresponds to λ_0. Let p_1, p_2, be polygonal regions which have the same area, and let p_1, p_2, be equivalent to normal rectangles n_1, n_2, respectively. Let the lengths of the sides of n_1, n_2, (under λ_0) be 1, x_1, and 1, x_2, respectively. Now we have $f_0(p_1) = f_0(p_2)$. Hence,

$$x_1 = f_0(n_1) = f_0(p_1) = f_0(p_2) = f_0(n_2) = x_2 .$$

Hence the rectangular regions n_1, n_2, are congruent, and therefore they are equivalent. It now follows, from the transitivity of equivalence, that $p_1 \sim n_1 \sim n_2 \sim p_2$, which completes the proof of Bolyai's Theorem.

Remarks:

1. If you review the proof of the existence of area functions for polygonal regions, you will probably agree that the most awkward part was Theorem 3-5.7, in which we proved that, for a triangulation $\{\delta_i\}$ of a triangular region δ, $\Sigma x_i x_i'$ was independent of the triangulation. The development used in Bolyai's Theorem might suggest to you that we could by-pass this proof by showing that AP-1 and AP-2 imply that equivalent regions must have the same "area", showing that every polygonal region is equivalent to a normal rectangular region, and then using the area theory for rectangular regions to define area functions for P. This can be done, but, in order to get area functions for P, we must show that equivalent polygonal regions are equivalent to a unique normal rectangular region (in effect, that equivalent normal rectangular regions are congruent) and the proof of this is, essentially, Theorem 3-5.7. We cannot use Bolyai's Theorem to prove that equivalent normal rectangular regions are congruent: Bolyai's Theorem requires the existence of area functions for polygonal regions, and hence this theorem cannot be used to establish the existence.

2. We can use normal rectangular regions to reinforce our earlier comment concerning the arbitrariness in the naming of area functions by using the length functions to which they correspond under the 1-1 correspondence η. Let σ_0 be a fixed segment, as in the proof of Bolyai's Theorem, and let λ be any length function. For a polygonal region p, we have seen that there is a unique (up to congruence) normal rectangular region n, such that $n \sim p$. One side of n is congruent to σ_0; let σ be the other side of n. Then we may define a function

$$g_\lambda : P \rightarrow R^+$$

by

$$g_\lambda(p) = \lambda(\sigma) .$$

It is easy to prove directly that g_λ satisfies AP-1 and AP-2, so that g_λ is an area function for P. Thus there is an area function $f : P \rightarrow R^+$, with $f = g_\lambda$. If λ is the inch function, it might be reasonable to call g_λ the "σ_0-inch" function.

Clearly each length function λ determines a unique area function g_λ, and (for fixed σ_0) we have a 1-1 correspondence $\lambda \leftrightarrow g_\lambda$ of length functions, λ, and area functions, g_λ. Moreover if f_0 is any area function, then f_0 corresponds under η to a length function λ_0, such that

$$f_0(p) = \lambda_0(\sigma_0) \cdot \lambda_0(\sigma) .$$

Hence if λ is the length function $\lambda = \lambda_0(\sigma_0) \cdot \lambda_0$ we have $g_\lambda(p) = \lambda(\sigma)$ $= \lambda_0(\sigma_0) \cdot \lambda_0(\sigma) = f_0(p)$. That is, the set $\{g_\lambda\}$ of functions (derived from the set Λ of all length functions) is the set of all area functions. In other words, the correspondence $\lambda \leftrightarrow g_\lambda$ is a 1-1 correspondence of length and area functions. This correspondence depends, of course, on σ_0, and it is uniquely determined by the congruence class of σ_0. Thus we get one such correspondence for each congruence class of segments. For a fixed segment σ, it will be convenient to denote the corresponding 1-1 correspondence by

$$\eta_\sigma : \Lambda \rightarrow \Delta ,$$

where Δ denotes the set of all area functions for polygonal regions.

We have introduced this alternate procedure for the "naming" of area functions in terms of related length functions, in order to show you the degree of arbitrariness in this "naming", and the fact that our usual procedure is not forced on us in any sense: the usual procedure just happens to be by far the most convenient. The fact that the usual procedure involves a convention is important in the understanding of the question of "dimension" in relation to measure functions. We shall return to this question again during the discussion of "dimension", but meanwhile you should note that, for area functions "named" by the correspondences η_σ, the usual change of units/change of scale "formulas" do not apply. For example, if for any fixed σ, $f_{\sigma\text{-in}}$ denotes the σ-inch area function, and $f_{\sigma\text{-ft}}$ denotes the σ-foot function, you should verify that $f_{\sigma\text{-in}} = 12 f_{\sigma\text{-ft}}$. In the everyday language of "units", this would be expressed as: 12 σ-inches = 1 σ-foot.

N. B. To avoid confusion, we adopt the usual convention that whenever we refer without qualification, to "the length function which corresponds to a given area function", or "the area function which is determined by a given length function", we are referring to the correspondence η. When we want to refer to any other correspondence (such as one of the η_σ correspondences) we shall do so explicitly. In general, if λ_1 and λ_2 $(= k\lambda_1)$ are two length functions, and if f_1, f_2 are the corresponding area functions under η_σ ; (i.e., $f_1 = \eta_\sigma(\lambda_1)$, $f_2 = \eta_\sigma(\lambda_2)$) then $f_2 = kf_1$. The following commutative diagrams indicate the relationship between the automorphisms $\bar{\bar{k}}$ of the sets Λ (length functions) and $\triangle$ (area functions) which are determined by composition with the similarity transformations $\bar{\bar{k}}$ of R^+, and the correspondences η and η_σ.

$$\begin{tikzcd}
\Lambda \arrow[r, leftrightarrow, "\bar{\bar{k}}"] \arrow[d, leftrightarrow, "\eta"'] & \Lambda \arrow[d, leftrightarrow, "\eta"] \\
\triangle \arrow[r, leftrightarrow, "\bar{\bar{k}}^2"] & \triangle
\end{tikzcd}
\qquad
\begin{tikzcd}
\Lambda \arrow[r, leftrightarrow, "\bar{\bar{k}}"] \arrow[d, leftrightarrow, "\eta_\sigma"'] & \Lambda \arrow[d, leftrightarrow, "\eta_\sigma"] \\
\triangle \arrow[r, leftrightarrow, "\bar{\bar{k}}"] & \triangle
\end{tikzcd}$$

If you like to see this sort of relationship expressed by a "formula" then the above diagrams yield: $\eta\bar{\bar{k}} = \bar{\bar{k}}^2\eta$; $\eta_\sigma\bar{\bar{k}} = \bar{\bar{k}}\eta_\sigma$; where $\bar{\bar{k}}$, $\bar{\bar{k}}^2$, are the functions indicated by the diagrams.

In each of the sets of functions Λ, $\triangle$, in addition to the scalar multiplication by positive constants k (which determines the automorphisms $\bar{\bar{k}}$) we have an operation of addition, and the sets are closed under this operation. It is natural to ask how the functions η, η_σ, behave with respect to this addition. This question, which is only of secondary interest, is answered in the following exercises.

Exercises 3-5

1. If r is a rectangular region with adjacent edges σ', σ'', and if λ_1, λ_2, are two length functions, prove that

   (i) $\lambda_1(\sigma') \cdot \lambda_2(\sigma'') = \lambda_1(\sigma'') \cdot \lambda_2(\sigma')$;

   (ii) the function $f_{12} : r \to \lambda_1(\sigma') \cdot \lambda_2(\sigma'')$ is an area function for the set P_r of rectangular regions;

   (iii) the function $f_{11} : r \to \lambda_1(\sigma') \cdot \lambda_1(\sigma'')$ is the area function $\eta(\lambda_1)$, which corresponds (under η) to λ_1;

   (iv) $\sqrt{\lambda_1\lambda_2}$ is a length function, and $\eta(\sqrt{\lambda_1\lambda_2}) = f_{12}$; (cf. Exercise 2-4.16);

   (v) $\eta(\lambda_1 + \lambda_2) = f_{11} + 2f_{12} + f_{22}$

   $$= \eta(\lambda_1) + 2\eta(\sqrt{\lambda_1\lambda_2}) + \eta(\lambda_2);$$

   (vi) $\eta_\sigma(\lambda_1 + \lambda_2) = \eta_\sigma(\lambda_1) + \eta_\sigma(\lambda_2)$, and η_σ is an isomorphism — from $(\Lambda, +)$ to $(\Delta, +)$.

2. (i) If f_1, f_2, are area functions for rectangular regions, with "units" s_1, s_2, respectively, (the regions s_1, s_2 are assumed disjoint), show that the area function for which $s_1 \cup s_2$ is the "unit", is $\dfrac{f_1 f_2}{f_1 + f_2}$;

   (ii) If f_1, f_2, . . . , f_n are area functions with "units" s_1, s_2, ... , s_n, then

   (a) the f_i are in arithmetic progression if and only if the s_i are in harmonic progression;

   (b) the f_i are in harmonic progression if and only if the s_i are in arithmetic progression.

(Refer to Exercises 2-4, Nos. 8, 9, 10, 11, and the remark which follows: the definitions of arithmetic and harmonic progression for area units and functions are similar to those for length functions, and results such as these hold for any class of similarity-related measure functions--i.e., for any ratio scale--whose functions all have units.)

3. Let A_Λ and A_Δ denote the groups (under composition) of automorphisms of Λ and Δ. Prove that

$$A_\Lambda = \{\bar{\bar{k}} : \bar{\bar{k}}(\lambda) = k\lambda \ , \ k \in R^+\}$$

and similarly for A_Δ; and that $\overline{\overline{k_1 k_2}} = \bar{\bar{k}}_1 \bar{\bar{k}}_2$. Hence show that the correspondence $k \to \bar{\bar{k}}$ is an isomorphism of $(R^+, \cdot)$ and each of the automorphism groups $(A_\Lambda , \circ)$, $(A_\Delta , \circ)$, where $\circ$ denotes composition of automorphisms. (Strictly speaking, we should not use $\bar{\bar{k}}$ to denote an element of A_Λ and also to denote an element of A_Δ, but this minor ambiguity is easily resolved from the context.)

4. (i) Define $\eta^* : A_\Lambda \to A_\Delta$ by $\eta^*(\bar{\bar{k}}) = \bar{\bar{k}}_1$, where $\bar{\bar{k}}_1(\eta(\lambda)) = \eta(\bar{\bar{k}}(\lambda))$. Show that $\eta^* : \bar{\bar{k}} \to \overline{\overline{k^2}}$, and that η^* is an isomorphism of A_Λ onto A_Δ;

(ii) Similarly define $\eta^*_\sigma : A_\Lambda \to A_\Delta$, show that $\eta^*_\sigma : \bar{\bar{k}} \to \bar{\bar{k}}$, and that η^*_σ is an isomorphism of A_Λ onto A_Δ.

Remark on Exercise 4: We shall refer to these results in the next chapter, but meanwhile you should observe that η^* (and, similarly, η^*_σ) is defined so as to make the following diagram commutative:

$$\begin{tikzcd}
\Lambda \arrow[r, leftrightarrow, "\bar{\bar{k}}"] \arrow[d, leftrightarrow, "\eta"'] & \Lambda \arrow[d, leftrightarrow, "\eta"] \\
\Delta \arrow[r, leftrightarrow, "\eta^*(\bar{\bar{k}})"'] & \Delta
\end{tikzcd}$$

That is, if $f \in \Delta$,

$$(\eta^*(\bar{\bar{k}}))(f) = \eta(\bar{\bar{k}}(\eta^{-1}(f)))$$

so that $\eta^*(\bar{\bar{k}}) = \eta\bar{\bar{k}}\eta^{-1}$.

Area and Language. We conclude the elementary (or geometric) study of area functions, by pointing out the relationship between our treatment, and some of the ways in which questions of area are described in everyday language. This situation for area is very much as it was for length.

First of all, we remind you that we do not have a definition of the separate word "area". We have defined "area function", and we can easily define exactly what we mean by such expressions as "the area of A is larger than the area of B "; or "X and Y have the same area". The latter expression, for example, is equivalent to saying that X and Y belong to the domain of some area function and the value of the function on X is the same as its value on Y . You should observe that this, and inequality statements concerning "area", do not depend on a particular choice of area function. (I.e., they are "unit-free" statements.)

When we have fixed on a particular area function, and it is clearly understood that we are referring to values with respect to this function, it is quite satisfactory to use an expression such as "the area of A is 7 ". In many situations in mathematics we wish to work with a fixed area function. For example, in calculus we almost always work with the area function which corresponds (under η) to the length function which is compatible with an assumed coordinate structure. Neither the length function nor the corresponding area function are normally named. Sometimes an attempt is made to indicate the (usually unstated) conventions by referring to length and area values as so many "units", and it is understood that the "units" for length and area "correspond", under the correspondence which is "dual" to η . That is, the unit of area is (represented by) the square region whose side represents the unit of length; and a statement such as "the area is 7 units" can be interpreted in terms of the scalar multiplication of domain classes by positive real numbers. As in the case of "length", such a multiplication can be defined for all positive real numbers whenever we have a ratio scale whose functions are onto R^+ ; and for suitable real numbers when the functions are not onto.

In some books you will find a distinction made between "the area of A " and "the measure of the area of A with respect to a particular area function". When this is done, the expression "the area of A " is really a statement about the domain of (all) area functions, and it can be interpreted to mean the set of all domain elements with the same "area" (i.e., same value under every area function) as A. That is, the "area of A " is the equivalence class to which A belongs under the equivalence relation which corresponds to "same area". The measure of the area of A under a particular area function f is then the number f(A). This distinction is usually dropped in everyday language, when we use such expressions as "the area of A is 7 square inches".

In Exercise 3-5.1 above, and in other places, we have deliberately adopted a notation which differs from the everyday convention. In everyday usage, if "foot" is a length unit, and we adopt the same name for the corresponding length function, then "$(ft)^2$" is used (with or without parentheses) for the area function which we have called $\eta(ft)$. (It is also used to denote the unit of that function.) If we were to adopt this notational convention, then for each length function λ, we should use λ^2 to denote the corresponding area function $\eta(\lambda)$. Similarly the area function which we called f_{12} (Exercise 3-5.1 (ii)) would be denoted by $\lambda_1\lambda_2$.

Let us see what some of the results proved (or given as exercises) above would look like in this notation:

(i) The result $\eta(k\lambda) = k^2\lambda$ becomes

$$(k\lambda)^2 = k^2\lambda^2 .$$

(ii) The result (Exercise 3-5.1 (iv)) $\eta(\sqrt{\lambda_1\lambda_2}) = f_{12}$, becomes

$$(\sqrt{\lambda_1\lambda_2})^2 = \lambda_1\lambda_2 .$$

(iii) The result (Exercise 3-5.1 (v)) $\eta(\lambda_1 + \lambda_2) = f_{11} + 2f_{12} + f_{22}$ becomes

$$(\lambda_1 + \lambda_2)^2 = \lambda_1^2 + 2\lambda_1\lambda_2 + \lambda_2^2 .$$

These results are part of the justification for the way in which scientists manipulate the symbols which name functions (or units) as if these symbols were themselves amenable to the laws of algebra. We have avoided using these highly suggestive notations, because it is easier to see what is involved in, and to prove, the relations expressed in (i) - (iii) above, if we avoid a too-suggestive notation.

Another reason by avoiding the notation "λ^2" for the area function $\eta(\lambda)$, is that λ^2 has a different and well-established meaning in the algebra of real-valued functions: λ^2 denotes the function whose domain is the same as that of λ , and whose values are the squares of corresponding values of λ . (E.g., consider the functions $f : x \to \sin x$, and $f^2 : x \to \sin^2 x$.)

A similar situation is reflected in the usual way in which one keeps track of functions and units in the everyday application of measurement "formulas". For example, if r is a rectangle with sides of length 3 inches and 4 inches, we often find the area calculation written as: area of $r = 3 \text{ in.} \times 4 \text{ in.} = 12 \text{ in.}^2$. The validity of this rests, of course, on the basic theorem concerning area for rectangles: if σ_1 ; σ_2 are adjacent sides of r , then we have $\lambda_{in}(\sigma_1) = 3$, $\lambda_{in}(\sigma_2) = 4$, and $f_{in^2}(r) = 3 \cdot 4 = 12$, where $f_{in^2} = \eta(\lambda_{in})$. If we drop the λ and the f , we have

$$\text{in.}(\sigma_1) = 3 , \ \text{in.}(\sigma_2) = 4 ,$$

$$\text{in.}^2(r) = [\text{in.}(\sigma_1)][\text{in.}(\sigma_2)] = 12 .$$

This is usually written as: length of $\sigma_1 = 3$ in. , length of $\sigma_2 = 4$ in. , area of $r = 3 \text{ in} \times 4 \text{ in} = 12 \text{ in}^2$; which often gets abbreviated to

$$L\sigma_1 = 3 \text{ in} , \ L\sigma_2 = 4 \text{ in} ; \ Ar = 12 \text{ in}^2$$

or even

$$\sigma_1 = 3 \text{ in} , \ \sigma_2 = 4 \text{ in} , \ r = 12 \text{ in}^2 .$$

From a mathematical point of view, no significance is yet attached to the symbol "in^2" except as the name of the particular area function which corresponds to the length function "in" under the defined correspondence η ; and the justification for the common abbreviations and symbol manipulations is that, when interpreted in this way, they correspond to valid results. In other words, we may regard them as a kind of shorthand notation for keeping track of functions (or units) and of the relationships between functions. [Later on we shall be able to give another interpretation, in terms of certain operations by means of which new ratio scales may be constructed as "powers" and "tensor products" of existing scales.]

For mnemonic purposes, we have such formulas as

$$A = \ell b. \ ; \ A = \frac{1}{2} b h .$$

for the "areas" of rectangular and triangular regions. In these formulas no particular area of length functions are usually mentioned, but it is understood that ℓ, b, h, A, are numbers derived from any length function λ, and the related area function $\eta(\lambda)$.

So far we have only defined area functions for polygonal regions. We shall need to extend the notion of area to a considerably larger domain. This is discussed in Section 3-6. It will turn out that the extended functions will be in 1-1 correspondence with those we have already defined, so we can use the same conventions in naming them. Moreover the relationships (discussed above) of area functions to one another, and between length and area functions, will still hold, as will the validity of the common language and symbol conventions.

3-6. Extension of The Domain For Area Functions

In Sections 2-8 and 2-9 we discussed in some detail the way in which the domain of length functions could be extended beyond the elementary domain of segments. The theory of area is capable of a similar treatment, to give a theory of area for (some) plane sets which are not polygonal regions, and for (some) nonplane sets; (e.g., subsets of so-called "curved surfaces"). Having given the flavor of this kind of extension problem for length functions, we do not intend to carry out the corresponding treatment for area in anything like the same detail.

Our treatment of area for polygonal regions does not even apply to such "simple" regions as circular regions (i.e., the union of a circle and its interior) or most of the regions "under the graphs of continuous functions", which we encounter in elementary integral calculus. We therefore wish to extend the domain of area functions to a larger class of subsets of the plane; and, in the extension, we would like to preserve at least some of the more important properties of area functions as defined for polygonal regions. As a result of our experience with linear measure, we do not approach the area extension problem with quite such a naive outlook. In particular, we are quite prepared to accept that some nonempty sets are likely to have zero "area measure", and that some sets are likely to be, in some sense, "unmeasurable".

If A, B, C, . . . denote subsets of the plane, properties of a generalized area function μ (on a domain D of plane sets) which we might regard as essential, are

A1. $\mu(A) \geq 0$ for every set A in the domain.

A2. $\mu(\emptyset) = 0$.

A3. (Congruence Property). If $A \cong B$, then $\mu(A) = \mu(B)$.

A4. (Additivity). If $A \cap B = \emptyset$, then $\mu(A \cup B) = \mu(A) + \mu(B)$.

A5. (Monotonicity). If $A \subset B$ then $\mu(A) \leq \mu(B)$.

A6. D contains P (the set of all polygonal regions); and, if A is a polygonal region, then $\mu(A) = f(A)$, for some area function f which depends only on μ, and not on A . (I.e., every measure function μ is an extension of some area function f for polygonal regions.)

A7. The correspondence $\mu \longleftrightarrow f$ (of A6.) is 1-1.

A8. If μ is a suitable measure function, then so is $k\mu$; and if $\mu \longleftrightarrow f$, then $k\mu \longleftrightarrow kf$.

We might go further than this and demand that similar sets (in the geometric sense) should have suitably-related measures; that the domain D should have a suitable structure (e.g., a Boolean ring) etc. The question we must then answer, is whether or not there are significant domains D on which such measure functions exist. There certainly are: if we trivially extenl the domain of polygonal regions to include the empty set (with measure zero) then all of the properties A1 - A8 are satisfied. Moreover, with the definition which we have adopted for area functions on the domain P of polygonal regions, such an area function is finitely additive for unions of regions whose boundaries may have a non-null intersection. This suggests immediately that we might be able to include such "degenerate" regions as points and line segments (and finite unions of such regions) in our domain, and give these regions measure zero. The next step would be to include such sets as the interiors of triangles, and finite "disjoint unions" of such sets. But this sort of haphazard approach does not lead easily to a satisfactory general theory. To get such a theory, we need to take a more general approach, such as that which we used for the definition of Lebesgue measure for subsets of the line. In adopting such an approach, we are, however, doing little more than reporting the end result of a process which undoubtedly involved a considerable amount of "trial and error" in its developmental stages.

It is possible to develop Lebesgue measure theory for subsets of the plane, for subsets of space, and, more generally, for subsets of real euclidean spaces of any finite dimension; and this can be done with very little more work than that involved in the development of Lebesgue measure theory for subsets of the line, if appropriate definitions are adopted for such terms as "open set", and "interval" ~~in the euclidean space~~ of the relevant dimension. But it is more instructive to deal first with the simpler (and less general) theory of Jordan measure (first developed in 1892) which was extended by Borel and Lebesgue about ten years later. The Jordan theory gives us a domain which includes circular regions, and the regions "under" the graphs of continuous functions, and it is therefore a satisfactory theory for the development of the (Riemann) definite integral in elementary calculus.

The Jordan Measure of Plane Sets. The idea involved in Jordan measure (which is sometimes called Jordan content) is to approximate a given subset X of the plane by polygonal regions in two ways: we take inner (or lower) approximations by polygonal regions p_* which are contained in X; and we take outer (or upper) approximations by polygonal regions p^* which contain X. For a fixed area function f for P, we define $\mu_*(X) = \sup\{f(p_*)\}$, (provided that this least upper bound exists) and we define $\mu^*(X) = \inf\{f(p^*)\}$. If there is no polygonal region contained in X, then we define $\mu_*(X) = 0$. (Actually we can consider values in an "extended" number system, so that the least upper bound will always exist.) The number $\mu_*(X)$ is called the *interior Jordan measure* of X; and $\mu^*(X)$ is the *exterior Jordan measure* of X. If $\mu_*(X) = \mu^*(X)$, we say that X is measurable in the sense of Jordan, or *Jordan measurable*. In this case the common value $\mu(X) = \mu_*(X) = \mu^*(X)$ is called the *Jordan measure* (or content) of X. We point out several things:

(i) The use of the definite article in the expression "*the* Jordan measure" is standard usage. But this function μ depends, of course, on f, and μ is the Jordan measure which corresponds to f.

(ii) This idea of inner and outer approximations by polygonal regions goes back to the Greeks, and it was used by them to determine (by plausible arguments) the areas of circular regions and other simple nonpolygonal regions.

(iii) The idea that a generalized area, or measure, (if it ~~exists~~) should lie between each inner and each outer approximation is, of course, very natural. From a formal point of view, it is a consequence of the monotonocity property and of the desire that the generalized measure function μ should be the same as the area function f on the domain of polygonal regions, that, if $p_* \subset X \subset p^*$, then $f(p_*) = \mu(p_*) \leq \mu(X) \leq \mu(p^*) = f(p^*)$. Thus if X is to have such a generalized area measure, and if $\sup\{f(p_*)\} = \inf\{f(p^*)\}$, then this common value is the only possible value for $\mu(X)$. If $\sup\{f(p_*)\} \neq \inf\{f(p^*)\}$, the question of whether or not it is possible to assign (in some "reasonable" way) a generalized area measure to X, is still left open.

(iv) The definition of Jordan measure need not involve the use of polygonal regions; it can be given in terms of simpler geometric regions, such as rectangular regions, or triangular regions. If either of these approaches is used, then we use finite unions in the formation of outer approximations, and finite unions of pairwise disjoint regions in the formation of inner approximations. The resulting theory will be exactly the same under each of these approaches. I.e., the set of measurable sets will be the same, and the measure function determined by a given f' (an area function for rectangular or triangular regions) will be the same as that obtained by first extending the area function to polygonal regions, and then building a Jordan measure function on this extended area function. Whatever the approach, the Jordan measure and the extended area functions agree on the set of all polygonal regions. (We do not intend to prove these statements, but we mention them because you will find Jordan measure treated in different (but equivalent) ways, in different books on real function theory and integration; and because the approaches using triangular or rectangular regions are closer to the further generalizations of Borel and Lebesgue.)

Properties of Jordan Measure. Let μ be the Jordan measure function which is determined by the area function f for polygonal regions. Then the following properties may be proved:

(a) Every polygonal region p is Jordan measurable, and $\mu(p) = f(p)$.

(b) (Monotonicity). If $p_1 \subset p_2$ then $\mu(p_1) \leq \mu(p_2)$.

(c) If X is the empty set, or a single point, or a segment, or a finite union of points and segments, then X is Jordan-measurable, and $\mu(X) = 0$.

(d) (Congruence Property). If X_1 and X_2 belong to D (the class of all Jordan measurable plane sets) and if $X_1 \cong X_2$, then $\mu(X_1) = \mu(X_2)$.

(e) (Similarity Property). If X_1 and X_2 are geometrically similar, Jordan measurable, sets, with a "similarity factor" $k > 0$. (i.e., all distances in X_1 are k times the corresponding distances in X_2) then $\mu(X_1) = k^2\mu(X_2)$.

(f) (Finite Additivity). If $X_i \in D$, $(i=1, 2, \ldots, n)$ and $X_i \cap X_j = \emptyset$ when $i \neq j$, then $\bigcup_{i=1}^{n} X_i \in D$, and $\mu(\bigcup_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mu(X_i)$.

(g) D is a Boolean ring; i.e.,

(i) if $X_i \in D$ $(i=1, 2, \ldots, n)$ then $\bigcup_{i=1}^{n} X_i \in D$; and

(ii) if $X_1, X_2 \in D$, then $X_1 - X_2 \in D$ and $X_1 \cap X_2 \in D$.

(h) D does not depend on the particular area function f (for P) from which μ was derived.

(i) If $X_1, X_2 \in D$ and $\mu(X_1 \cap X_2) = 0$, then $\mu(X_1 \cup X_2) = \mu(X_1) + \mu(X_2)$.

(j) If f_1, and $f_2 = kf_1$, are area functions for P, and if μ_1, μ_2, are the corresponding Jordan measure functions for D, then $\mu_2 = k\mu_1$.

(k) D includes all circular regions and all circular sectors; the Jordan measure of a circular region of radius r (under a length function λ), is πr^2 (under the measure function μ which is derived from the area function $\eta(\lambda)$).

(l) Let $h(x)$ and $g(x)$ be two continuous real valued functions defined on an interval $[a, b]$, $(a < b)$, with $h(x) \leq g(x)$ for all $x \in [a, b]$. Let X be the subset of the cartesian plane defined by $X = \{(x, y) : a \leq x \leq b ; h(x) \leq y \leq g(x)\}$. Then X is Jordan measurable, and,

$$\mu(X) = \int_a^b [g(x) - h(x)]dx ,$$

where μ is determined by the area function $f = \eta(\lambda)$, and λ is the length function which is compatible with the coordinate function for the given plane. The integral on the right is the so-called definite integral of Riemann, the standard definite integral of elementary calculus.

The proofs of most of these properties are quite straightforward. Many of them can be found (along with a much more detailed treatment than we have given) in Chapters 21, 22, of [14]. For others see [19].

Sets Which Are Not Jordan Measurable: Lebesgue Measure. The theory of Jordan measure which we outlined above for plane sets, has exact analogues for subsets of the line, and for subsets of higher dimensional euclidean spaces. In our treatment of linear measure we did not discuss the Jordan theory, because it does not take us very far beyond sets which are merely finite unions of segments. [In case you wish to look at the theory of Jordan measure on the line, the outer measure of a set is defined as for area, in terms of the greatest lower bound of the sum of the lengths of those finite unions of intervals which contain the set, and the inner measure is defined in terms of the least upper bound of the sum of the lengths of those finite unions of pairwise disjoint intervals which are contained in the set.]

Jordan measure has serious technical deficiencies, which result, essentially, from the restriction to finite unions in its definition. For example, you can easily show that the set S_0 of rational points on the interval $[0,1]$ is not measurable in the sense of Jordan linear measure. [It is not hard to see that S_0 contains no interval of real numbers, so that its interior measure is zero; and that every finite union of intervals which contains S_0 also contains the whole interval $[0,1]$, so that the exterior measure of S_0 is 1.] Similarly, the set of points (x,y) in the unit square $[0,1] \times [0,1]$ of the cartesian plane, such that x and y are each rational (we call this the set of rational points in the unit square) is not measurable in the sense of Jordan area measure.

If $f(x)$ is a bounded, nonnegative, real valued function defined on an interval $[a,b]$, and if S denotes the set $S = \{(x, y) : x \in [a,b] , 0 \leq y \leq f(x)\}$ (the so-called "ordinate-set" of f; or "the region under the graph of f") then it can be shown that f is Riemann-integrable on $[a,b]$ if and only if S is Jordan measurable, and (provided that we use the "natural" Jordan measure function which is related to the coordinate structure of the plane)

$$\int_a^b f(x)dx = \mu(S) .$$

The set of rational points on a line interval is Jordan measurable as a plane set, and its Jordan "area" is zero. [You may prove this by considering a sequence of rectangular regions r_n (of respective widths $\frac{1}{n}$, and each with the same length as the given interval) each of which contains the interval.] However, the set of rational points in the unit square (which is a countable union of sets, each of which is congruent to the set of rational points on a unit interval) is not Jordan measurable. This is the characteristic defect of Jordan measure: that countable unions of Jordan measurable sets need not be Jordan measurable, even when the union is bounded. This defect has an exact counterpart for Riemann integration: the limit of a sequence of Riemann-integrable functions (defined on the same interval) need not be Riemann integrable, even when the limit function is bounded.

These deficiencies are removed by the more general theories of Lebesgue measure and Lebesgue integration. The theory of Lebesgue measure for plane sets can be developed almost exactly as for linear Lebesgue measure, but we shall not develop it here. As far as the Lebesgue outer measure of a set S is concerned, the essential difference from the Jordan outer measure is that we take the greatest lower bound of the set of "sums" of the areas of the "intervals" in those countable (not just finite) unions of "intervals", such that the union contains S. It is easy to show from this that the Lebesgue outer measure is always less than or equal to the Jordan outer measure. Moreover, for the set of rational points in a unit square the Lebesgue outer measure is easily shown to be zero. [The set is countable; name all of the points of the set in some order, $x_1, x_2, \ldots, x_n, \ldots$; for each $\mathcal{E} > 0$, "cover" the point X_n with an "interval" (a rectangular region) of area less than $\mathcal{E}/2^n$; then the series of area values converges to a "sum" which is less than $\mathcal{E}$; hence the greatest lower bound of all such "sums" is zero.] It follows that the set of all rational points in the unit square (which was not Jordan measurable) is Lebesgue-measurable, and that its Lebesgue measure is zero.

Generally speaking, the Lebesgue measure has all of the desirable properties of the Jordan measure, as well as the additional property (with respect to countable unions) mentioned above. In particular, the two measures (derived, of course, from the same underlying area function for rectangular regions) agree on every Jordan measurable set, so that Lebesgue measure is a true generalization of Jordan measure. There are, of course, plane sets which are not Lebesgue measurable, and these can be very complicated. However one such set can be easily described using a basic property of Lebesgue measure. This property is that if X is a subset of a line (which we take as the x-axis) and if f is the so-called characteristic function of X (the function which has value "1" on points of X, and "0" elsewhere) then X is measurable in the sense of Lebesgue linear measure, if and only if the ordinate set of f is measurable in the sense of the Lebesgue "area" measure. Moreover (with the natural assumptions concerning the correspondence of the underlying length and area functions) the two measures agree. It follows that if X is the non-Lebesgue-measurable subset of the line described in Section 2-9, then the ordinate set of its characteristic function is not Lebesgue-measurable as a plane set.

Surface Area. So far we have only discussed the measurement of area for subsets of a plane. This does not enable us to deal with such simple surfaces (surface is a term which we shall not attempt to define) as spheres, circular cylinders, cones, surfaces of revolution, etc.

In mathematics, as elsewhere, analogy is a powerful device for suggesting ways of dealing with new situations, and sometimes (as with Lebesgue measure) we are able to adopt suitable definitions so as to enable us to deal simultaneously with a whole class of analogous problems. When we come to consider the notion of the area of a region on a surface which is not necessarily plane, we might expect that our experience with the notion of curve length would be of some help. Let us see how far this kind of thinking can take us.

To begin with, we have to decide on at least some of the properties which area functions for suitably defined surface regions should have: Among these we would surely include the "natural" properties:

(i) The definition of "surface region" should include plane regions, and the area functions for surface regions should be extensions of those for plane regions.

(ii) Congruent regions should have the same area.

(iii) A suitable area function should be finitely additive for disjoint unions of regions.

Guided by our experience with space curves, we might define a simple surface to be a continuous image of some suitably restricted plane region (such as a polygonal region) under a function which is 1-1 except possibly on the boundary. This definition would include such regions as polyhedral surfaces (see below), spherical surfaces, and cylindrical surfaces. [As we have not given a definition of the essentially topological notion of continuity, we cannot prove these statements, but it is not hard to make them plausible.] By a polyhedral surface we mean a surface which is like a polygonal region, except that we relax the condition that it be a subset of a plane. That is, we define a polyhedral surface to be a subset of space which can be expressed as a finite union of triangular regions with pairwise disjoint interiors. Thus a polyhedral surface is "piecewise plane", and the plane pieces are polygonal regions. (Cf. the notion of broken segment.) Thus, for some "triangulation", a polyhedral surface is piecewise congruent to a finite set of disjoint triangular regions on a plane (such a set is a polygonal region), and the combination of these congruences gives a continuous function which is 1-1 except possibly on boundary points.

To envisage a spherical surface as such a continuous image of a polygonal region, first take the surface of a cube (this is easily shown to be a polyhedral surface) which is concentric with, and contained in the sphere, and then "project" points of the cube surface onto the sphere by "projection" from the center; this "projection" is a 1-1 continuous function on the cube surface, and the composite of the "cube function" with this projection is the desired function. Similar procedures can be used to show that circular cylinders, cones, and other simple surfaces would be included in such a definition of simple surface, but of course such a definition also gives many surfaces which are certainly not "simple" in the everyday sense!

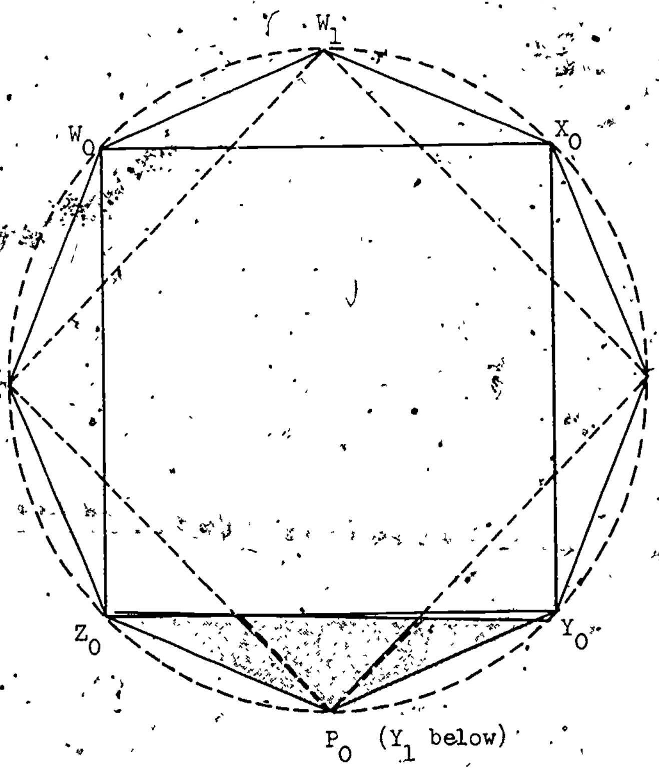
Our next step is to consider how to define suitable area functions for such surfaces. For polyhedral surfaces (as for broken segments), there is really no choice: given any area function for polygonal regions, its extension to polyhedral surfaces must be such that it is additive with respect to the areas of the triangular regions in any triangulation, and the proof of the single-valuedness of this definition (i.e., independence of triangulation) is essentially the same as that for polygonal regions.

After polyhedral surfaces, the next class of surface regions which we might consider are those which are called "developable". To make this idea precise we would need some concepts from differential geometry, but the intuitive idea is simple enough. Intuitively, a simple developable surface is one which can be "flattened out" to a plane surface without tearing or stretching (i.e., without changing what we intuitively think of as the surface area) and a developable surface is a finite union of such simple developable surfaces, with (at most) boundary points of intersection. Simple and well-known examples are: the curved surface of a cylinder; (to see this, cut on a line parallel to the axis, and "unroll"); the curved surface of a cone; (to see this cut on a line through the vertex). We are not likely to be satisfied with any definition of area function for surface regions, which does not give the "expected values" for developable surfaces. For example, consider a right circular cylinder of height h and base of radius r (under a given length function λ, with related area function $\eta(\lambda) = f$). We would expect that any reasonable extension of the domain of f should include cylindrical surfaces, and that the extended function must have the value $2\pi rh$ on the curved surface of the cylinder.

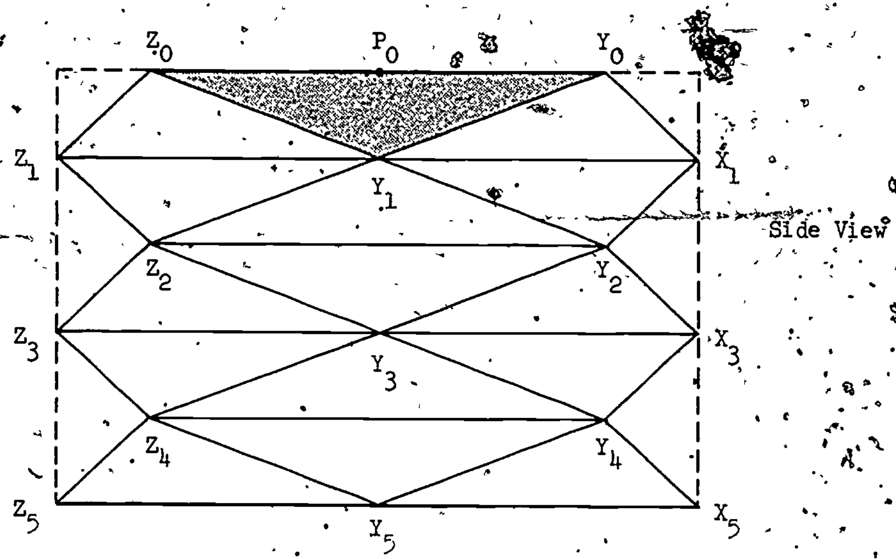
As a next step (motivated by our treatment of curve length for simple curves, where we took the least upper bound of the lengths of approximating broken segments) we might try to define surface area in terms of approximating polyhedral surfaces. The most natural way to do this for a surface S, given as a suitable continuous image $(S = g(p))$ of a polygonal region p, would be to use the triangulations of p. That is, if $(p, \{\delta_i\})$ is a triangulation of p, with $\delta_i = \blacktriangle A_iB_iC_i$, $(i=1, 2, \ldots, n)$, and if f is any area function for polygonal regions, we could consider the polyhedral region $P = \bigcup_{i=1}^{n} \blacktriangle g(A_i)g(B_i)g(C_i)$ (whose vertices $g(A_i)$, $g(B_i)$, $g(C_i)$, $(i=1, 2, \ldots, n)$ all lie on S) as a "polyhedral approximation" to S; and we could consider the area $f(P) = \sum_{i=1}^{n} f(\blacktriangle g(A_i)g(B_i)g(C_i))$ as an approximation to the desired area value of f on S. We would then try to define $f(S)$ as the limit (in some sense; e.g., the supremum) of the $\{f(P)\}$ for all triangulations of p. Unfortunately this apparently natural idea just doesn't work, as we can show you by means of a simple example.

The Failure of Polyhedral Approximation Methods in The Treatment of Surface Area. For our example we consider the curved surface S of a right circular cylinder with radius a and height h, using some fixed length function λ, and the related area function $(\eta(\lambda) = f)$ for polygonal regions. We shall describe polyhedral approximations to S directly on the surface, but it is not difficult to express this surface as the image of a suitable function defined on a plane polygonal region (e.g., the rectangular region with sides of lengths $2\pi a$ and h, obtained by "cutting and unrolling" the curved cylinder surface) and to show how the polyhedral approximation which we describe, derives from a suitable triangulation of the plane region.

Take $n + 1$ equally spaced parallel planes (at distances h/n apart, and each parallel to the circular base of the cylinder) cutting the surface in the congruent circles C_0, C_1, ..., C_n, where C_0, C_n, are the "ends" of the cylinder. On each of these circles C_i take four equally-spaced points (i.e., at the vertices of an inscribed square) W_i, X_i, Y_i, Z_i, with the points on every second circle located "above" the mid-points of the arcs of the circle which is immediately below it. (See diagram, plan view, showing the points on circles C_0 and C_1.) Connect up these



Plan View



points as indicated on the diagram to give a polyhedral approximation P_n to S, consisting of the union of $8n$ congruent isosceles triangular regions. (In the diagrams, the cylinder itself is indicated by dotted lines, and the accordian-like polyhedral approximation by full lines.) The set of all such polyhedral regions, for all positive integers n, is contained in the set of all polyhedral approximations to S. Hence the least upper bound (if it exists) of the numbers $8nf(\delta_n)$ (where δ_n denotes one of the congruent triangular regions in the polyhedral approximation P_n) must be less than or equal to the corresponding least upper bound taken over all polyhedral approximations. But it is easy to see that the set of numbers $\{8nf(\delta_n)\}$ is unbounded (see below for detail). Hence $\sup_n\{8nf(\delta_n)\}$ does not exist. It follows that the corresponding least upper bound over all polyhedral approximations cannot exist. But we have already agreed that any reasonable extension of f must have the value $2\pi ah$ on S. It follows that any attempt to define the area of S (with respect to f) to be the least upper bound of the areas of all polyhedral approximations, is quite unworkable.

We still have to show that the set of numbers $\{8nf(\delta_n)\}$ is unbounded. This can be done without getting involved in a detailed notation. Let P_0 be the point on the circle C_0 which is immediately above Y_1. (See diagrams.) Then the triangular region $\blacktriangle Z_0Y_0P_0$ is a "projection" of

$\blacktriangle Z_0Y_0Y_1$. For every n, $\overline{Z_0Y_0}$ is a common base for these triangles, and the corresponding altitude of $\Delta Z_0Y_0P_0$ is the "projection" of the corresponding altitude of $\Delta Z_0Y_0Y_1$. Hence the area $f(\blacktriangle Z_0Y_0P_0) < f(\blacktriangle Z_0Y_0Y_1) = f(\delta_n)$. This is true for every value of n, so that we have, for all n, $f(\delta_n) > f(\blacktriangle Z_0Y_0P_0) = c$ (say), where the positive number c does not depend on n. It follows that the area $f(P_n) > 8nc$. Hence, from the archimedean property of the real numbers, the set $\{f(P_n)\}$ is unbounded, and hence has no least upper bound.

We might analyze the above example in an attempt to discover why the "obvious" procedure fails. This analysis might suggest that we should have restricted ourselves to polyhedral approximations with "small" triangular regions, and taken some sort of limit as the "mesh" of such a triangulation tends to zero. By this we mean that the "diameter" of each triangle (i.e., the length of the largest segment contained in the triangle) should tend to zero. We could accomplish this by letting m (instead of 4) be the number of equally-spaced points on each circle C_i, and looking for some sort of limit as $m \to \infty$, $n \to \infty$. There are various ways of formulating such a limit notion, but unfortunately, although a limit exists for some formulations, the value of the limit depends on which formulation is used. [You might like to look at our example with $m = n$, and consider the limit as $m \to \infty$; in this case the "correct" value for the area is obtained for the limit. The amount of trigonometry needed to show this is not very great. If n is taken equal to m^2, the corresponding limit as $m \to \infty$ also exists, but it has a value which is greater than $2\pi ah$.] It is a fact that, with a more careful (and more complicated!) definition of "surface" and with suitable definitions of the necessary limit concepts, the "correct" value for the area is obtained as the greatest lower bound of the set of such limits, but this treatment involves some difficult and subtle questions which go beyond the scope of this book. Perhaps the essential source of the difficulties lies in the fact that, whereas the interpolation of segments (in a broken segment approximation to a curve) involves approximating the curve by a piecewise linear function which agrees with the curve function on the boundaries of the small segments, polyhedral approximation does not similarly agree on the boundaries of the small triangular regions, but only on the vertices of these regions.

In most treatments of surface area in advanced calculus and differential geometry, the very real difficulties are overcome by adopting a much restricted definition of surface and of surface area. (Roughly speaking, these restrictions require the surface to be "smooth" and to have a unique tangent plane, at "most" points.) This restricted category of surfaces includes such simple surfaces as polyhedral surfaces, developable surfaces, spheres, and many other surfaces of revolution; and the area functions agree, of course, with the earlier defined area functions for polyhedral surfaces.

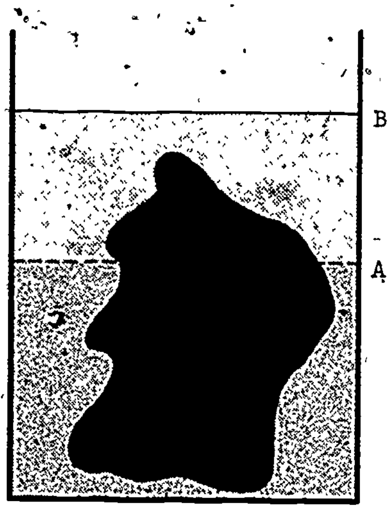
3-7. The Measurement of Volume

In its relationship to the measurement of length, the measurement of volume has much in common with the measurement of area. For this reason we shall not go through the same sort of details again, but we shall concentrate instead on those aspects of volume measurement which differ in some significant way from the corresponding area situation. For a more detailed treatment (especially concerning the volumes of many simple solid regions) you are referred to the excellent chapter on volume in [14].

The Empirical Measurement of Volume. In the discussion of area measurement we pointed out that there is no useful general method for the direct empirical measurement of area, and that most area measurement is carried out indirectly -- usually by the assumption of a mathematical model, and by calculation from certain length measurements. A great deal of volume measurement is also carried out indirectly, but there is an important direct procedure with a wide range of applicability, for the empirical measurement of the volumes of both solids (by liquid displacement) and liquids. We assume that you are generally familiar with these ideas, and we shall not go into them in detail. As usual with empirical measures, a number of physical assumptions are made concerning the "rigidity" of containers, the invariance of the volume of liquids when transferred from place to place, the volume invariance of a solid when immersed in a liquid and the volume equivalence of such a solid and the amount of liquid which it "displaces" when fully immersed, and so on.

From the point of view which we have adopted in this book, the first requirements for the establishment of an (empirical) volume function are the recognition of a category of domain elements which possess "volume", and the establishment of procedures for deciding equivalence and ordering, with respect to volume. A notion of "combining" volumes must then be introduced, and we look for volume functions, on the selected domain, which have positive real number values, which preserve equivalence and order, and which are "additive" in the sense that the value of the function on the "combination" of two domain elements is the sum of the values on the separate elements. We assume that the situation will be very much as it was for length measurement; namely, that the domain will be, in effect, an ordered abelian semigroup; that there will be infinitely many suitable volume functions, all of which are similar to one another; and that a particular function may be determined and named by the selection of a unit. (That is, by assigning the value "1" to some arbitrarily selected equivalence class of domain elements.) There is no doubt that, even if we had no prior ideas about length, a workable direct method of volume measurement, with a significant domain, could be established in this way. In fact, as we remarked above, such procedures are the basis for a great deal of practical volume measurement, with the additional feature that the names of the units/functions are frequently determined by assuming a specific relationship of volume and length functions. (Thus we have a "cubic inch" function, whose "unit" is a cube whose side has a length of one inch.) The existence and properties of this 1-1 correspondence of length and volume functions are more easily discussed in a mathematical (geometrical) context, but we want to show you by means of an example (which we shall refer to later in connection with the notion of dimension) that there are other ways of relating length and volume functions; and that, from an empirical standpoint, one of these arises quite naturally.

Let us assume that we have established (as suggested above) an empirical procedure for the comparison of the volumes of certain solids, and certain "quantities" of liquids, and that we wish to proceed to the final step of actually setting up specific volume functions. Assume that we have available cylinders of constant cross section. (E.g., right cylinders whose normal plane cross-sections are all congruent to a fixed plane region r.) Assume also that we have a well-developed procedure for length measurement, and that λ is a particular length function. For each object d (in the domain of objects to which our methods are applicable) we "measure" its volume by immersing it in liquid in one of our cylindrical containers, and assigning to d the value $v_r(d) = \lambda(\overline{AB})$, where $\overline{AB}$ is a segment representing the difference in the original and the final level of the liquid in the container. (See diagram.) Without going into details, it is fairly apparent that v_r will be a perfectly satisfactory volume function for the domain of objects considered; (i.e., it will preserve the domain structure). Each



length function λ will determine a unique volume function v_r, which will also depend, of course, on the cross-section region r. The resulting function $\nu_r : \lambda \to v_r$ from the set of all length functions to the set of all volume functions will be a 1-1 correspondence, with the property that $\nu_r(k\lambda) = k\nu_r(\lambda)$. You will recognize the similarity of this example to the one which we used in Section 3-5 (motivated by the construction used in the proof of Bolyai's Theorem) to show the existence of essentially-different 1-1 correspondences of length and area functions, leading to different isomorphisms of the automorphism groups of the sets of length functions and area functions.

<u>The Theory of Volume For Rectangular Parallelepipeds</u>. The theory of volume for the domain D_r of rectangular parallelepipeds can be carried through (in the context of euclidean geometry) in exactly the same way as the theory of area for rectangular regions, except that it is more difficult to draw relevant diagrams. We shall state the main results and leave the details to you.

A <u>volume function</u> $v : D_r \to R^+$ is defined to be a function which has the same value on congruent rectangular parallelepipeds, and which is finitely additive with respect to rectangular parallelepipeds which are expressed as finite unions of rectangular parallelepipeds with disjoint interiors. Let p_0 be a fixed cubical region with edges congruent to a segment σ_0, let λ_0 be the length function with unit σ_0; and let p be a rectangular parallelepiped whose adjacent edges are the segments σ_1, σ_2, σ_3. Then, if v is a volume function for D_r, (cf. Theorem 3-5.1; the existence of such a function has, of course, to be shown) it is easy to prove that:

(i) $v(p) = [\lambda_0(\sigma_1)] [\lambda_0(\sigma_2)] [\lambda_0(\sigma_3)] [v(p_0)]$.

(ii) There is a unique volume function v_0 for which $v_0(p_0) = 1$.

Guided by these results, we test the function v_0, defined by

$$v_0 : p \rightarrow [\lambda_0(\sigma_1)] [\lambda_0(\sigma_2)] [\lambda_0(\sigma_3)] ,$$

and we may prove, as in Section 3-5, that

(iii) v_0 satisfies the congruence and finite additivity requirements, and hence v_0 is a volume function for D_r.

(iv) Every volume function maps D_r onto R^+, and it also maps the sub-domain D_c of cubical regions onto R^+.

(v) Each two volume functions for D_r are similar, and the set of all volume functions for D_r is a ratio scale.

(vi) The correspondence $\nu : \lambda_0 \rightarrow v_0$ determines a 1-1 function from the set of all length functions onto the set of all volume functions.

(vii) $\nu(k\lambda) = k^3\nu(\lambda)$, for all $k \in R^+$, and all length functions λ.

The correspondence ν is the one from which volume functions are usually named, as cubic inch (or in^3), cubic centimeter (cm^3) etc. The result (vii) indicates the well-known way in which this system of nomenclature is related to changes of "scale", or "unit".

Volume Functions For Polyhedral Regions. We can give a definition of polyhedral region which is analogous to that for polygonal region (i.e., as a finite union of tetrahedral regions with pairwise disjoint interiors) and attempt to carry through a treatment of volume for polyhedral regions, which is comparable to the elementary theory of area functions. A volume function is defined to be one which has same value on congruent polyhedral regions, and which is additive with respect to the finite union of regions with pairwise disjoint interiors. The next step is to attempt to derive the "form" of possible volume functions for polyhedral regions, by showing that the volume "formula" for a tetrahedron must be $\frac{1}{3} \times$ area of base $\times$ height (using any length function λ, and the related area and volume functions $\eta(\lambda)$ and $\nu(\lambda)$). Very quickly we encounter a difficulty: we do not seem to be able to prove that the expected (and well-known) analogue of Theorem 3-5.6 (i.e., that tetrahedral regions with congruent bases and congruent corresponding altitudes must have the same volume) must follow from our congruence and additivity assumptions. We can prove (without using Cavalieri's Principle - see below) that every triangular prism is equivalent (in the sense of piecewise congruence under decomposition) to a right triangular prism on its "normal section" as base, and that this is equivalent to a right rectangular prism (i.e., a rectangular parallelepiped). Hence (using the naturally-related length, area, and volume functions) we can prove the volume "formula" for a triangular prism (area of base $\times$ altitude). Moreover, given any tetrahedron, we can find a triangular prism with a decomposition into three tetrahedra, one of which is the given one, and each pair of which have congruent bases and altitudes. But we cannot prove that tetrahedra with congruent bases and altitudes must have the same volume. The question as to whether this could be proved (within the framework of classical geometry) as a consequence of our congruence and additivity assumptions for volume functions, was posed by Gauss, and solved (in the negative) by Dehn in 1902.

The difficulty represented by Dehn's negative result is usually overcome, in elementary work, by an assumption which is known as Cavalieri's Principle. Roughly stated, this principle asserts that two solids have the same volume if, for each plane parallel to some fixed plane, the two "cross-sections" have the same area. (For a precise statement of this principle, see [14].) Cavalieri was an Italian mathematician of the early seventeenth century, who attempted to put on a better mathematical basis the kind of plausible "limit" arguments used by Pythagoras in deriving the (correct) area and volume formulas for many simple regions. Cavalieri was not successful in this, and his so-called "principle" was not proved until the later development of calculus and measure theory, in which, with suitable definitions of the terms involved, it becomes a theorem.

Although we now know (from Dehn's Theorem) that any elementary attempt to prove the volume formula for a tetrahedron is bound to fail, this does not prevent us from "finding" the formula by the plausible methods of Pythagoras and Cavalieri, and then going back to test whether or not the corresponding function for polyhedral regions has the same value for every tetrahedral decomposition of a given region. It is not hard (after you have obtained suitable diagrams) to prove the analogue of Theorem 3-5.5 (in effect, that the volume "formula" for a tetrahedron does not depend on the choice of base) but it would be a formidable task indeed to give a detailed proof of the analogue of Theorem 3-5.7, and we shall certainly not attempt to do this. It is much easier to prove that volume functions do exist for polyhedral regions by first establishing the more general theories of Jordan or Lebesgue, and then proving that polyhedral regions are measurable sets in either of these theories. The theory of Jordan "volume" measure, and the more general theory of Lebesgue, can be carried out quite similarly to the corresponding measure theories for the line and the plane.

We conclude this very brief discussion of volume by remarking that the negative result of Dehn is closely related to the question of a possible "3-dimensional" equivalent of Bolyai's Theorem. It can be proved (but the proof is quite difficult) that there is no such equivalent, and that polyhedral regions may have the same volume without being piecewise congruent under finite polyhedral decomposition. In fact it can be shown that there are infinitely many equivalence classes of polyhedra (with respect to the equivalence relation of piecewise congruence under finite decomposition) which have the same volume.

Chapter 4

MEASUREMENT AND DIMENSION

4-1 Introduction

The word "dimension" is used in everyday life, as well as in mathematics and science, in a variety of ways. For example, we have such everyday usages as "the dimensions of this box are 3 feet, 2 feet, and 18 inches"; "the dimensions of this room are 20 feet by 12 feet"; "a draftsman dimensions his drawing"; "the problem of airport noise will take on added dimensions with the introduction of supersonic aircraft".

In mathematics, we speak of the dimension of a euclidean space; the dimension of a vector space; and the dimension of a topological space. Each of these three ideas (euclidean dimension, linear dimension, and topological dimension) has its own definition, and each is applicable to a certain set of "objects". In other words, each is a function with its own well-defined domain and each might reasonably be considered to be a measure function in the broad sense. The values of these functions are integers or cardinal numbers, and the definitions of these functions are such that (unlike, for example, length functions and area functions) they are unique.

In this chapter we are going to discuss another usage of the word "dimension", which arises whenever we have ratio scales (i.e., sets of similarity-related functions, such as length functions, or area functions) and a certain type of relationship between such sets. Speaking rather informally at this stage, the "dimension" of one class of such functions with respect to another will be a function of a specific relationship between the classes. Thus it will turn out that the dimension of the class of area functions with respect to the class of length functions, under the "natural" correspondence η defined in the last chapter, will be "2"; while the dimension of the class of area functions with respect to the class of length functions under each correspondence η_σ, will be "1".

Classes of similar functions with correspondences between them, arise typically in measurement situations (especially in science, but also, as we have seen in a purely mathematical context, originally motivated by, but not logically dependent on, empirical ideas). Hence the notion of dimension which we will discuss here is motivated by measurement questions. There is no simple common name (such as topological dimension, linear dimension) for this kind of dimension notion (itself a measure function in the broad sense) so we refer to it as <u>measure dimension</u>. There are connections between all of the common dimension notions, including measure dimension, but we do not intend to explore these connections in this book.

You have undoubtedly encountered the notion of measure dimension in your courses in mathematics and physics, in such expressions as

"area has dimension 2 with respect to length";

"velocity has dimension 1 in length and dimension -1 in time".

These statements are often expressed symbolically by such notations as $A = L^2$ and $V = LT^{-1}$. You will also be familiar with the fact that "dimensional relationships" of this type can be quite complicated, and that there often appears to be something mysterious about them: the nature of the symbols $A, L, T, V, \ldots,$ is usually left rather obscure, and yet they are often manipulated as if they belong to some simple algebraic system, such as a multiplicative group. It is one of the purposes of this book to make clear just what is involved in the notion of measure dimension, what is the domain of a measure dimension function, and what is the equivalent, in the language of this book, of such symbolic statements as $V = LT^{-1}$. As you will see, the basic mathematical ideas are relatively simple, but by no means trivial.

You might be surprised that we are initiating the discussion of measure dimension when we have only four ratio scales "on hand"; namely, the scales for length, angularity, area, and volume. The reason we are doing this is that most of the mathematically significant aspects of the notion of measure dimension can be discussed with these simple examples in mind, and that we are unlikely to understand the more complicated situations if we cannot understand the simple ones.

We must emphasize that our discussion is almost entirely mathematical. The length functions and the area functions that we refer to will be those which we have described above, whose domains are subsets of euclidean space. The properties which we need in order to discuss the notion of measure dimension are properties which we have proved or shall prove. Whether such properties hold for the corresponding empirical functions (and for the other empirical measure functions used in science) is an empirical, and not a mathematical question. It is, of course a scientific hypothesis that such properties do hold in those situations to which the mathematical treatment is relevant.

Before getting involved in details, there are a few points which should be made clear. Most books on measurement--in fact most books on physics--deal with the subject of dimension. But their objective is generally to achieve a working familiarity with the notion, and no serious attempt is generally made to develop a mathematical theory. The "facts" are generally summarized, and illustrated with many examples. Our objective is, in a sense, complementary to this kind of treatment, as our principal concern is to explain the mathematics which lies behind the "facts".

Among the many books and papers devoted to this subject, Bridgman's "Dimensional Analysis" [17], now over forty years old, stands out as a classic, and we commend it to your attention. Other books, which you will find particularly useful as sources of applications are [3] and [18]. In addition there is an extensive and still-growing literature on this question, (books, and articles in scientific, philosophical, and occasionally mathematical, journals) and it is safe to say that all problems--mathematical, empirical, and philosophical--are by no means settled.

4-2 Ratio Scales

Before proceeding to a formal definition of measure dimension, let us review the notion of what (following Stevens) we have called a ratio scale. In our discussion of length functions we found that (for the domain D of segments) there were many (in fact infinitely many) length functions, each of which was a function $\lambda : D \to R^+$. (We assume henceforth the archimedean and Cantor-Dedekind postulates, so that each λ is onto R^+:) We proved that if λ is any length function, and $k \in R^+$, then $k\lambda$ was also a length function. Moreover we proved that any length function could be obtained from any other in this way. (I.e., by multiplication with any k, in terms of the multiplication in the algebra of functions; or equivalently, by composition with the corresponding automorphism $\bar{k} : x \to kx$ of $(R^+, +)$.)

We also noted earlier that if Λ denotes the set of all length functions, then, corresponding to each $k \in R^+$, the function

$$\bar{\bar{k}} : \Lambda \to \Lambda$$

defined by $\bar{\bar{k}}(\lambda) = k\lambda = \bar{k}\lambda$ is a 1-1 correspondence of Λ onto Λ, and $\bar{\bar{k}}$ has the properties that, for all $k' \in R^+$, and all λ_1, and $\lambda_2 \in \Lambda$

(i) $\bar{\bar{k}}(\lambda_1 + \lambda_2) = \bar{\bar{k}}(\lambda_1) + \bar{\bar{k}}(\lambda_2)$;

(ii) $\bar{\bar{k}}(k'\lambda) = k'(\bar{\bar{k}}(\lambda))$.

That is, $\bar{\bar{k}}$ is an automorphism of the R^+ - semimodule $(\Lambda, +, R^+)$. It is not difficult to prove that

(i) $\{\bar{\bar{k}} : k \in R^+\}$ is the set of all automorphisms of $(\Lambda, +, R^+)$;

(ii) this set of automorphisms (which we denote by A_Λ) is a group, which is naturally isomorphic to the group of automorphisms of $(R^+, +)$. As we proved in Section 2-2, this group of automorphisms is isomorphic to the group $(R^+, \cdot)$. Hence the function $\varphi_\Lambda : k \to \bar{\bar{k}}$ is an isomorphism of the groups $(R^+, \cdot)$ and $(A_\Lambda, \circ)$. (Here the symbol "$\circ$" denotes composition of automorphisms.)

Whenever we have a set of functions (not necessarily arising in a specific measurement situation) with a common domain, which are related to one another like length functions, we refer to such a set as a ratio scale. Specifically, a ratio scale is a set M of functions, with values in R^+, such that

(i) each function in M has the same domain (it is sometimes convenient to call this the domain of the ratio scale M);

(ii) composition of any function in M with a similarity transformation $\bar{k}$ of R^+ (i.e., an automorphism of $(R^+,+)$) gives another function in M; (equivalently, M is closed under multiplication by positive real numbers);

(iii) every two functions in M are similar; (i.e., related by composition with an automorphism of $(R^+,+)$).

It follows that if D denotes the domain of M, then

(iv) if a_1, a_2, belong to D, the ratio $f(a_1)/f(a_2)$ has the same value for every $f \in M$; hence two elements a_1, a_2, of D, have the same value under one function in M if and only if they have the same value under every other function in M. That is, the relation defined by: $a_1 \sim a_2$ if and only if $f(a_1) = f(a_2)$ is an equivalence relation on D, and this relation is the same for every $f \in M$;

(v) if f_1, f_2, belong to M, the ratio $f_1(a)/f_2(a)$ has the same value for every $a \in D$;

(vi) composition of functions in M with a fixed automorphism of $(R^+ , +)$, yields an automorphism of $(M , +)$;

(vii) the groups of automorphisms of $(R^+ , +)$ and of $(M , +)$ are isomorphic, and each is naturally isomorphic to the group $(R^+ , \cdot)$; this isomorphism can be used to give an ordering to the automorphism group of the set of functions, M ;

(viii) every function f in M "generates" M , in the sense that, for each f , $M = \{kf : k \in R^+\}$;

(ix) in addition to functional addition in M , there is a "scalar multiplication" by any positive real number, and $(M , + , R^+)$ is an R^+ - semimodule;

(x) if $\tilde{D}$ denotes the set of equivalence classes of elements of D under the relation defined in (iv) ~~above, a~~ "scalar multiplication" of elements of $\tilde{D}$ by some positive real numbers can be defined (as in Section 2-5), and an "addition" can sometimes be ~~defined by~~ "working backwards". (I.e., if $f(a_1) + f(a_2) = f(a_3)$, define $\tilde{a}_1 + \tilde{a}_2 = \tilde{a}_3$.) As far as these operations are defined, $\tilde{D}$ has a structure like an R^+ - semimodule. Moreover there is a natural 1-1 mapping from $\tilde{D}$ to the dual-space of $(M , + , R^+)$.

If each function in M is onto R^+ , we say that the ratio scale is complete; otherwise it is incomplete. Most of the ratio scales that we have encountered in a mathematical context, are complete. The angular measure scale for simple angles is an example of an incomplete ratio scale. Some of the theorems below concerning ratio scales, hold only when the scale is complete. As we pointed out in connection with length scales, it is easy to prove that a ratio scale is complete if and only if every function in the scale has a "unit". In this case the structure of $\tilde{D}$ is also "complete"; i.e., $(\tilde{D} , + , R^+)$ is also an R^+ - semimodule, and $(M , + , R^+)$ is, in effect, the dual space of $(\tilde{D} , + , R^+)$. The fact that $(M , + , R^+)$ is an R^+ - semimodule whether or not M is a complete scale, is one of the reasons why it is easier to work with M than with $\tilde{D}$, in a discussion of measure dimension.

There are two rather different aspects of the question of completeness. One concerns the existence of arbitrarily large values and arbitrarily small values for each function. We encountered this question in connection with angular measure functions for simple angles, where the range was only an initial segment of positive real numbers; the procedure there adopted (of formally enlarging the domain by considering finite sets -- or "formal sums" -- of domain elements) can sometimes be applied to extend the scale to a complete scale. The other aspect of the onto-ness question, is whether, given any two real numbers in the range, all real numbers between these also appear as values. We remarked earlier that, in the actual recording of measurement "readings", rational numbers are usually used. And, as these are a dense subset of the reals (this is a topological property, which means, roughly speaking, that rational numbers can be used to approximate every real number arbitrarily closely) you might think that we could manage with scales whose values were restricted to the set of positive rational numbers. But, in the mathematical analyses which arise from empirical situations, we are frequently concerned with functions whose domains and ranges are real numbers, and these real numbers are the values of measure functions. In these analyses we need to deal with power functions for non-integral powers, exponential and logarithmic functions, differentiation and integration, and so on; and a restriction to rational arguments and values would prevent the use of most of these functions and operations.

Relationships Between Ratio Scales. The fact that the set Λ of length functions was a ratio scale was shown in considerable detail. The set Γ of angular measure functions on the set of generalized angles is also a ratio scale, as is the set Δ_r of area functions on the domain of rectangular regions, the set Δ_p of area functions on the domain of polygonal regions, and even the set Δ_s of area functions on the domain of square regions. It follows that the automorphism groups of each of these ratio scales are all isomorphic (as ordered groups) to $(R^+, \cdot, <)$.

In many situations which arise in connection with measurement, we are concerned with relationships between ratio scales. These relationships between ratio scales can be conveniently expressed in terms of monotone homomorphisms (which are usually isomorphisms) of their automorphism groups. (The monotone condition implies continuity, and conversely. In many scientific contexts continuity is possibly the more intuitive idea, but it is simpler for us to use the monotone condition rather than a continuity condition; cf. the relationship $\eta : \Lambda \rightarrow \Delta$ of length and area functions, which induced the mapping $\bar{k} \rightarrow \bar{k}^2$. The corresponding power function $k \rightarrow k^2$ on $(R^+, \cdot)$ is, of course, a monotone, continuous isomorphism.) This is not really surprising, because in considering relationships between ratio scales, we are not generally interested in particular functions in the scales, but in the scales "as a whole". This suggests that we should look at functions (such as $\eta : \Lambda \rightarrow \Delta$) which relate the scales themselves. But when we do this, and examine the properties of such functions, we find that they do not generally preserve the additive structure of Λ, but they do preserve the "compositional", or automorphism structure. In mathematics it is quite common to study a system in terms of its automorphisms, so it should not surprise us that significant relationships between ratio scales should involve mappings (i.e., monotone homomorphisms) of their groups of automorphisms.

In mathematical situations, the properties of these relationships will have to be proved. For example, the relationship $\eta : \Lambda \rightarrow \Delta_s$ of length functions (for segments) and area functions for square regions, defined by $\eta(\lambda) : s \rightarrow [\lambda(\sigma)]^2$ (where σ is a side of s) has the property (which follows from Theorem 3-5.1) that $\eta(k\lambda) = k^2\eta(\lambda)$. Thus, if A_Λ, A_{Δ_s}, denote the respective automorphism groups, η induces $\eta^* : A_\Lambda \rightarrow A_{\Delta_s}$, defined by $\eta^* : \bar{\bar{k}} \rightarrow \bar{\bar{k}}^2$. It is easily verified that η^* is a monotone isomorphism.

In empirical situations, the fact that such a monotone homomorphism exists will require justification. In many cases appeal is made to a "principle" which Bridgman calls "the absolute significance of relative magnitude". (See Chapter 2 of [17].) Like most scientific writers, Bridgman works mainly with units and values, and does not consider explicitly the structural relations on, and between, sets of measure functions. His book [17] is an excellent source of ideas, but, as you might find some difficulty in translating from his language to ours, we shall prove (in the next section) that his "principle", and other assumptions which he makes, imply that, in certain circumstances, there is a mapping from one ratio scale to another, and that the properties of this mapping are such that it induces a homomorphism on the corresponding automorphism groups. We recall that these automorphism groups are all naturally isomorphic to the well-known ordered group $(R^+, \cdot, <)$, so that we can reduce many questions concerning relationships between ratio scales to questions about the monotone (order preserving or reversing) homomorphisms of $(R^+, \cdot)$. Hence we can use the well known properties of such homomorphisms (they are all power functions) which we proved in Section 2-2. In this way we shall establish (Section 4-4) a simple mathematical foundation for the study of the subject of measure dimension.

4-3 Primary and Secondary Quantities

In Bridgman's book [17] the word "quantity", and the expressions "primary quantity" and "secondary quantity" are not defined explicitly. The word "quantity" (sometimes expanded to "physical quantity", or "measurable physical quantity") refers to such concepts as "length", "area", and "volume", whose measurement leads to the establishment of ratio scales. You will recall that we did not attempt to define such terms as "length", and "area": for our purposes it was quite sufficient to define "length function", and "area function", and an important property of the resulting sets of functions was that each set was a ratio scale. So, without attempting to define the word "quantity", we require of each "quantity" that its measurement should determine a unique ratio scale. This scale has a domain, which we refer to as the domain of the relevant "quantity".

A "primary quantity" involves the direct establishment of measure functions (leading to a ratio scale, which we sometimes refer to as a primary scale, of measure functions) by "operations" on elements of its domain. In the measurement of the value of a "secondary quantity" on a particular element b of the domain of that quantity, certain domain elements, a_i, of one or more primary quantities are associated with b, the primary quantity values (x_i) of these a_i are obtained by direct measurement, and the value (y) of the secondary quantity b is calculated by some "rule". (If you wish to have something specific in mind, think of area for squares as a secondary quantity, and length for segments as a primary quantity; for a square b, a is its side, x is the length of its side under a particular length function, and $y(=x^2)$ is the calculated value for the area of b.) [Some writers prefer to use the terms "basic", or "fundamental", where we have used "primary"; and the term "derived", where we have used "secondary".]

It is important to keep in mind that the designation of a "quantity" as primary or secondary is, to a considerable extent, arbitrary; and that a complete "system of measurement" involves the classification of certain quantities as primary or secondary, direct procedures for the measurement of the primary quantities, and explicit "rules" for the indirect "measurement" of the secondary quantities by the use of primary quantity measurement, and calculation. These rules must not only specify the calculations to be performed on the primary values, but they must also specify the rules of association for the domain elements.

Bridgman assumes that every primary quantity is "measured" by a set of functions, which form a ratio scale. In addition he assumes as a principle, "the absolute significance of relative magnitude". This principle states that, for any two elements b_1, b_2, in the domain of a secondary quantity, the ratio of the corresponding values (i.e., secondary measures) for b_1, b_2, should be the same, no matter which particular set (one from each scale) of primary scale functions were used in the process of finding the secondary measures. Now if B is the set of all elements b in the domain of the secondary quantity, then, corresponding to each choice of primary functions, we get a function $g : B \to R^+$, and the principle of "absolute significance of relative magnitude" simply says that for two such functions, g_1, g_2, and any two domain elements b_1, b_2,

$$\frac{g_1(b_1)}{g_1(b_2)} = \frac{g_2(b_1)}{g_2(b_2)} .$$

It is also assumed that all of these numbers are positive, so that, for every two domain elements b_1, b_2,

$$\frac{g_1(b_1)}{g_2(b_1)} = \frac{g_1(b_2)}{g_2(b_2)} .$$

Hence, for some $k > 0$, $g_2 = kg_1$. In other words these "g" functions for the measurement of the secondary quantity are similar. Hence they belong to, and determine, a unique ratio scale. In most books you will find that the question of whether the set of calculated functions comprise, or are contained in, a ratio scale, is not made explicit. However, if you recall the situation for radian measure in connection with the measurement of angularity, you will see that the set of "secondary" or "calculated" functions need not be the full ratio scale--in fact it can be a single function. The ratio scale determined by a secondary quantity is called a secondary scale, but this term has no absolute significance: the same scale can be a primary scale in another context. We point out that Bridgman's condition implies that all secondary functions must be similar, and hence determine a unique ratio scale. The converse is easily shown to hold; i.e., if we assume that all secondary functions are similar, then the "relative magnitude of secondary domain elements has an absolute significance". This "relative magnitude" of b_1 and b_2 is, of course, a positive real number, which is customarily denoted by $b_1 : b_2$. The function $p : B \times B \to R^+$, defined by $p(b_1, b_2) = b_1 : b_2$ is easily shown to be a ratio operation for B; the ratio operation which corresponds to the secondary ratio scale, as discussed in Section 2-5.

The Relationship of Primary and Secondary Quantities. Let $F_1, F_2, \ldots, F_n$ ($F_i = \{f_i\}$) denote the primary scales used in the description of a secondary quantity, and let $G = \{g\}$ denote the corresponding secondary scale. For any choice of a set of primary functions, the rules for the measurement of the secondary quantity give us a secondary measure function. Thus we get a mapping

$$\gamma : F_1 \times F_2 \times \ldots \times F_n \rightarrow G$$

where. $F_1 \times F_2 \times \ldots \times F_n$ is simply the cartesian product of the sets of functions F_i .

In order to simplify the discussion of the relationship of primary and secondary quantities, let us first look at the properties of such a function for the case where there is only a single primary quantity. That is, suppose that the rules for the measurement of the secondary quantity have given us a function

$$\gamma : F \rightarrow G$$

from the ratio scale F to the ratio scale G . Now if we make a simple assumption (which you will usually not find stated explicitly) then we shall see that γ must have a certain important property. The assumption which we make (and which we shall weaken later) is that, for each primary function f in F ~~, and~~ each positive real number x , there is a "b" in the domain of the secondary quantity whose associated "a" in the domain of the primary scale F satisfies $f(a) = x$. (This is not the same as assuming that every "a" appears in association with some b ; it merely says that enough such "a's" must appear so that the set of their values, under each f , is all of R^+ .)

We can picture this condition with the aid of the following diagrams, which are trivially commutative because g is defined by composition to be $\varphi f \delta$:

$$\begin{tikzcd}
A \arrow[r, "f"] & R^+ \arrow[d, "\varphi"] \\
B \arrow[u, "\delta"] \arrow[r, dashed, "g"] & R^+
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
a \arrow[r] & f(a) = x \arrow[d] \\
b \arrow[u] \arrow[r, dashed] & g(b) = \varphi(x)
\end{tikzcd}$$

δ is the function which associates a with b ; f is any function in the primary scale F ; and φ is the function determined by the "rule" for calculating $g(b)$ from $f(a)$. The assumption stated above, is that $f(\delta(B)) = R^+$ for every f. (Or, equivalently, that $f|\delta(B)$ is onto R^+.) We have already seen (by considering area functions for square regions as secondary quantities, determined from the ratio scale of length functions for segments as a primary quantity) that the function φ may be the power function $\varphi: x \to x^2$. The question of what kind of functions φ may be used in the derivation of secondary quantities, has considerable interest. We shall see later in this section that, provided that we make a few quite reasonable assumptions, the function φ must always be a constant multiple of a power function. (That is, there exist $c > 0$ and $\alpha \in R$, such that $\varphi: x \to cx^\alpha$ for every $x \in R^+$.)

Assume now that f_1, f_2, are any functions in F. Then, for some $k \in R^+$, $f_2 = kf_1$. Moreover, we have shown that Bridgman's principle of "the absolute significance of relative magnitude" implies that the functions $g_1 = \gamma(f_1)$, $g_2 = \gamma(f_2)$, are similar. I.e., there is a $k' \in R^+$ such that $g_2 = k'g_1$. We assert that (with the assumptions we have made), this k' is uniquely determined by k, and not by the specific functions f_1 and $f_2 = kf_1$. This property is stated precisely, and proved, in the following theorem:

Theorem 4-3.1. F is a ratio scale with domain A ; B is any set; δ is a mapping $\delta : B \to A$ such that $f\delta$ maps B onto R^+ for every $f \in F$; φ is a mapping on R^+ with values in R^+ ; and, for each $f \in F$, a mapping $g : B \to R^+$ is defined (by composition) as $g = \varphi f\delta$. If the functions $g = \varphi f\delta$ are all similar, so that the set of functions $\{g : g = \varphi f\delta, f \in F\}$ is part of a ratio scale G whose domain is B, and if $\gamma : F \to G$ is defined by $\gamma(f) = g$, then to each $k \in R^+$ there corresponds a unique $k' \in R^+$, such that, for every $f \in F$, $\gamma(kf) = k'\gamma(f)$.

Remark: The statement of the theorem looks quite formidable, because we have listed all of the assumptions. If you keep in mind the following diagram (which, as you can easily check, is commutative except possibly for the right "trapezoid") then you will see that the assumptions are not really complicated.

$$\begin{tikzcd}
A \arrow[rr, "kf"] \arrow[dr, "f"] & & R^{+} \arrow[ddd, "\varphi"] \\
& R^{+} \arrow[ur, leftrightarrow, "\overline{k}"'] \arrow[d, "\varphi"] & \\
& R^{+} \arrow[dr, leftrightarrow, "\overline{k_f}"] & \\
B \arrow[uuu, "\delta"] \arrow[ur, "\gamma(f) = \varphi f\delta"] \arrow[rr, "\gamma(kf) = \varphi(kf)\delta = k_f\,\gamma(f)"'] & & R^{+}
\end{tikzcd}$$

Proof. Let $k \in R^{+}$, let $f \in F$, and let b_1 and b_2 be elements of B such that $(f\delta)(b_2) = k(f\delta)(b_1)$. [This is where we use the assumption that $f\delta$ is onto R^{+}; this assumption can be weakened: see below.] Then

$$\varphi[(f\delta)(b_2)] = (\varphi f\delta)(b_2) = \varphi[k(f\delta)(b_1)] = (\varphi(kf)\delta)(b_1) = k_f(\varphi f\delta)(b_1)$$

where k_f is the similarity factor connecting the similar functions $\varphi f\delta$ and $\varphi(kf)\delta$. Hence

$$k_f = (\varphi f\delta)(b_2)/(\varphi f\delta)(b_1) .$$

But this ratio of the values of b_2 and b_1 is the same for all functions in a ratio scale, hence, in particular, it does not depend on f. That is, the number $k' = k_f$ is determined uniquely by k, δ, and φ. Thus $\gamma(kf) = \varphi(kf)\delta = k'(\varphi f\delta) = k'\gamma(f)$ for every $f \in F$.

Remark: If you examine the proof carefully, you will see that we did not use the full force of the assumption that $f\delta$ mapped B onto R^{+}. What we needed was that for every $k \in R^{+}$, there must be elements b_1, $b_2 \in B$, such that $(f\delta)(b_2) = k(f\delta)(b_1)$. In other words, the range $X = (f\delta)B$ of $f\delta$ in R^{+} must have the property that the set of all quotients of numbers in X must be the whole of R^{+}. Thus, in the proof of this theorem, we could replace the condition $(f\delta)(B) = R^{+}$, by the weaker condition

$$(*) \qquad \{x_1/x_2 : x_1 , x_2 \in (f\delta)(B)\} = R^{+} \quad ;$$

You can easily verify that this condition does not depend on a particular f; i.e., if it holds for one $f \in F$, then it holds for every other $f \in F$. Thus it is a property of δ and of the scale F. This condition will be satisfied if $(f\delta)(B)$ is an initial segment of R^+ (cf. radian measure on simple angles), and also if $(f\delta)(B)$ is a terminal segment of R^+; (i.e., the set of all positive real numbers greater than some fixed positive number); but it is not satisfied if $(f\delta)(B)$ contains only rational numbers. You can easily find other "solutions" X of the equation $\{x_1/x_2 : x_1, x_2 \in X\} = R^+$, but it is not clear whether these have any significance in the discussion of secondary measure functions.

Exercises 4-3

1. With the notation of the above Theorem, but without either the assumption (*), or the stronger assumption that $f\delta$ is onto, prove that $\delta_*F(=\{f\delta : f \in F\})$ is a ratio scale with domain B.

2. Let $\rho : B \times B \to R^+$ be the ratio operation which corresponds to the ratio scale δ_*F. (I.e., $\rho(b_1,b_2) = \dfrac{(f\delta)(b_1)}{(f\delta)(b_2)}$ for every $f \in F$.) Prove that the condition (*) holds if and only if ρ is onto R^+.

3. Prove that (*) is satisfied if $(f\delta)(B)$ is either an initial segment or a terminal segment of R^+, with or without the relevant endpoint in each case.

4. Prove that (*) is satisfied if $(f\delta)(B)$ contains all positive rational numbers and also an open interval $(a,b)(a \neq b)$ of positive reals. (Hint: prove that for each $x \in R^+$, $\{xq : q \in Q^+\} \cap (a,b) \neq \emptyset$.)

Diagram Chasing. This is a very convenient place to introduce you to an activity (associated with the use of commutative diagrams) which mathematicians refer to rather flippantly as "diagram chasing". The proof of Theorem 4-3.1 (assuming the property (*), instead of the hypothesis $f\delta$ onto) is a good example of a proof which can be visualized very simply with the aid of an appropriate diagram. We first prove a simple lemma, which gives a useful equivalent form for the condition (*).

Lemma. For every f in F, $\{x_1/x_2 ; x_1, x_2 \in (f\delta)(B)\} = R^+$ if, and only if, for every $f_1, f_2 \in F$, $(f_1\delta)(B) \cap (f_2\delta)(B) \neq \emptyset$.

Proof. If $(*)$ holds, then for any f_1, $f_2(= kf_1) \in F$, there exist b_1, $b_2 \in B$, such that $(f_1\delta)(b_1) = x_1$, $(f_1\delta)(b_2) = x_2$, $x_2/x_1 = k$. Therefore $(f_2\delta)(b_1) = (kf_1\delta)(b_1) = kx_1 = x_2 \in (f_2\delta)(B) \cap (f_1\delta)(B)$.

Conversely, for any $k \in R^+$, take f_1, $f_2(= kf_1) \in F$. Then $(f_1\delta)(B) \cap (f_2\delta)(B) \neq \emptyset$. Hence there exist b_1, $b_2 \in B$, such that $(f_1\delta)(b_1) = (f_2\delta)(b_2) = (kf_1\delta)(b_2) = k(f_1\delta)(b_2)$. Therefore $(f_1\delta)(b_1)/(f_1\delta)(b_2) = k$. Hence $(*)$ holds.

Alternate Proof of Theorem 4-3.1 Under The Weaker Hypothesis $(*)$. Consider the following diagram:

$$\begin{tikzcd}
A \arrow[rr, "kf"] \arrow[dr, "f"] & & R^+ \arrow[ddd, "\varphi"] \\
& R^+ \arrow[ur, "\bar{k}"'] \arrow[d, "\varphi"] & \\
& R^+ \arrow[dr, dashed, leftrightarrow, "?\rho"] & \\
B \arrow[uuu, "\delta"] \arrow[ur, "\varphi f\delta"] \arrow[rr, "\varphi(kf)\delta"] & & R^+
\end{tikzcd}$$

We sketch the proof, by means of a sequence of statements, all of which are easy to prove. You should follow the proof by drawing appropriate sub-diagrams. The proof is much easier to follow than to write.

(i) The top "triangle" is commutative. [Definition of kf.]

(ii) The left "trapezoid" is commutative. [Trivially: $\varphi f\delta$ is a composite function.]

(iii) The outer "rectangle" is commutative. [Trivially, as for (ii).]

(iv) For a fixed f, there is a unique similarity transformation $\rho(= \bar{k}_f$, say) which makes the lower "triangle" commutative. [The functions $\varphi f\delta$ are all similar.]

(v) The right "trapezoid" is "partially commutative". Specifically, the "sub-diagram" (in which φ' , φ'' , denote the appropriate restrictions of φ)

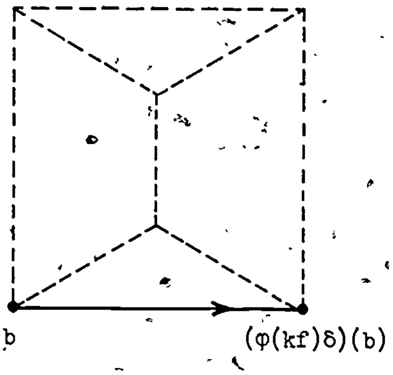
$$\begin{tikzcd}
(f\delta)(B) \arrow[r, leftrightarrow, "\bar{k}"] \arrow[d, "\varphi'"] & (\bar{k}f\delta)B \arrow[d, "\varphi''"] \\
R^{+} \arrow[r, leftrightarrow, "\bar{k}_f"'] & R^{+}
\end{tikzcd}$$

is commutative. [This is where you must indulge in "diagram chasing", to build up the following diagram:

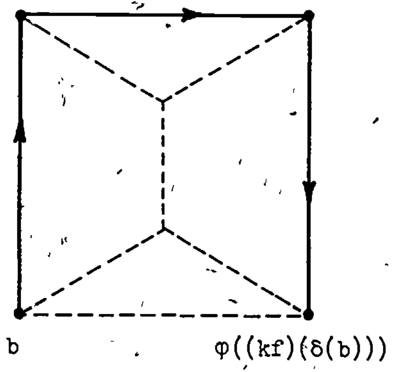
$$\begin{tikzcd}
\delta(b) \arrow[rr] \arrow[dr] & & (\bar{k}f\delta)(b) \arrow[ddd] \\
& (f\delta)(b) \arrow[ur] \arrow[d] & \\
& (\varphi f\delta)(b) \arrow[dr] & \\
b \arrow[uuu] \arrow[ur] \arrow[rr] & & (\varphi\bar{k}f\delta)(b) = (\bar{k}_f\varphi f\delta)(b)
\end{tikzcd}$$

The essential point is that, given any element in $(f\delta)(B)$, we can track it "back" to a (not necessarily unique) element $b \in B$, and then "forward" to either $(\bar{k}_f \varphi f\delta)(b)$ or $(\varphi\bar{k}f\delta)(b)$, which must be the same, because of the commutativity of the outer rectangle, the two triangles, and the left trapezoid.] The idea of this "diagram chasing" is conveyed by the following set of diagrams:

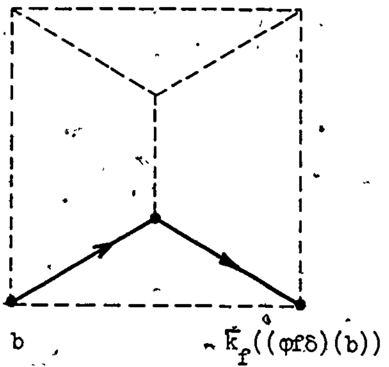
Schematically, for every $b \in B$ (with " $\doteq$ " meaning that the functions represented by the marked "function paths" must agree where they "meet")


b
$(\varphi(kf)\delta)(b)$

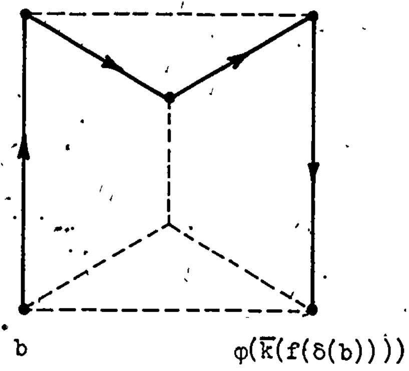
=


b
$\varphi((kf)(\delta(b)))$

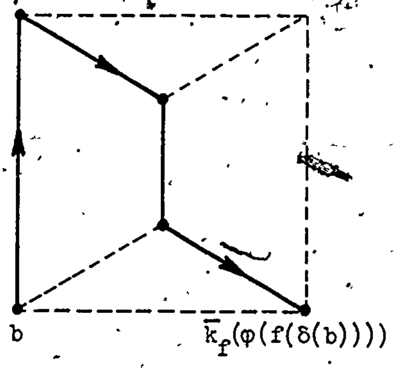
∴


b
$\bar{k}_f((\varphi f\delta)(b))$

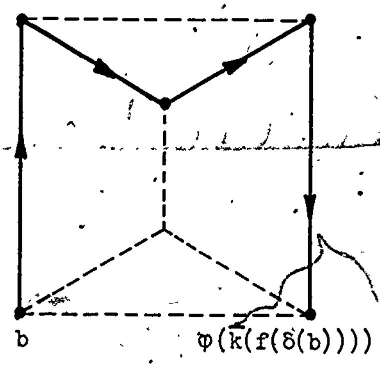
=


b
$\varphi(\bar{k}(f(\delta(b))))$

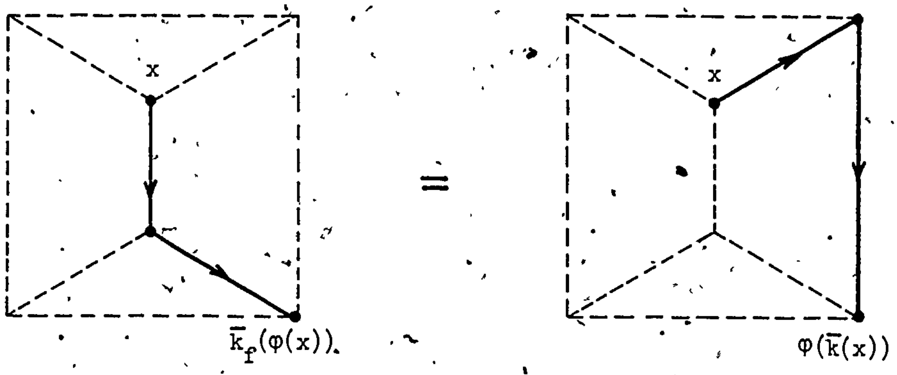
∴


b
$\bar{k}_f(\varphi(f(\delta(b))))$

=


b
$\varphi(\bar{k}(f(\delta(b))))$

Therefore, for every $x \in (f\delta)(B)$



(vi) For any $f_1, f_2 (= kf_1) \in F$, condition (*) implies that $k_{f_1} = k_{f_2}$. [(*) implies that there is an $x \in (f_1\delta)(B) \cap (f_2\delta)(B)$. From (v), $\varphi(kx) = k_{f_1}\varphi(x) = k_{f_2}\varphi(x)$, hence $k_{f_1} = k_{f_2}$.]

(vii) With $\rho = \bar{k}' = \bar{k}_f$ for every f, the right "trapezoid" (and hence the whole diagram) is commutative for every $f \in F$. [(v) and (vi) imply that (with $\rho = \bar{k}'$) the right trapezoid is commutative for $\bigcup_{f\in F} (f\delta)(B)$ in the upper left position. This union is easily seen to be the whole of R^+.]

(viii) For given k, there exists k', such that $\gamma(kf) = k'\gamma(f)$ for every $f \in F$. [This is just $\varphi\bar{k}f\delta = k'\varphi f\delta$ for every $f \in F$, which follows from (vii).]

Corollary. With the notation of the theorem, but with the assumption (*) (or the stronger assumption that $f\delta$ is onto) replaced by the much weaker assumption $B \neq \emptyset$, the existence of k' such that $\gamma(kf) = k'\gamma(f)$ for every $f \in F$, is equivalent to the existence of k' such that the diagram below is commutative for every $f \in F$.

$$\begin{tikzcd}
A \arrow[rr, "kf"] \arrow[dr, "f"'] & & R^+ \arrow[dl, leftrightarrow, "\bar{k}"] \arrow[ddd, "\varphi"] \\
& R^+ \arrow[d, "\varphi"] & \\
& R^+ \arrow[dr, leftrightarrow, "\bar{k}'"] & \\
B \arrow[uuu, "\delta"] \arrow[ur, "\gamma(f)"] \arrow[rr, "\gamma(kf)"'] & & R^+
\end{tikzcd}$$

(We leave the proof to you.)

Remarks:

1. Notice how the condition (*) (in the equivalent form $(f_1\delta)(B) \cap (f_2\delta)(B) \neq \emptyset$) enabled us to "link together" the possibly-different numbers k_f for different $f \in F$, to get a uniform value $k_f = k'$ for every f.
2. It is easy to see that the condition (*) is not a necessary condition for the conclusion of the theorem (which is equivalent to the commutativity of the above diagram). For example, if B contains only one element, (*) cannot possibly hold, but the resulting diagram can still be commutative. For this reason (see later) we will not want to require (*) in the definition of a secondary quantity which is derived from a single primary quantity. The reason for the introduction of (*) was that it does, in fact, hold in certain well-known cases; and it is a sufficient condition to insure the result of Theorem 4-3.1.

Functions Connecting Ratio Scales. A function (from one ratio scale to another) which has the property proved in Theorem 4-3.1, will be called a uniform function. More specifically, if F and G are ratio scales, a function $\gamma : F \rightarrow G$ is a uniform function if for every $k \in R^+$, there exists a unique $k' \in R^+$, such that $\gamma(kf) = k'\gamma(f)$ for all $f \in F$. The terminology seems to be a natural and suggestive use of the word "uniform", in view of the fact that k' is uniformly determined by k, and does not depend on f. Thus if a secondary quantity is defined in terms of a single primary quantity, and if condition (*) is satisfied, then the description of the secondary quantity determines a uniform function from the primary scale to the secondary scale.

The condition of uniformity may be regarded as a commutativity condition with respect to the automorphisms of the respective scales. If $\gamma : F \rightarrow G$ is a uniform function, and $\overline{\overline{k}}$ denotes, as usual, the automorphism of F determined by $k > 0$, then the uniformity of γ is equivalent to the existence of $k' > 0$ such that the following diagram is commutative:

$$\begin{tikzcd}
F \arrow[r, two heads, "\overline{\overline{k}}"] \arrow[d, "\gamma"'] & F \arrow[d, "\gamma"] \\
G \arrow[r, two heads, "\overline{\overline{k'}}"] & G
\end{tikzcd}$$

Our definition of "uniform function" used the "scalar multiplication" in F and in G. In view of our earlier remarks concerning the relationship of ratio operations and scalar multiplication, we should not be surprised to find that there is another simple way of looking at a uniform function. If F is a ratio scale, then there is a ratio operation on F $(\rho_F^* : F \times F \rightarrow R^+)$ defined by $\rho_F^* : (f_1,f_2) \rightarrow f_1/f_2$. If γ is a uniform function from F to G, then it is easy to show that there exists a related function $\overline{\gamma} : R^+ \rightarrow R^+$ which makes the following diagram commutative:

$$\begin{tikzcd}
F \times F \arrow[r, "\rho_F^*"] \arrow[d, "\gamma \times \gamma"] & R^+ \arrow[d, "\overline{\gamma}"] \\
G \times G \arrow[r, "\rho_G^*"'] & R^+
\end{tikzcd}$$

Conversely, if such a $\overline{\gamma}$ exists, then it is easy to show that [illegible] uniform function. This suggests that we may look at a uniform function on ratio scales as a function which has the property that it takes pairs of functions in F which have equal ratios into pairs of functions in G which have equal ratios. I.e., γ preserves equality of ratios.

We shall return later to the discussion of secondary quantities which are measured by the use of more than one primary quantity, but meanwhile we pursue the study of the relationships between ratio scales.

Theorem 4-3.2. If F and G are ratio scales, and if $\gamma : F \to G$ is a uniform function, then the function $\bar{\gamma} : k \to k'$ determined by γ is a homomorphism $\bar{\gamma} : (R^+, \cdot) \to (R^+, \cdot)$. (Equivalently, since k, k', determine unique automorphisms of F and G respectively, γ determines a homomorphism $\bar{\bar{\gamma}} : A_F \to A_G$ of the respective automorphism groups of F and G.).

Proof. Let $k_1, k_2 \in R^+$, with $\bar{\gamma}(k_1) = k_1'$; $\bar{\gamma}(k_2) = k_2'$. Then, for all

$$f \in F \text{ , } \gamma(k_1 f) = k_1' \gamma(f) \text{ .}$$

Hence, for all f

$$\gamma(k_1 k_2 f) = \gamma(k_1(k_2 f)) = k_1' \gamma(k_2 f) = k_1' k_2' \gamma(f)$$

i.e.,

$$\bar{\gamma}(k_1 k_2) = k_1' k_2' = \bar{\gamma}(k_1)\, \bar{\gamma}(k_2) \text{ ,}$$

so that $\bar{\gamma}$ is a homomorphism.

To further clarify the properties of uniform functions, it is natural to appeal to properties of the homomorphisms of $(R^+, \cdot)$ which we proved in Section 2-2. But, if you refer back, you will see that we did not consider all homomorphisms, but only those which are monotone (alternatively: continuous). In the case of secondary quantities, and the uniform functions on ratio scales which are derived from the measurement of secondary quantities, it is generally assumed that, for every element in the domain of the secondary quantity, the values of the secondary quantity vary continuously with the values of each of the associated primary quantities, when the primary measure functions are varied within their ratio scales. (Bridgman assumes the stronger condition of differentiability, but it was shown, in 1946, by Martinot-Lagarde that continuity would suffice.) Translated back into our language, this implies that the secondary scale-change factor, k', is a continuous function in each of the primary scale-change factors. (Scale-change refers, of course, to a change of functions within the relevant ratio-scale.) It can be shown that this assumption, together with the homomorphism property, implies monotonicity, and conversely, that a monotone homomorphism is continuous.

We shall be interested in arbitrary monotone uniform functions, defined on ratio scales, and with values in a set of similar functions. As such a set can be uniquely imbedded in the set of all similar functions, which is a (not necessarily complete) ratio scale, we might as well assume that the value space (but not necessarily the range) is also a ratio scale. Let $\gamma : F \to G$ be a uniform function from the ratio scale F to the ratio scale G. Then, from the theorem above, γ induces a homomorphism $\bar{\gamma} : (R^+ , \cdot) \to (R^+ , \cdot)$. If in addition, the induced homomorphism $\bar{\gamma} : (R^+ , \cdot) \to (R^+ , \cdot)$ is monotone, then we say that γ is a <u>monotone uniform function</u> from F to G. [It follows from the above discussion that (assuming the continuity condition for the induced function $k \to k'$) the functions on ratio scales which are determined by the measurement of secondary quantities, are monotone uniform functions.]

The following theorem gives the fundamental property of a monotone uniform function:

<u>Theorem 4-3.3.</u> If G is a set of similar functions, F is a ratio scale, and $\gamma : F \to G$ is a monotone uniform function, then the induced homomorphism $\bar{\gamma} : (R^+ , \cdot) \to (R^+ , \cdot)$ is a power function. That is, there exists $\alpha \in R$ (uniquely determined by γ) such that $\bar{\gamma}(k) = k^{\alpha}$, for every $k \in R^+$.

<u>Proof</u>. Since γ is a monotone uniform function, $\bar{\gamma}$ is a monotone homomorphism from $(R^+ , \cdot)$ to $(R^+ , \cdot)$. Hence, (from Theorem 2-2.3) there is a unique $\alpha \in R$ such that for every $k \in R^+$, $\bar{\gamma} : k \to k^{\alpha}$. (The number α is, of course, $\log_k (\bar{\gamma}(k))$, for every $k \neq 1$.)

<u>Remarks</u>:

1. This result can be conveniently indicated by the commutative diagram:

$$\begin{tikzcd}
F \arrow[r, "\bar{\bar{k}}"] \arrow[d, "\gamma"'] & F \arrow[d, "\gamma"] \\
G \arrow[r, "\bar{\bar{k}}^{\alpha}"] & G
\end{tikzcd}$$

The result of the theorem is that, if γ is a monotone uniform function, there exists a unique α such that this diagram is commutative for every k. That is, α does not depend on k, but only on γ.

2. We have now shown that the function $\gamma : F \to G$ satisfies the condition $\gamma(kf) = k^{\alpha}\gamma(f)$, for all $f \in F$, and all $k \in R^+$. If you recall the definition of "homogeneous function" for functions of real variables you will see that it is now reasonable to call a monotone uniform function (from a ratio scale to a set of similar functions) a <u>homogeneous function</u>. This terminology (or something very much like it) is widely used by scientists in connection with the description of secondary quantities.

3. If $\alpha = 0$, then $\overline{\gamma}$ is the constant function $\overline{\gamma} : k \to 1$, and hence γ is a "constant" function which maps F onto a single element of G. (Cf. the situation for angular measure, with radian measure considered as a secondary measure.)

   If $\alpha \neq 0$, then $\overline{\gamma}$ is a monotone continuous isomorphism (i.e., a continuous automorphism) of $(R^+ , \cdot)$, and (as you may easily prove) γ is 1-1 and onto. In this case G must be a ratio scale, but it is not necessary that G be complete even if F is complete; in fact it is possible for every function in G to be a constant function.

4. Without the assumption of monotonicity (or continuity) the question of "finding" the homomorphisms from $(R^+ , \cdot)$ to $(R^+ , \cdot)$ becomes quite difficult; in fact this problem is closely connected with the questions of Lebesgue measure on R, measurable sets, and measurable functions. It can be shown that if such a homomorphism is discontinuous at any point, then it is everywhere discontinuous. [This last result is equivalent to: any endomorphism of $(R^+ , \cdot)$ which is continuous in some open interval of R^+, is continuous everywhere. This is closely related to the result of Exercise 2.12.13.]

5. The exponent α, which is uniquely determined by γ, is called the <u>degree</u>, or <u>dimension</u> of the homogeneous function γ. We denote this by $\dim \gamma$. Thus "dim" is a function $\dim : \{\gamma\} \to R$, defined on the set of all homogeneous functions between ratio scales. Whenever it is necessary to distinguish this concept of dimension from others, we refer to it as the <u>measure dimension</u> of γ. The concept of homogeneous function is, of course, definable for any $\overset{o}{R}{}^+$ - semimodule, and the notion of dimension for a homogeneous ratio scale mapping thus coincides with the general concept of the degree of a homogeneous function.

6. If G is a ratio scale, $\dim \gamma$ is often referred to as the <u>dimension of G with respect to F and γ</u>. If it is clear that a particular

homogeneous function γ is involved, this is frequently abbreviated to "the dimension of G with respect to F", or "the dimension of G in F"; this is often done when G is the scale defined by the measurement of a secondary quantity, for in this case the description of the secondary quantity specifies the functions δ and φ, and hence determines a particular γ. But, as we have seen in the case of area and length, exactly the same ratio scale (e.g., the area scale) can result from two different "secondary quantity" descriptions, and the corresponding homogeneous functions may have different dimensions. This suggests that we should use the abbreviated terminology "the dimension of G in F" with considerable caution.

Theorem 4-3.4. Let $c \in R^+$, let $\gamma : F \to G$ be a homogeneous function, and let $c\gamma$ denote the function $c\gamma : f \to c\gamma(f)$. Then

(i) $c\gamma$ is a homogeneous function;

(ii) $\dim (c\gamma) = \dim \gamma$;

(iii) if γ_1, γ_2 are homogeneous functions from F to G, with $\dim \gamma_1 = \dim \gamma_2$, then for some $c \in R^+$, $\gamma_2 = c\gamma_1$.

Proof.

(i) Given k, there exists k' such that $\gamma(kf) = k'\gamma(f)$ for every $f \in F$. Hence

$$c\gamma : kf \to c\gamma(kf) = ck'\gamma(f) = k'(c\gamma)(f)$$

so that $c\gamma$ is uniform, and determines the same correspondence $k \to k'$ as γ. That is $\overline{\gamma} = \overline{c\gamma}$, and hence $c\gamma$ is a homogeneous function.

(ii) This is immediate, since $\overline{\gamma} = \overline{c\gamma}$.

(iii) Let $\dim \gamma_1 = \dim \gamma_2 = \alpha$; and let $f_0 \in F$. Then $\gamma_1(f_0)$ and $\gamma_2(f_0)$ belong to G, and hence there exists $c \in R^+$, such that

$$\gamma_2(f_0) = c\gamma_1(f_0) .$$

We have to show that c does not depend on f_0. Let f be any other function in F. Then for some $k \in R^+$, $f = kf_0$.

Hence

$$\begin{aligned}\gamma_2(f) = \gamma_2(kf_0) &= k^\alpha \gamma_2(f_0) \\ &= k^\alpha c\gamma_1(f_0) \\ &= ck^\alpha \gamma_1(f_0) \\ &= c\gamma_1(kf_0) \\ &= c\gamma_1(f) .\end{aligned}$$

That is, $\gamma_2(f) = c\gamma_1(f)$ for every f, so that $\gamma_2 = c\gamma_1$, and the proof is complete.

Remark: This theorem tells us that a homogeneous function γ between two ratio scales F, G, is completely determined by two things:

(i) the value $\gamma(f_0)$ for any $f_0 \in F$; and

(ii) the dimension of γ.

Moreover a homogeneous function γ is determined "up to a constant factor" by its dimension. Alternatively, γ is completely determined by its value on two elements of F; for if f_1, $f_2 = kf$, are two such elements, then there exists k', such that $\gamma(f_2) = k'\gamma(f_1)$, and $\dim \gamma = \log_k k'$.

The proof of the following theorem is quite straightforward:

Theorem 4-3.5. If $\gamma_1 : F \to G$ and $\gamma_2 : G \to H$ are uniform functions, then $\gamma_2\gamma_1$ is a uniform function; if γ_1 and γ_2 are homogeneous, then so is $\gamma_2\gamma_1$, and $\dim(\gamma_2\gamma_1) = \dim \gamma_2 \dim \gamma_1$.

The Measurement of Secondary Quantities. Now that we have established some of the basic properties of homogeneous functions, we can easily discover the nature of those functions φ on R^+ which may be used in the "rules" which define secondary measure functions, if these rules are to lead to homogeneous functions on ratio scales. To keep things simple, let us look first at the case where we have a domain B of objects, and associated with each $b \in B$ there is a unique object $a = \delta(b)$ in the domain A of a (primary) ratio scale, F. Then a "rule" for the determination of a measure function for a secondary quantity, is a combination of the function δ, the selection of a function $f \in F$, and a function $\varphi : R^+ \to R^+$. Corresponding to each $f \in F$ we get a secondary measure function g, such that $g(b) = \varphi(f(\delta(b)))$. This is easily pictured from the diagrams

$$\begin{tikzcd}
A \arrow[r, "f"] & R^+ \arrow[d, "\varphi"] \\
B \arrow[u, "\delta"] \arrow[r, dashed, "g"] & R^+
\end{tikzcd}
\qquad
\begin{tikzcd}
a \arrow[r, "f"] & f(a) = x \arrow[d, "\varphi"] \\
b \arrow[u, "\delta"] \arrow[r, dashed] & g(b) = \varphi(x)
\end{tikzcd}$$

(Because g is defined by composition, the first diagram is trivially commutative.) Our objective is to find those functions φ which will make the function $f \to \varphi f \delta$ a homogeneous function on F.

Let us assume that the functions δ and φ determine a ratio scale G, and a homogeneous function $\gamma : F \to G$. Then there exists $\alpha \in R$, such that for every $b \in B$, every $f \in F$, and every $k \in R^+$

$$(\gamma(kf))(b) = k^\alpha[(\gamma(f)(b)] .$$

Hence

$$\begin{aligned}(\varphi(kf)\delta)(b) &= \varphi(k(f(\delta(b)))) \\ &= k^\alpha \varphi(f(\delta(b))) .\end{aligned}$$

That is, for all $x \in f(\delta(B))$,

$$\varphi(kx) = k^\alpha \varphi(x) .$$

This means that φ is a homogeneous function in the usual sense. We assume that $B \neq \emptyset$; this implies that $\{x : \text{for at least one } f \in F, x \in f(\delta(b))\} = R^+$. Hence φ must be defined on all of R^+; and, putting $x = 1$, φ must satisfy

$$\begin{aligned}\varphi(k) &= k^\alpha \varphi(1) , \text{ for every } k \in R^+ . \\ &= ck^\alpha \text{ (say) }, \text{ where } c = \varphi(1) .\end{aligned}$$

Conversely, it is easy to verify that if (for any $c \in R^+$, and any $\alpha \in R$) we define $\varphi : R^+ \to R^+$ by $\varphi : x \to cx^\alpha$, then the functions $\varphi f \delta$ are all similar, and therefore determine a ratio scale G with domain B;

and the resulting mapping from F to G is a homogeneous function with dimension α. We recall that φ is a homogeneous function (in the usual sense) with degree α, so that φ and γ are both homogeneous, and both have the same degree. That is, $\dim \gamma = \deg \varphi$.

We incorporate this result in

Theorem 4-3.6.

(i) Let $F = \{f\}$, be a (primary) ratio scale, with domain A; let δ be a mapping from a set $B\ (\neq \emptyset)$ to A; and let φ be a function from R^+ to R^+. Then the composite functions $g : B \to R^+$ (defined by $g = \varphi f \delta$, $f \in F$) are all similar and the induced function $f \to \varphi f \delta$ is homogeneous, if, and only if, for some fixed $\alpha \in R$ and some fixed $c \in R^+$, $\varphi : x \to cx^{\alpha}$.

(ii) The function $\varphi : R^+ \to R^+$ defined by $\varphi : x \to cx^{\alpha}$ is homogeneous of degree α; and if $\gamma : F \to G$ is the corresponding homogeneous function on F, $\dim \gamma = \deg \varphi$.

Remarks:

1. If we use Bridgman's language, and refer to the "quantity" measured by F as a "primary quantity", and to the "quantity" measured by the "g"-functions as a "secondary quantity", then α is called the dimension of the secondary quantity with respect to the primary quantity.

2. For fixed F, B, δ, and α, the secondary measure functions determined by $\varphi : x \to cx^{\alpha}$ (for different positive c) all belong to the same (secondary) ratio scale. If $\alpha \neq 0$, this secondary scale is the set of functions $\{\varphi f \delta : f \in F, \varphi : x \to x^{\alpha}\}$; if $\alpha = 0$, there is only one function in the latter set, and this function "generates" the secondary scale.

It is important to keep in mind that the determination of the measure functions $\{g\}$ depends on δ as well as on φ. Thus if different functions "δ" are used, we may get quite different sets of secondary measure functions on B. Whether or not these sets are different, they must be regarded as derived from different "secondary quantities". For a fixed B, by using different "δ" and "φ" we may obtain the same set of secondary measure functions in different ways. These must be regarded as being determined by different "secondary quantities". I.e., a "secondary quantity" is more than just a set of similar functions, or a ratio scale; the concept of "secondary quantity" also includes the rules by means of which the secondary ratio scale is obtained. Thus our set Δ of area functions (on the domain of polygonal regions, say) is not itself a "secondary quantity", but there are many (in fact infinitely many) "secondary quantities" which determine the same set of area functions. If these all had the same dimension (2) we could say that "area is a secondary quantity of dimension 2 in length". But, as we have seen, there are also homogeneous functions of dimension 1 which relate the ratio scales for length and area, and these homogeneous functions are associated with secondary measure functions.

This difficulty with area functions is typical of the confusion which arises if one attempts to associate a dimension with a set of secondary measure functions, instead of with the homogeneous function (either on the scale, or on R^+), which is determined by the procedure for "measuring" the secondary quantity.

The construction of the η_σ functions (see Section 3-5) relating the length and area scales, has its counterpart in connection with the relationship of the length and volume scales. Recall the procedure which we described for the empirical measurement of volumes of liquids, and for the measurement of the volumes of solids by displacement of a liquid, in a container of constant cross-section. This established a homogeneous function of dimension 1 from the length scale to the volume scale. You might regard the length/area example as rather "pathological", but the length/volume situation is certainly quite natural. In fact a great deal of volume measurement is carried out in just this way, with the additional feature that the resulting "linear scale" is "calibrated" by naming the marks on the scale in terms of the corresponding "cubic" measure function.

Our purpose in giving these examples is to emphasize that there are implicit conventions in everyday usage. When it is stated that "area has dimension 2 in length", it is implied that a "dimension 2" correspondence has been agreed upon. Usually this will be the homogeneous function η' of Chapter 3, but it could be any homogeneous function which differs from η' by a positive constant factor. (Recall that, from Theorem 4-3.4, these are the only homogeneous functions of dimension 2 from the length scale to the area scale.)

The Definition of a Secondary Quantity. You have possibly noticed that we have not yet given a genuine definition of "secondary quantity". The better scientific treatments of measurement and dimension insist that a secondary quantity (in an empirical sense) cannot be considered except in the context of a specific measurement procedure, and that even though two such procedures might lead to exactly the same secondary ratio scale, they must be considered as determining different secondary quantities. We certainly agree with this point of view, because without it all discussion of "the dimension of a secondary quantity" would become meaningless.

It follows that, whatever definition we adopt for "secondary quantity", the definition must preclude such vague statements as "area is a secondary quantity", and it must also preclude such statements as "the area scale (which is well defined after we specify the domain) is a secondary quantity".

The above discussion suggests that, in the case where a single (primary) scale is involved, the most useful procedure would be to define a simple secondary quantity to be a set $\{F , B , \delta ; \varphi \}$ (where F is a ratio scale with domain A , δ is a function $B \to A$, and φ is a function $R^+ \to R^+$) such that .

SQ-1. The functions $\varphi f \delta$ are all similar, and the function $\gamma : f \to \varphi f \delta$ is homogeneous.

As we have seen above, SQ-1 is equivalent to either of the conditions:

SQ-2. φ is homogeneous;

or

SQ-3. There exists $c > 0$, $\alpha \in R$, such that $\varphi : x \to cx^\alpha$;

so we may use any one of these three conditions in the definition. Thus defined, the dimension of the simple secondary quantity may be defined to be the degree of γ , or, equivalently, the degree α of φ .

We return later to the question of an appropriate definition for a secondary quantity derived from more than one (primary) ratio scale.

Secondary Quantities Defined by Using Several Primary Quantities. A secondary quantity is said to be determined by several primary quantities if (but not necessarily only if) the following conditions are satisfied:

(i) There is a finite set of primary quantities, corresponding to the ratio scales F_i $(i = 1, 2, \ldots, n)$, with domains A_i.

(ii) There is a mapping δ from the domain B of the secondary quantity into the cartesian product $A = A_1 \times A_2 \times \ldots \times A_n$ of the domains of the primary quantities. (Intuitively, each domain element b determines a set $(a_1, a_2, \ldots, a_n)$ of elements, one from each domain A_i.)

(iii) There is a function $\varphi : R_1^+ \times R_2^+ \times \ldots \times R_n^+ \to R^+$ (where each R_i^+ is equal to R^+), from which a secondary measure function g for B is calculated by the following "rule": let $f_1, f_2, \ldots, f_n$ be any set of measure functions from the respective primary scales, and let $b \in B$. If $\delta(b) = (a_1, a_2, \ldots, a_n)$ then $g : B \to R^+$ is defined by $g(b) = \varphi(f(a_1), f(a_2), \ldots, f(a_n))$. This may be pictured from the commutative diagram

$$\begin{tikzcd}
A \arrow[r, "\prod f_i"] & \prod R_i^+ \arrow[d, "\varphi"] \\
B \arrow[u, "\delta"] \arrow[r, dashed, "g"'] & R^+
\end{tikzcd}$$

where g is defined by composition (so that the diagram is trivially commutative) and $\prod f_i = f_1 \times f_2 \times \ldots \times f_n$, $\prod R_i^+ = R_1^+ \times R_2^+ \times \ldots \times R_n^+$.

(iv) For each choice of functions f_i, there will be a unique (but not necessarily different) "g" function. It is assumed that all such "g" functions are similar, and hence they belong to (in fact, determine) the same ratio scale. (The set need not be the whole scale: cf. radian measure; and the scale need not be complete.)

(v) The composite function $(\prod f_i)\delta$ is onto $\prod R_i^+$ for each set of functions $f_1, f_2, \dots, f_n$.

[Remark: As in the case of Theorem 4-3.1, condition (v) could be weakened, but we leave it in this form in order to simplify the following discussion. If you follow the discussion carefully to see where we make use of this condition and its consequences, you will see how it could be weakened.]

Let $\bar{x}_j \in R_j^+$ be fixed $(j \neq i)$, let $\prod f_i \in \prod F_i$ and define $B_i = \{b : (\prod f_i)(\delta(b)) = (\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_n), x_i \in R_i^+\}$. Then B_i depends on the choice of the $\bar{x}_j$, and on the functions f_j $(j \neq i)$, but with these fixed, B_i is the same for all $f_i \in F_i$. Let $h_i : x_i \to (\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_n)$. Then the functions h_i depend on the choice of the $\bar{x}_j$ $(j \neq i)$. For each choice of $\bar{x}_j$ $(j \neq i)$, and each choice of f_j $(j \neq i)$ the diagram below is commutative for every $f_i \in F_i$:

$$\begin{tikzcd}
A_i \arrow[r, "f_i"] & R_i^+ \arrow[d, "h_i"] \\
\delta(B_i) \arrow[u, "p_i"] \arrow[r, "\prod f_i"] & \prod R_i^+ \arrow[d, "\varphi"] \\
B_i \arrow[u, "\delta"] \arrow[r, "g_i"] & R_i^+
\end{tikzcd}$$

(p_i denotes the "projection", $p_i : (a_1, \dots, a_i, \dots, a_n) \to a_i$.)

Condition (v) implies that $(f_i\, p_i\, \delta)(B_i) = R_i^+$. For each choice of the $\bar{x}_j$ and the f_j, $(j \neq i)$, and for each i, there is a function $g : B \to R^+$ such that the resulting mapping g_i is equal to $g|B_i$. This function g is uniquely determined by g_i. (Recall that the "g" functions are similar, so that two of them must be the same if they agree on a single element of B.)

Hence if we keep $\bar{x}_j$ and f_j fixed $(j \neq i)$, Theorem 4-3.1 applies to the situation represented by the following diagram:

$$\begin{tikzcd}
A_i \arrow[r, "f_i"] & R_i^+ \arrow[d, "\varphi_i = \varphi h_i"] \\
B_i \arrow[u, "\delta_i = p_i\delta"] \arrow[r, "g_i"'] & R^+
\end{tikzcd}$$

This determines a uniform function $F_i \to G_i$ (where $G_i = \{g_i\}$) and hence a uniform function $\gamma_i : F_i \to G$, which might depend on the choices made for $\bar{x}_j$ and f_j $(j \neq i)$.

We can now use Theorem 4-3.6 (which dealt with the nature of φ for the case of a secondary quantity determined by a single quantity) to discover the nature of φ for the general case represented by the above discussion.

<u>Theorem 4-3.7</u>. Let $\varphi : R_1^+ \times R_2^+ \times \ldots \times R_n^+ \to R^+$ be a function used in the determination of a secondary quantity in accordance with conditions (i)-(v) above, and assume that for every i, and for every choice of the $\bar{\kappa}_j$ and $f_j (j \neq i)$, the resulting uniform function γ_i is monotone. Then there exist unique real numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$, and a unique positive real constant c, such that

$$\varphi : (x_1, x_2, \ldots, x_n) \to c x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n} .$$

<u>Proof</u>. With the notation of the above diagram, and with the assumptions of the theorem, the uniform functions γ_i are homogeneous. Hence Theorem 4-3.6 applies, so that, for some real number α_i, and some positive constant $\bar{c}_i$ (both possibly depending on the fixed $\bar{x}_j$ and the fixed functions f_j)

$$\varphi_i : x_i \to \bar{c}_i x_i^{\alpha_i}$$

That is,

$$\varphi: (\bar{x}_1 , \ldots , \bar{x}_{i-1} , x_i , \bar{x}_{i+1} , \ldots , \bar{x}_n) \to \bar{c}_i \, x_i^{\alpha_i} .$$

Hence, for $i = 1$ and $x_2 = x_3 = \ldots = x_n = 1$, there exist $c > o$ and $\alpha_1 \in R$ such that, for all $x_1 \in R_1^+$,

$$\varphi : (x_1 , 1 , \ldots , 1) \to c x_1^{\alpha_1} .$$

Keeping x_1 fixed at $\bar{x}_1$, and with $x_3 = x_4 = \ldots = x_n = 1$, a repetition of the argument implies that there exist $\alpha_2 \in R$ and $c_2 > 0$, such that

$\varphi : (\bar{x}_1 , x_2, 1, .., 1) \to c_2 x_2^{\alpha_2}$, for all $x_2 \in R_2^+$; (α_2 and the constant c_2 may depend on $\bar{x}_1$). But, when $x_2 = 1$, the right hand side $(c_2 x_2^{\alpha_2} = c_2)$ must be $c\bar{x}_1^{\alpha_1}$. Hence $c_2 = c\bar{x}_1^{\alpha_1}$, and hence

$$\varphi : (x_1 , x_2 , 1 , \ldots , 1) \to c x_1^{\alpha_1} x_2^{\alpha_2}$$

for all $x_1 \in R_1^+ , x_2 \in R_2^+$. Repeating this process, we obtain real numbers $\alpha_1 , \alpha_2 , \ldots , \alpha_n$, such that

$$\varphi: (x_1 , x_2 , \ldots , x_n) \to c x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n} .$$

Clearly $c = \varphi(1 , 1 , \ldots , 1)$ is unique. To prove the uniqueness of the numbers α_i , assume that $\varphi(x_1 , x_2 , \ldots , x_n) = c x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n} = c x_1^{\alpha'_1} x_2^{\alpha'_2} \ldots x_n^{\alpha'_n}$. Then (with $x_j = 1 , j \neq i$) we get $x_i^{\alpha_i} = x_i^{\alpha'_i}$ for each i . Hence $x_i^{\alpha_i - \alpha'_i} = 1$, so that $\alpha_i = \alpha'_i$, for each i . Hence the expression $c x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$ is unique, and the numbers α_i , and the uniform functions γ_i whose dimensions are α_i , do not, in fact, depend on the choices made during the proof.

Remarks:

1. The functions $\varphi: (x_1 , x_2 , \ldots , x_n) \to c x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$ are homogeneous in the standard sense (with degree $\Sigma \, \alpha_i$) but they are not

the only such homogeneous functions. They also satisfy the stronger "homogeneity" condition

$$\varphi : (k_1x_1 , k_2x_2 , \ldots , k_nx_n) \rightarrow k_1^{\alpha_1} k_2^{\alpha_2} \ldots k_n^{\alpha_n}[\varphi(x_1 , x_2 , \ldots , x_n)] .$$

That is, these functions φ are multihomogeneous functions.

You can easily verify that these functions are the only functions $\varphi : R_1^+ \times R_2^+ \times \ldots \times R_n^+ \rightarrow R^+$ which are multihomogeneous.

2. For fixed B , δ , $\{F_i\}$, and $\{\alpha_i\}$, the same secondary scale is generated by the functions $\varphi(\prod f_i)\delta$ (where $\varphi : (x_1 , x_2 , \ldots , x_n) \rightarrow cx_1^{\alpha_1}x_2^{\alpha_2} \ldots x_n^{\alpha_n}$) for every $c > 0$; i.e., the scale is determined by B , δ , the primary scales, the α_i , and any particular value of c : e.g., $c = 1$. The corresponding function, φ , can be "factored" as the composite of the cartesian product of the separate homogeneous functions $\varphi_i : x_i \rightarrow x_i^{\alpha_i}$, and the ordinary real number multiplication function (μ) on the cartesian product, as indicated in the commutative diagram:

$$\begin{tikzcd}
(x_1 , x_2 , \ldots , x_n) \arrow[r, "\prod \varphi_i"] \arrow[dr, "\varphi"'] & (x_1^{\alpha_1} , x_2^{\alpha_2} , \ldots , x_n^{\alpha_n}) \arrow[d, "\mu"] \\
 & x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}
\end{tikzcd}$$

If the scales F_i are all different, then we say that α_i is the <u>dimension of the secondary quantity</u> in the <u>primary quantity</u> F_i . If the primary scales are not all distinct, then the dimension of the secondary quantity in a primary quantity F is <u>defined</u> to be the sum of the exponents α_i which correspond to those F_i which are equal to F . (The reason for this latter definition will become more clear as we proceed. See Theorem 4-3.9 below.) Meanwhile you should recall that the establishment of area functions (as secondary quantities) for rectangular regions can be considered in this latter category. For if b is a rectangular region, with sides (segments) a_1 , a_2 then $\delta b = (a_1,a_2)$. And if f_1 , f_2 are any two length functions from the ratio scale F of length functions for segments, the function $g : b \to f_1(a_1) \cdot f_2(a_2)$ is an area function for rectangles. Moreover all of the above requirements for the definition of a secondary quantity are satisfied (you should check this) so that we obtain the area scale for rectangles from a secondary quantity of dimension $1 + 1 = 2$ in length. (But, as we have already remarked, the same area scale can also be obtained from a secondary quantity of dimension 1 in length; so we do not refer to the area scale itself as having a well-defined dimension in length.)

Remark: Perhaps you were surprised that, in the above example regarding the area of rectangular regions as a secondary measure, we selected different functions f_1, $f_2 \in F$. There was a very good reason for this: condition (iv) in the definition of a secondary quantity from a set of primary quantities, required that all g functions (for separate choices of the f_i in each factor F_i of the cartesian product) must be similar. Hence the hypotheses of Theorem 4-3.7 would not necessarily be satisfied (in the case $n = 2$, $F_1 = F_2 = F$) if you were to consider only those "g" functions

$$g : b \to f(a_1) \cdot f(a_2)$$

for which the same f has been used in each "factor". If you retrace the proof of Theorem 4-3.7 with this point in mind, you will see where we needed condition (iv). This does not mean that the set of functions $g : B \to R^+$, defined by

$$g : b \to f(a_1) \cdot f(a_2)$$

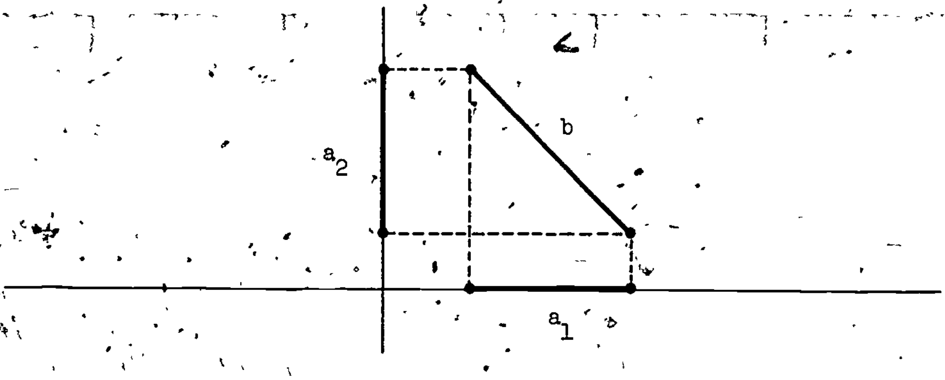
does not, in some sense, correspond to a secondary quantity of dimension 2 in length. Of course it does, but (at present) we can only conclude that this is so because they are part of the (possibly larger) set of functions, determined by $g : b \to f_1(a_1) \cdot f_2(a_2)$, to which Theorem 4-3.7 applies.

The source of this difficulty lies in the fact that our treatment of secondary quantities was primarily designed to apply to the case where all primary quantities were distinct, and we only considered the case of non-distinct primary quantities as a special case. But, in practice, it might well happen that several elements of the domain A of a single primary quantity are associated with an element of the domain of a secondary quantity, and, in the definition of a secondary measure function, we will usually wish to use the same primary function for each of these primary domain elements. This leads (see below) to a simple extension of the notion of secondary quantity, in which our choice of functions $f_i \in F_i$ is more restricted, but the choice of functions φ is less restricted. [You will recall that, in considering area functions for square regions, we were able to avoid this problem because of the congruence of the sides of a square, so that we could consider the secondary measure function for square regions without using the cartesian product of the domains of length functions. I.e., we obtained the area functions for square regions from a secondary quantity which was defined in terms of a single primary quantity, with a single domain element associated with each domain element of the secondary quantity.]

To clarify this question, we look at an example which, at first sight, appears to contradict the result of Theorem 4-3.7. (This illustrates the importance of checking hypotheses.) Most books either ignore this question of what to do when there is more than one associated domain element from a primary scale, or they imply that it can always be handled (as we were able to handle area for rectangles) by the use of cartesian products. The following simple example illustrates that this might not always be the case.

Example. Suppose that we have a fixed plane, with a coordinate system, and that b is any segment in the plane. Let a_1, a_2, be the projections of b on the coordinate axes. (It is possible that one of these might be a single point. If so we give it length "0" under every length function. As we saw in Chapter 2, this was a natural (and in a sense, the only reasonable) way of extending the domain to include single points. This minor addition is not significant in our example.) If $F = \{f\}$ is the set of length functions for segments, let us define a new set of functions

$$G = \{g_f : g_f(b) = \sqrt{[f(a_1)]^2 + [f(a_2)]^2}\ ,\ f \in F\}\ :$$



Then, from our knowledge of geometry, we know that the set G is just the set of length functions for all segments in the plane. In particular, G is certainly a ratio scale.

Let us put this in the context of the definition of a secondary quantity, and attempt to handle it by using cartesian products, where $\delta(b) = (a_1,a_2)$. The following diagram will help you to visualize this:

$$\begin{tikzcd}
A_1 \times A_2 \arrow[r, "f \times f"] & R_1^+ \times R_2^+ \arrow[d, "\varphi"] \\
B \arrow[u, "\delta"] \arrow[r, "g"'] & R^+
\end{tikzcd}$$

A_1 and A_2 are the sets of segments on the respective coordinate axes, and the set of all length functions on each of these sets is, of course, a ratio scale. (These scales are obtained by restricting the domain of F. We use the same symbols, f , for these restricted functions.) The function φ is defined by

$$\varphi : (x_1,x_2) \rightarrow \sqrt{x_1^2 + x_2^2} .$$

Here we have a perfectly good secondary scale G , defined by a procedure very similar to that used in the definition of a secondary measure in terms of several primary measures. But the function Φ is certainly not in the form which Theorem 4-3.7 leads us to expect, so we know that something must be "wrong". Careful examination shows that what is missing is that the <u>all</u> in condition (iv) of the definition of a secondary measure has been ignored. In order to be a secondary measure (as previously described) in F_1 and F_2 (the length scales on the respective axes) the set $\overline{G}$ of <u>all</u> functions

$$\overline{g} : b \rightarrow (\varphi(f_1 \times f_2)\delta)(b)$$

must be similar, where the f_1 and f_2 need not be the "same" function. (By "same", we mean that they are the restrictions to A_1 , A_2 , of the same length function from F.) It is not hard to show directly that the set $\overline{G}$ cannot belong to a ratio scale (e.g., find two elements b_1 , b_2 , which have the same value under one $\overline{g}$-function , but not under another) a fact which we could also infer indirectly if we were certain that all of the remaining conditions of Theorem 4-3.7 were satisfied.

This leaves us with the problem of what to do in this case. After all, the set of functions G is a perfectly good ratio scale, and we feel that it should be derivable from an appropriate definition of a secondary quantity. In addition, we probably feel that its dimension in "length" should be "1".

Secondary Quantities Derived From Several Arguments in a Single Primary Scale. To keep things reasonably simple, let us consider the situation where we have a single primary ratio scale F , which leads (as indicated in the diagram below) to a set of similar functions g , which determine a ratio scale G . (The domain of F is A , each $A_i = A$, and, of course, each $R_i^+ = R^+$.)

$$\begin{tikzcd}
A_1 \times A_2 \times \cdots \times A_n \arrow[r, "\prod f"] & \prod R_i^+ \arrow[d, "\varphi"] \\
B \arrow[u, "\delta"] \arrow[r, dashed, "g"'] & R^+
\end{tikzcd}$$

Each function g is defined by $g : b \to (\varphi(\prod f)\delta)(b)$. In other words, an ordered set $(a_1 , a_2 , \ldots , a_n)$ of elements of A is associated with each element of B , the values $f(a_i) = x_i$ are "measured", using the same $f \in F$, and then $\varphi(x_1 , x_2 , \ldots , x_n)$ is "calculated". We are interested in discovering what must be the nature of the functions φ , in order that G should be a ratio scale, with the related function $(f \to g = \varphi(\prod f)\delta)$ homogeneous. This question is answered in the following theorem:

Theorem 4-3.8. With the above notation, let $\gamma : f \to \varphi(\prod f)\delta$, and assume that $(\prod f)\delta$ is onto. (As before, this condition may be weakened.) If the functions $\gamma(f)$ are similar, and if γ is a homogeneous mapping, then φ is a homogeneous function, $\prod R_i^+ \to R^+$ (in the standard sense). Conversely, if φ is homogeneous, then (without the assumption that $(\prod f)\delta$ is onto), the functions $\gamma(f)$ are similar, and γ is homogeneous.

Proof. If the functions $\gamma(f)$ are similar and if γ is homogeneous, then, for each $k \in R^+$, there exists $\alpha \in R$, such that $\gamma(kf) = k^{\alpha}\gamma(f)$. Hence

$\varphi(kx_1, kx_2, \dots, kx_n) = k^{\alpha}\varphi(x_1, x_2, \dots, x_n)$, so that φ is homogeneous of degree α. The proof of the converse is quite straightforward.

Remarks:

1. If we apply this result to our simple example on segment length in the plane, we see that the function $\varphi : (x_1, x_2) \to \sqrt{x_1^2 + x_2^2}$ satisfies the condition of homogeneity, and that its degree is 1. This function is not, however, bihomogeneous, as you may easily verify.

2. We are now in a position to give a general definition of "secondary quantity". A secondary quantity is a set $\{\{F_i\}, B, \delta, \varphi\}$, where $\{F_i\}$ is a finite set of ratio scales (not necessarily distinct); δ is a function from B to a cartesian product of the domains of the scales F_i, with the domain of each factor F_i repeated j_i times; and φ is a function, on the "corresponding" cartesian product

$$\varphi : R^+_{11} \times \dots \times R^+_{1j_1} \times R^+_{21} \dots \times R^+_{2j_2} \times \dots \times R^+_{n1} \times \dots \times R^+_{nj_n} \to R^+$$

such that

$$\varphi : (x_{11}, \dots, x_{1j_1}, x_{21}, \dots, x_{2j_2}, \dots, x_{n1}, \dots, x_{nj_n})$$

$$\to c\,\varphi_1(x_{11}, \dots, x_{1j_1})\,\varphi_2(x_{21}, \dots, x_{2j_2}) \dots \varphi_n(x_{n1}, \dots, x_{nj_n}),$$

where the φ_i are homogeneous (in the ordinary sense) with degrees $\alpha_1, \alpha_2, \dots, \alpha_n$, respectively.

With this general definition of secondary quantity, we can give an unambiguous definition of dimension. The dimension of the secondary quantity in the primary scale F, is the sum of the degrees α_i for those φ_i whose corresponding $F_i = F$. You should check that this definition agrees with the definitions given earlier for special cases.

3. If you apply Theorem 4-3.8 to the case of angular measure in relation to linear measure, as in the ~~construct~~ion of the radian measure function, you will see that this construction gives a secondary quantity of dimension 0 in length. The function φ was given by $\varphi : (x_1, x_2) \to (x_1/x_2)$; this is homogeneous of degree 0. It is also bihomogeneous,

of degree $(1,-1)$: this means that if we were to use the radian measure construction, but with length functions for arcs and segments which did not have to be the "same", then the set of all secondary measure functions obtained would also belong to a ratio scale. This set contains the radian measure function Φ ; it is, of course, the full ratio scale which the radian measure function determines.

4. If, in the development of angular measure, we had used a fixed circle, and set up angular measure functions in terms of the arc lengths intercepted on this fixed circle (or a congruent circle), then the resulting angular measure scale (as a whole) would have been the same, but the associated homogeneous function from the length scale to the angular measure scale would then have dimension 1 . (This is quite comparable to our earlier examples showing that homogeneous functions of different dimension could be established from the length scale to the area scale; and also from the length scale to the volume scale.)

5. When there are two ways of looking at the same dimension question by considering it either in terms of several arguments in a single scale and (standard) homogeneous functions, or in terms of separate but equal scales with a single argument in each scale and multihomogeneous functions, then we would expect that the different definitions of dimension should agree. This is, in fact, the case, and this was our motivation for the definition of the dimension of a secondary quantity which we gave (at the end of Theorem 4-3.7) for the case of the F_i not all different. (Recall that we defined the dimension of the secondary quantity in F to be the sum of the α_i which correspond to those scales F_i which were the same as F .) To avoid complicated notations, we prove the simplest case of this result; the general case presents no additional difficulties.

Theorem 4-3.9. If G is a secondary measure scale defined in terms of the primary scales F_1 , F_2 , $(F_1 = F_2 = F)$ as in Theorem 4-3.7, then there is an associated function $\varphi : (x_1, x_2) \to cx_1^{\alpha_1}x_2^{\alpha_2}$. This function is homogeneous (in the standard sense) in (x_1, x_2) , and has degree $\alpha_1 + \alpha_2$; and it is multihomogeneous of degree (α_1, α_2) . Hence the two definitions of dimension agree.

Proof.
$$\begin{aligned}\varphi(kx_1, kx_2) &= c(kx_1)^{\alpha_1}(kx_2)^{\alpha_2}\\ &= ck^{\alpha_1+\alpha_2}\, x_1^{\alpha_1}\, x_2^{\alpha_2}\\ &= k^{\alpha_1+\alpha_2}\cdot \varphi(x_1, x_2)\,.\end{aligned}$$

Hence φ is homogeneous in the standard sense, and has degree $\alpha_1 + \alpha_2$. Similarly show that φ is multihomogeneous of degree (α_1, α_2), so that the two definitions of dimension for the dimension of the secondary quantity in F, are in agreement.

While we have certainly not exhausted even the elementary study of dimension in relation to the question of primary and secondary quantities, we have probably covered most of the basic ideas, with particular emphasis on some of the simple but fundamental ideas which are usually passed over very lightly. We use these ideas as motivation for the next section, in which we study the genesis, and the inter-relationships, of ratio scales, from a more mathematical standpoint.

4-4 A Mathematical Theory of Ratio Scales

In earlier sections we have studied a number of measurement situations, both mathematical and empirical, in which the concept of a ratio scale arose quite naturally: in a variety of situations we found that the set of all measure functions satisfying certain mathematical, or empirically suggested, conditions, satisfied the requirements of a ratio scale. In the last section we studied the empirically motivated notions of primary and secondary quantity, and we saw that the requirements for the definition of a secondary quantity were such that the concept of a secondary quantity implied

(a) the existence of a "secondary" ratio scale;

(b) the existence of homogeneous functions from the primary scales to the secondary scale.

In this section we will study ratio scales, and homogeneous functions on ratio scales, in a more general mathematical context, drawing together and adding to the results of the last section.

As only a small handful of ratio scales are encountered in practice, you might think that the development of a general mathematical theory (involving infinitely many scales) is a waste of time. But it is necessary to do this if we are to go beyond the study of special cases. The situation is rather like that for the scales themselves: in practice, we only use a very small number of particular length functions, but a theory of length must concern itself with the whole ratio scale of (infinitely many) length functions.

In the development of a mathematical theory, the concepts of primary and secondary scale will virtually lose their significance: in relation to a homogeneous function their roles are rather like that of domain and range, and the same scale can be the domain of one homogeneous function and the range of another. Moreover, all homogeneous scale functions except those of dimension zero, are 1-1, and therefore have inverses. For this reason we shall have little use for the terms "primary quantity" and "secondary quantity", but we shall certainly use the ideas which they have suggested.

Any reader who is familiar with the unifying notions of category, morphism, and functor, will notice familiar ideas, but no attempt has been made to develop a theory of ratio scales and dimension in the most abstract form.

The Generation of Ratio Scales. As we observed earlier, if A is any set, and f any function from A to R^+, then there is a unique ratio scale "generated" by or containing, f. This tells us two things: that the class of all ratio scales is very numerous; and that a ratio scale is fully determined by any one of the functions which it contains. (This corresponds to the fact that similarity is an equivalence relation on the set of functions from A to R^+, and to the fact that each equivalence class of similar functions is a ratio scale.) We wish to consider various ways in which "new" ratio scales can be "generated" from existing scales. At the same time, we would like the procedure for generating the new scale to suggest (in a natural way) homogeneous mappings between the old scales and the new. Among the various procedures which may be used are the following three, which were suggested by the discussion in the last section, and which may be combined in a wide variety of ways:

(i) Post-composition. If $F = \{f : f{:}A \to R^+\}$ is a ratio scale, we may compose each mapping in the scale with the same function $\varphi : R^+ \to R^+$. For a suitable φ, this gives us a new scale $\{\varphi f : f \in F\}$, with the same domain as F.

(ii) Pre-composition. With F as in (i), if B is any set, and if δ is any mapping $\delta : B \to A$, we obtain a new scale $\{f\delta : f \in F\}$ with domain B. A simple but important case of this occurs when $B \subset A$, and $\delta : b \to b$; the new scale is then the set of restricted functions $f|B$. (This situation is encountered in "extension of the domain" considerations, where we have the functions $f|B$, and seek to extend them to a larger domain A, so as to obtain an "extended" scale.)

(iii) Products. If $F_i = \{f_i\}$, $i=1, 2, \ldots, n$, are ratio scales, we can form a "product scale", which we will denote by $F_1 \underline{\times} F_2 \underline{\times} \ldots \underline{\times} F_n$. (We use the underlined symbol to distinguish the product scale from the cartesian product of the sets F_i.)

Properties of Homogeneous Functions. Before looking at the above three methods in detail, we summarize the main properties of homogeneous functions on ratio scales.

(i) A function $\gamma : F \to G$ from a ratio scale F to a ratio scale G is a homogeneous function, if and only if there exists $\alpha \in R$, such that for every $f \in F$ and $k > 0$,

$$\gamma(kf) = k^{\alpha}\gamma(f) .$$

(ii) If $\alpha = 0$, then γ is a "constant" function.

(iii) If $\alpha \neq 0$, γ is 1-1, and its inverse (γ^{-1}) is also a homogeneous function, with dimension $\frac{1}{\alpha}$. That is

$$\dim \gamma^{-1} = \frac{1}{\dim \gamma} .$$

(iv) If $\gamma_1 : F \to G$, and $\gamma_2 : G \to H$ are homogeneous functions, then $\gamma_2\gamma_1$ is a homogeneous function, and

$$\dim (\gamma_2\gamma_1) = (\dim \gamma_2) (\dim \gamma_1) .$$

(v) An automorphism $\bar{\bar{k}} : F \to F$, $(\bar{\bar{k}} : f \to kf)$ is, of course, a homogeneous function, and $\dim \bar{\bar{k}} = 1$.

(vi) If $\gamma_1, \gamma_2, \ldots, \gamma_n$, and $\gamma'_1, \gamma'_2, \ldots, \gamma'_n$, are homogeneous functions such that

$$\gamma_n \gamma_{n-1} \cdots \gamma_1 = \gamma'_m \gamma'_{m-1} \cdots \gamma'_1$$ (functional composition)

then the composite functions are homogeneous, and

$$\prod_{i=1}^{n} \dim \gamma_i = \prod_{j=1}^{m} \dim \gamma'_j .$$

(vii) With a notation similar to (vi), if two composite homogeneous functions $\gamma_n \gamma_{n-1} \cdots \gamma_1$ and $\gamma'_m \gamma'_{m-1} \cdots \gamma'_1$ have the same domain scale, and the same image scale, and the same dimension, then they differ only by an automorphism $\bar{\bar{c}}$ of the range scale. (We say that such functions are the same "up to a constant factor".)

Formation of New Scales by Post-Composition. Let F be a set of functions with domain A and range in R^+, and let $\varphi : R^+ \to R^+$. Then if f $(\in F)$ and φ are composed, we not only obtain a new function $\varphi f : A \to R^+$, but we also obtain, in a natural way, a mapping $\varphi_* : F \to F'$, where $F' = \{\varphi f : f \in F\}$, and $\varphi_* : f \to \varphi f$. The question we wish to answer is: what functions φ have the properties that, if F is a ratio scale, then F' is also a ratio scale and φ_* is a homogeneous function? This was implicitly answered in the last section, but we answer it explicitly in the following theorem:

Theorem 4-4.1.

(i) With the above notation, F' is a ratio scale and φ_* is a homogeneous function, if and only if $\varphi : x \to cx^{\alpha}$, $(c > 0 , \alpha \neq 0)$; and, in this case, $\dim \varphi_* = \alpha$.

(ii) $F' = F$ if and only if $\alpha = 1$; in this case φ_* is simply an automorphism of F .

Proof.

(i) If $c > 0$, $\alpha \neq 0$, and $\varphi : x \to cx^{\alpha}$, then for every $a \in A$, $[\varphi_* (kf)] (a) = \varphi((kf)(a)) = ck^{\alpha}(f(a))^{\alpha} = k^{\alpha}[\varphi_* (f)](a)$. Hence $\varphi_* (kf) = k^{\alpha} \varphi_* (f)$, and therefore $\varphi_* (k^{1/\alpha} f) = k \varphi_* (f)$. Thus every two functions in F' are similar, and any function which is similar to a function $\varphi_* (f)$ in F' , also belong to F' . Thus F' is a ratio scale. Clearly

φ_* is a homogeneous function, with $\dim \varphi_* = \alpha$.

The converse is proved as for Theorem 4-3.6.

(ii) If $\alpha = 1$, φ is a similarity transformation of R^+ , hence $\varphi_*(f) = \varphi f \in F$, and φ_* is just an automorphism of F . Conversely, suppose that $F' = F$. This means that if $f \in F$, then $\varphi_*(f) \in F$. That is, there exists $k \in R^+$, such that for every x in the range of some $f \in F$ (i.e., every $x \in R^+$) $cx^\alpha = kx$. Hence $c = k$, and $\alpha = 1$.

Corollary. Let F be a complete ratio scale. Then, for every $\alpha \neq 0$ or 1 , the functions f and f^α ($f^\alpha : a \to [f(a)]^\alpha$ for every $a \in A$) have the same "unit", but they do not agree on any other element of their common domain. More generally, if $\alpha \neq 0$, the functions f and cf^α $(c > 0)$ determine the same equivalence relation ($a_1 \sim a_2$ if and only if the function has the same value as a_1 and a_2) on their common domain; if, in addition, $\alpha \neq 1$, then they agree on exactly one equivalence class of domain elements. (We leave to you the proof of these statements.)

Remarks:

1. We saw in Theorem 4-3.4 that if φ_* is a homogeneous function with $\dim \varphi_* = \alpha$, then $c\varphi$ is also a homogeneous function with the same dimension α ; and that any two homogeneous functions from F to F' with the same dimension α , must differ by composition with a "constant" homogeneous function, $\bar{\bar{c}} : f \to cf$. We also observe that if $c > 0$, $\alpha \neq 0$, and $\varphi : x \to x^\alpha$, then $F' = \{\varphi f : f \in F\} = \{(c\varphi)f : f \in F\}$, because φf and $(c\varphi)f = c(\varphi f)$ belong to the same ratio scale. In the consideration of ratio scales which are derived from (or "generated" by) other ratio scales, we usually want to keep in mind not only the derived scale, but also the associated homogeneous functions. For this reason, if $F' = \{\varphi f\} = \{(c\varphi)f\}$, we might like to adopt the convention that the notation cF^α denotes the pair of objects consisting of the ratio scale F' and the homogeneous function $\varphi_* : F \to F'$, where $\varphi : x \to cx^\alpha$. However, with $(\alpha \neq 0)$ and $\varphi : x \to x^\alpha$, the sets of functions $\{\varphi f : f \in F\}$ and $\{(c\varphi)f : f \in F\}$ are the same, and it is

more useful to use the notations F^α , cF^α , to simply denote these sets of functions. From the definition of set equality, $F^\alpha = cF^\alpha$ for every $c > 0$. Thus we do not complicate the notation by trying to include the homogeneous function φ_* within it, but whenever a homogeneous function from F to F^α is implicit in the discussion of a derived scale F^α (particularly in connection with questions of dimension), it is understood that $\varphi_* : f \rightarrow cf^\alpha$ (for some $c > 0$) is the implied homogeneous function; these homogeneous functions have the same dimension, α , for all c .

2. It is a simple consequence of the theorem, that $F^1 = F$, and that $F^\alpha = F^\beta$ if and only if $\alpha = \beta$.

3. We observe that if $\varphi : x \rightarrow x^\alpha$ and $f \in F$, then $\varphi f = f^\alpha$. Hence the notation F^α is consistent with the usual usage in the algebra of real valued functions, where f^α denotes the function $f^\alpha : a \rightarrow [f(a)]^\alpha$ Because we are dealing with positive-valued functions only, f^α is well-defined for every real α ; and if $\alpha \neq 0$, $F^\alpha = \{f^\alpha : f \in F\} = \{cf^\alpha : f \in F\}$.

4. We have observed that $F^\alpha = cF^\alpha$, if $\alpha \neq 0$. Hence if $\alpha \neq 0$, $F^\alpha = \bigcup_{c \in R^+} (cF^\alpha)$. It is convenient to define $F^0 = \bigcup_{c \in R^+} (cF^0) = \{cf^0 : f \in F , c \in R^+\}$. That is, F^0 is the very simple (incomplete) ratio scale, consisting of all constant positive valued functions on A . In spite of its apparent triviality, we shall find examples of the occurrence of such scales in connection with the question of "dimensional constants".

5. Notice that we are not saying that the ratio scales F^α are the only scales with the same domain A . The set of all scales with domain A is, in general, vastly larger. The "power" scales F^α are just those which are derived from one particular scale F by the process described.

If $\varphi' : x \rightarrow c'x^\alpha$, and $\varphi'' : x \rightarrow c''x^\beta$, we may compose and get $\varphi'' \varphi' : x \rightarrow c''(c'x^\alpha)^\beta = c''(c')^\beta x^{\alpha\beta}$. Hence we get

Theorem 4-4.2. For any ratio scale F , with φ' , φ'' as above,

(i) $(F^\alpha)^\beta = F^{\alpha\beta} = (F^\beta)^\alpha$.

(ii) If $\alpha \neq 0$, and $F^\alpha = G$, then $G^{1/\alpha} = F$, and $\{F^\alpha : \alpha \in R\} = \{G^\alpha : \alpha \in R\}$.

(iii) $(\varphi'' \cdot \varphi')_* = \varphi''_* \ \varphi'_*$.

(iv) $\dim(\varphi'' \ \varphi')_* = \dim \varphi''_* \cdot \dim \varphi'_* = \alpha\beta$.

Proof. This is a straightforward exercise in the use of exponents.

Remark: The expression "F^α has dimension α in F" is frequently used. This involves the convention that we are really referring to the dimension of φ_* , where $\varphi : x \to cx^\alpha$ for some $c > 0$. Sometimes the statement "F^α has dimension α" is used, but we shall avoid this form of expression, which is a hangover from past attempts to associate a dimension with a scale, rather than with a homogeneous function between scales.

A Notational Convention. In the study of ratio scales the function $x \to x^\alpha$ arises so frequently that it is convenient to give it a standard name. In Chapter 1 we used the symbol I_{R^+} to denote the identity function on R^+ . We could use a corresponding notation $I^\alpha_{R^+} : x \to x^\alpha$, but this is rather cumbersome. For the present we leave the domain out of the notation, and use the abbreviated notation $I^\alpha : x \to x^\alpha$. The basic properties of I^α are, of course,

(i) $I^\beta \cdot I^\alpha = I^{\alpha+\beta}$;

(ii) $I^\beta \circ I^\alpha = I^\alpha \circ I^\beta = I^{\alpha\beta}$;

(iii) if $c > 0$, $(cI)^\alpha = c^\alpha I^\alpha$.

The " $\cdot$ " denotes multiplication of functions; the " $\circ$ " denotes composition of functions. We denote (as before) the induced homogeneous functions from a ratio scale to the derived scale (with the same domain) by I^α_* . (Where it is important to distinguish different homogeneous functions formed in this way, we use I^α_{1*} , I^α_{2*} , etc..) With this notation, the basic property of these induced homogeneous functions is

$$(I^\alpha \circ I^\beta)_* = I^\alpha_* \circ I^\beta_* = I^{\alpha\beta}_* .$$

Formation of New Scales by Pre-Composition. Let $F = \{f : f : A \to R^+\}$ be a ratio scale, and let δ be a function $\delta : B \to A$. Then we have

Theorem 4-4.3.

(i) $\{f\delta : f \in F\}$ is a ratio scale with domain B ;

(ii) the function $\delta^* : F \to \{f\delta : f \in F\} = \delta^* F$, defined by $\delta^*(f) = f\delta$, is a homogeneous function;

(iii) $\dim \delta^* = 1$;

(iv) $(\delta^* F)^\alpha = \delta^*(F^\alpha)$ for every $\alpha \in R$.

Proof. Clearly $\delta^*(kf) = (kf)\delta = k(f\delta) = k\delta^*(f)$, from which (i) - (iii) follow directly. (iv) follows from the associativity of functional composition: $(\delta^* f)^\alpha = I^\alpha(f\delta) = (I^\alpha f)\delta = \delta^*(f^\alpha)$.

Remarks:

1. You should notice that the homogeneous scale function δ^* is in the opposite direction (i.e., contravariant to) the domain mapping δ.

2. Whenever δ is a "domain mapping", and we refer to the scale $\delta^*(F)$ in a situation where a homogeneous scale function is implied, unless anything is specified to the contrary it is understood that the homogeneous function δ^* is the one involved. Thus if we ever use such an expression as "F and $\delta^*(F)$ have the same 'dimension'" (we probably won't, but other writers do) it is to be clearly understood that this is only another way of saying that the dimension of the "natural" homogeneous function $\delta^* : F \to \delta^*(F)$, is 1.

3. When we wish to distinguish between the homogeneous functions induced by δ on different scales with the same domain, we use notations such as δ_1^*, δ_2^*, etc.

4. If B is the domain of a ratio scale G, and A the domain of a ratio scale F, a domain function $\delta : B \to A$ does not generally induce a homogeneous scale function from F to G ; but there are important situations where such a connection does not exist. (See Example 2 below, and the discussion at the end of this section.)

Combination of Pre-Composition and Post-Composition. Suppose that we have mappings

$$C \xrightarrow{\bar{\delta}} B \xrightarrow{\delta} A \xrightarrow{f} R^+ \xrightarrow{I^\alpha} R^+ ,$$

where F is a ratio scale with domain A, and $f \in F$. Then, clearly,

(i) $(\delta\,\bar{\delta})^* = \bar{\delta}^*\,\delta^*$

so that the diagram below is commutative

$$\begin{tikzcd}
 & F \arrow[dl, "\delta^*"'] \arrow[dr, "(\delta\bar{\delta})^*"] & \\
\delta^*(F) \arrow[rr, "\bar{\delta}^*"'] & & \bar{\delta}^*(\delta^*(F)) = (\delta\,\bar{\delta})^*F
\end{tikzcd}$$

For each $f \in F$, part (iv) of Theorem 4-4.3 is equivalent to

$$I_*^\alpha(\delta^*(f)) = I_*^\alpha(f\delta) = I^\alpha f\delta = (I_*^\alpha(f))\delta = \delta^*(I_*^\alpha(f))$$

so that, with a slightly imprecise notation,

(ii) $I_*^\alpha\delta^* = \delta^* I_*^\alpha$.

This is more accurately represented in the following commutative diagram (where we distinguish different induced functions by subscripts)

$$\begin{tikzcd}
F \arrow[r, "I^\alpha_{1*}"] \arrow[d, "\delta_1^*"'] & F^\alpha \arrow[d, "\delta_2^*"] \\
\delta_1^*(F) \arrow[r, "I^\alpha_{2*}"'] & {[\delta_1^*(F)]^\alpha = \delta_2^*(F^\alpha)}
\end{tikzcd}$$

The simple "commutativity" properties (i) and (ii) above, together with the properties developed earlier for the composition of " I^{α} " functions, are of course just special cases of general properties associated with the composition of functions. They are fundamental in the development of more complicated relationships among ratio scales.

Example 1. As an example involving pre-composition, any "extension of domain" situation may be considered. Assume that a ratio scale F' with domain B has been obtained, and that it is desired to "extend" this scale to give a new scale F with some domain $A \supset B$. The requirements of an "extension" are that

(i) if $f' : B \to R^{+}$ belongs to the ratio scale F' on B, then there exists an $f(f : A \to R^{+})$ such that $f|B = f'$. If (using the notation of the above discussion) δ denotes the inclusion mapping (or injection) $\delta : B \hookrightarrow A$ (i.e., $\delta : b \to b$) then the condition $f|B = f'$ is equivalent to $f' = f\delta$;

(ii) the set of all such extensions f shall constitute a ratio scale F with domain A.

These conditions imply that:

(iii) $\delta^{*}(F) = F'$. $[\delta^{*}(f) = f\delta = f|B.]$ (That is, the function induced by the injection δ is a homogeneous function of dimension 1.);

(iv) if $f_1|B = f_1'$, and $f_2|B = f_2' = kf_1'$, then $f_2 = kf_1$.
More briefly, we can write this condition

$$\operatorname{ext}(kf') = k \operatorname{ext}(f') .$$

It is easy to verify that if (i) and (iv) are satisfied, then so are the other conditions. In practice, ext(f') is usually defined for each $f' \in F'$, in such a way that the function $f' \to \operatorname{ext} f'$ is obviously 1-1, and only condition (iv) needs to be verified. If you recall the various examples of extensions given in Chapters 2 and 3, you will see that condition (iv) may follow from the distributive property of multiplication over addition, or it may follow from properties of the least upper bound and the greatest lower bound.

Any scale F, related to a scale F' by "extension of the domain" as above, will be called an extension of F'.

It is a trivial matter to verify that, if $F = \{f\}$ is an extension of a ratio scale $F' = \{f'\}$, then the function $(\delta^*)^{-1} : F' \rightarrow F$ defined by $(\delta^{+})^{-1} : f' \rightarrow \text{ext } f'$ is inverse to δ^*, and $(\delta^*)^{-1}$ is a homogeneous function, whose dimension is also 1. In a certain sense this correspondence of F and F' is "natural". In most scientific work it is customary to suppress the functions δ^* arising in extension situations (i.e., induced by inclusion mappings of the domains) and to talk about (for example) <u>the</u> length scale, and <u>the</u> area scale, irrespective of the domain. As the suppressed homogeneous functions have dimension 1, this does not produce any errors in dimensional arguments, provided that it is clearly understood that the missing functions are the "natural" ones.

<u>Example</u> 2. As an example involving pre-composition and post-composition, let B_s be the set of square regions, let A be the set of segments, let δ be the function from a square region in B_s to a side (in A) and let F be the length scale for segments. Then $I_*^2(F) = F^2$, and $\delta^*(I_*^2(F)) = \delta^*(F^2) = G$ is the area scale for square regions.

Frequently, in scientific work, one encounters a "dimensional formula"

$$A = L^2$$

where, in some sense which is not usually completely clear, A stands for "area", and L for "length". If A is interpreted as the area scale (for square regions, say), and L as the length scale (for segments) this formula would not have any clear meaning in our terminology, because the two scales A and L^2 have different domains; and the difference is more significant than it would be if one domain were merely an extension of the other. Our equivalent for this "formula" is

$$A = \delta^*(L^2) = \delta^*(I_*^2(L))$$

where the value of δ on a square region is a segment (or the congruence class of segments) which is a side of the region. We could consider $A = L^2$ as an abbreviation. In either form, the formula does not assert that "area has dimension 2 in length": such an assertion has no meaning except in relation to a specific homogeneous function connecting the length and area scales. In practice the function "understood" is the above function $\delta^* I_*^2$, which has dimension 2. The importance of this observation is emphasized (but not necessarily in our terminology) in the better books which deal with measurement, in such statements as "there is no meaning in saying the dimensions of a physical quantity, until we have also specified the system of measurement with respect to which the dimensions are determined" (Bridgman, [17]); "dimensions are characteristic of magnitudes; they are not characteristic of physical quantities themselves". (Focken [3].) Unfortunately not all writers keep this point clearly in mind.

Remark: Generally speaking, where no confusion can result we shall not distinguish between different ratio scales which are naturally related by inclusion mappings of their domains; (i.e., between a scale and its extensions). However, while on the subject of extensions, let us look at the relationships of length scales for segments and curves (with domains A_s , A_c ; i ; $A_s \subset A_c$) and area scales for squares, rectangles, and polygons (with domains B_s , B_r , B_p ; $j_1 : B_s \subset B_r$; $j_2 : B_r \subset B_p$) , where the connection between length and area scales is made through the function δ which maps square regions onto their sides, and through the function $I^2 : x \to x^2$. Area functions are denoted by "g's" , and length functions by "f's" .

$$\begin{tikzcd}
B_p \arrow[rdd, "g_p"] & \\
B_r \arrow[u, "j_2"] \arrow[rd, "g_r"] & \\
B_s \arrow[u, "j_1"] \arrow[r, "g_s"] \arrow[d, "\delta"] & R^+ \\
A_s \arrow[r, "f_s"] \arrow[d, "i"] & R^+ \arrow[u, "I^2"] \\
A_c \arrow[ru, "f_c"] &
\end{tikzcd}$$

The diagram is commutative. (Notice that there are ratio scales shown for which we have no particular names or use (e.g., the scale $\{I^2 f_c\}$, with domain A_c).) If F_s , F_c , denote the length scales for segments and curves respectively and if G_s , G_r , G_p denote the area scales for squares, rectangles, and polygons respectively, the induced homogeneous functions on these scales are indicated in the diagram:

$$\begin{tikzcd}
F_c \arrow[r, leftrightarrow, "i^*"] & F_s \arrow[r, leftrightarrow, "I_*^2"] & F_s^2 \arrow[r, leftrightarrow, "\delta^*"] & G_s \arrow[r, leftrightarrow, "j_1^*"] & G_r \arrow[r, leftrightarrow, "j_2^*"] & G_p
\end{tikzcd}$$

All of these homogeneous functions are 1-1 and have inverses.

Products of Ratio Scales. Let F_1, F_2, ..., F_n be ratio scales, with domains A_1, A_2, ..., A_n respectively. Denote functions in F_i by f_i, f_i', f_i'', etc. We shall describe a new scale whose domain is the cartesian product of the domains A_i.

Let $\Pi A_i = A_1 \times A_2 \times \ldots \times A_n$; let $a = (a_1, a_2, \ldots, a_n) \in \Pi A_i$; and let $f_i \in F_i$, i=1, 2, ..., n. Then the "product" function, $\Pi f_i = f_1 \times f_2 \times \ldots \times f_n$, defined by

$$\Pi f_i : (a_1, a_2, \ldots, a_n) \rightarrow (f_1(a_1), f_2(a_2), \ldots, f_n(a_n))$$

maps ΠA_i into $\Pi R_i^+ = R_1^+ \times R_2^+ \times . . . R_n^+$. (Each $R_i^+ = R^+$.)

Let $\mu : \Pi R_i^+ \rightarrow R^+$ denote the "multiplication" function which maps the element $(x_1, x_2, \ldots, x_n)$ of ΠR_i^+ into the real-number product $x_1 x_2 \ldots x_n$. Where no confusion can arise we denote this product by $\prod_{i=1}^{n} x_i$, or, more briefly, simply as Πx_i. Thus, for each element $a = (a_1, a_2, \ldots, a_n) \in \Pi A_i$, we have $[\mu(\Pi f_i)](a) = \Pi(f_i(a_i))$.

The composite functions $\mu(\Pi f_i) : \Pi A_i \rightarrow R^+$ are going to be the functions of a ratio scale, which we will denote by $F_1 \underline{\times} F_2 \underline{\times} \ldots \underline{\times} F_n$, with domain ΠA_i. But first we make a very important observation: although the functions Πf_i are all different (i.e., $\Pi f_i \neq \Pi f_i'$ if, for at least one i, $f_i \neq f_i'$) this is not true for the composite functions $\mu(\Pi f_i)$. To see this, consider the simple case where $i = 2$. Then, clearly, the function $\mu(2f_1 \times f_2)$ and the function $\mu(f_1 \times 2f_2)$ agree on every element of $A_1 \times A_2$, hence, as functions from $A_1 \times A_2$ to R^+, they are the same function. A little thought will show you that every function in the set

$F_1 \underline{\times} F_2 \underline{\times} \dots \underline{\times} F_n = \{\mu(\Pi f_i) : f_i \in F_i\}$ will have infinitely many designations or "representations". This is one of the reasons why we used the notation " $\underline{\times}$ " to distinguish the product scale from the much larger cartesian product of the sets F_i .

To see the general situation, let f_i , $f_i' \in F_i$, $(i=1, 2, \dots, n)$ and let $f_i' = \bar{k}_i f_i$. (We use the notation $\bar{k}_i f_i$, representing composition with an automorphism of $(R_i^+, +)$, rather than the "scalar product" $k_i f_i$, because it makes the diagram below easier to follow.) Then the relationship of the functions $\mu(\Pi f_i)$ and $\mu(\Pi f_i') = \mu(\Pi(\bar{k} f_i))$ is shown in the following diagram, which is easily shown to be commutative:

$$\begin{tikzcd}
 & \Pi A_i \arrow[dl, "\Pi f_i"'] \arrow[dr, "\Pi(\bar{k}_i f_i)"] & \\
\Pi R_i^+ \arrow[rr, leftrightarrow, "\Pi \bar{k}_i"] \arrow[d, "\mu"'] & & \Pi R_i^+ \arrow[d, "\mu"] \\
R^+ \arrow[rr, leftrightarrow, "\overline{\Pi k_i}"] & & R^+
\end{tikzcd}$$

Hence we have

$$\mu(\Pi \bar{k} f_i) = \overline{\Pi k_i}\ (\mu(\Pi f_i)) ,$$

so that $\mu(\Pi \bar{k}_i f_i) = \mu(\Pi f_i)$ if and only if $\Pi k_i = 1$. This property determines an equivalence relation (functional equality) on the set of functional representations. When we refer to the set of functions $\{\mu(\Pi f_i)\}$ we are not concerned, of course, with any particular representation for each function in the set.

We can now prove quite simply that the set of functions $\{\mu(\Pi f_i')\}$ is a ratio scale.

<u>Theorem 4-4.4.</u>

(i) $F_1 \underline{\times} F_2 \underline{\times} \dots \underline{\times} F_n = \{\mu(\Pi f_i) : f_i \in F_i\}$ is a ratio scale with domain ΠA_i .

(ii) If $\bar{f}_j$ is a fixed function in each $F_j (j \neq i)$, then every function $\mu(\Pi f_i)$ in $F_1 \underline{\times} F_2 \underline{\times} \dots \underline{\times} F_n$ can be written uniquely as

$$\mu(\Pi f_i) = \mu(\bar{f}_1 \times \dots \times \bar{f}_{i-1} \times f_i' \times \bar{f}_{i+1} \times \dots \times \bar{f}_n).$$

(In other words, we get each function in the "product" scale exactly once, if we consider only those representations in which the "factors" are fixed in all but one of the "factor" scales.)

Proof.

(i) If $\mu(\Pi f_i)$ and $\mu(\Pi f_i') = \mu(\Pi(\bar{k}_i f_i))$ are any two (not necessarily different) functions in $F_1 \underline{\times} F_2 \underline{\times} \dots \underline{\times} F_n$, then, from the above, $\mu(\Pi f_i') = \Pi k_i \, (\mu(\Pi f_i))$, so that the two functions are similar. On the other hand, if $k(\mu(\Pi f_i))$ is any function similar to $\mu(\Pi f_i)$, then $k_1, k_2, \dots, k_n$ can be chosen (in infinitely many ways) so that $\Pi k_i = k$. For each choice we get the same function $\mu(\Pi(\bar{k}_i f_i))$, and this function is $\bar{k}(\mu(\Pi f_i))$. Hence $F_1 \underline{\times} F_2 \underline{\times} \dots \underline{\times} F_n$ is a ratio scale.

(ii) The proof of this useful result is now quite straightforward, and it is left for you to complete for yourself. (This result is closely related to Theorem 4-3.7, and the fact that the choices made in the proof did not affect the result.)

The ratio scale $F_1 \underline{\times} F_2 \underline{\times} \dots \underline{\times} F_n$ is called the <u>product of the scales</u> F_i, and each F_i in the product is called a <u>factor</u> of the product. (It might be better to say <u>a</u> product, but the order is not really significant: see below.) It is sometimes convenient to abbreviate the notation to $\prod_{i=1}^{n} F_i$, or simply to $\underline{\Pi} F_i$ but we must not confuse this with the cartesian product of the sets F_i.

The following theorem gives some of the formal relationships between the three different procedures which we have described for generating ratio scales.

You should draw the relevant diagrams; with "parallel" diagrams indicating the behavior for a single domain element. As you will find, the diagrams are commutative; and the proofs are either trivial, or depend on well-known properties of products and powers of positive real numbers.

Theorem 4-4.5. For each $i (i=1, 2, \ldots, n)$ let F_i be a ratio scale; let α_i be a real number; let A_i be the domain of F_i; and let δ_i be a function from a set B_i to A_i. Then

(i) $(\underline{\Pi} F_i)^{\alpha} = \underline{\Pi} F_i^{\alpha}$

(ii) $\underline{\Pi}(\delta_i^* F_i) = (\Pi \delta_i)^* (\underline{\Pi} F_i)$.

The Order of Terms in the Product. Let F_i, $(i=1, 2, \ldots, n)$ be ratio scales, and $\underline{\Pi} F_i$ and $\underline{\Pi}' F_i$ be the products formed by using two different orders; (i.e., one order is a permutation of the other). Then the permutation gives a natural 1-1 correspondence (a homogeneous function of dimension 1) connecting $\underline{\Pi} F_i$ and $\underline{\Pi}' F_i$. It is convenient to "suppress" this dimension 1 function, and "identify" the scales $\underline{\Pi} F_i$ and $\underline{\Pi}' F_i$. That is, we treat the product as a commutative associative operation on ratio scales.

Homogeneous Functions Related to Products of Ratio Scales. In our discussion of post-composition and pre-composition, we emphasized that we were not only interested in obtaining new scales, but we were also interested in homogeneous functions relating the new and the old scales. The situation is similar (but more complicated) with respect to products of scales. Let $\underline{\Pi} F_i$ be the product of ratio scales F_i, $i=1, 2, \ldots, n$. Then, as we have already seen, if we consider fixed functions $\bar{f}_j (j \neq i)$ in all but one of the factors, then each function $\mu(\Pi f_i)$ in $\underline{\Pi} F_i$ can be uniquely expressed as $\mu(\Pi f_i')$, where $f_j' = \bar{f}_j$ for each $j \neq i$. It is natural to examine the resulting function

$$p_i : \underline{\Pi} F_i \to F_i$$

defined by $p_i : \mu(\Pi f_i) \to f_i'$, to see if p_i is a homogeneous function. It is trivial to verify that this is a homogeneous function, with dimension 1, but you should bear in mind that p_i depends on the choices of $\bar{f}_j (j \neq i)$. For each choice of the $\bar{f}_j$, we will get a "projection" p_i, but of course this correspondence of choices and projections is not 1-1. Moreover, as all projections have dimension 1, we know (from Theorem 4-3.4) that any two of them can differ only by a "constant", a fact which we could also verify directly. Thus, from a dimensional point of view, they are not significantly different.

Each such projection p_i has an inverse p_i^{-1}, which is also a homogeneous function with dimension 1. This function (which might be called an "injection") can be defined directly by

$$p_i^{-1} : f_i \to \mu(\bar{f}_1 \times \ldots \times \bar{f}_{i-1} \times f_i \times \bar{f}_{i+1} \times \ldots \times \bar{f}_n) .$$

Of course p_i^{-1} also depends on the choice of $\bar{f}_j (j \neq i)$; but different functions p_i^{-1} can differ at most by composition with a "constant". (I.e., by an automorphism of $(\Pi F_i , +)$.)

Another way in which a relationship can be established between factor and product scales is by pre-composition with a suitable mapping of the respective domains. With the notation used earlier, let $F_i (i=1, 2, \ldots, n)$ be ratio scales with domains A_i, and let $\bar{a}_2 , \bar{a}_3 , \ldots , \bar{a}_n$ be fixed elements of the domains $A_2 , \ldots , A_n$ respectively. (We are fixing a_i in all domain factors except the first to simply notation: of course any other factor than the first could be singled out.) Let δ be the function $A_1 \to \Pi A_i$ defined by $\delta : a_1 \to (a_1 , \bar{a}_2 , \bar{a}_3 , \ldots , \bar{a}_n)$. Then δ determines a ratio scale $\delta^*(\Pi F_i)$ and a homogeneous function δ^* from ΠF_i onto this scale. We shall show that $\delta^*(\Pi F_i)$ is actually F_1, and that δ^* is one of the projections onto F_1, as defined earlier.

Theorem 4-4.6.

(i) $\delta^*(\Pi F_i) = F_1$;

(ii) δ^* is the projection determined by those functions $\bar{f}_2 , \ldots , \bar{f}_n$ which have $\bar{a}_2 , \ldots , \bar{a}_n$ as units.

Proof.

(i) Let $\overline{f}_2 , \ldots , \overline{f}_n$ be the (unique) functions in $F_2 , \ldots ,$ F_n, respectively, which have $\overline{a}_2 , \ldots , \overline{a}_n$ as units. Then, from the definition of δ^*, for any $a_1 \in A_1$ and any $\mu(\Pi f_i) \in \underline{\Pi F_i}$,

$$[\delta^*(\mu(\Pi f_i))](a_1) = \mu[(\Pi f_i)(\delta(a_1))]$$

$$= \mu[(\Pi f_i)(a_1 , \overline{a}_2 , \ldots , \overline{a}_n)]$$

$$= \mu[(f_1' \times \overline{f}_2 \times \ldots \times \overline{f}_n)(a_1 , \overline{a}_2 , \ldots , \overline{a}_n)]$$

$$= f_1'(a_1) ;$$

where f_1' is the unique function in F_1 such that $\mu(\Pi f_i) = \mu(f_1' \times \overline{f}_2 \times \ldots \times \overline{f}_n)$. Hence $\delta^*(\mu(\Pi f_i)) = f_1'$, so that $\delta^*(\underline{\Pi F_i}) \subset F_1$. It is easy to show that δ^* is 1-1 and onto.

(ii) The fact that δ^* is the projection determined by the functions $\overline{f}_j (j \neq 1)$ follows immediately from the definition of projections given earlier.

Remark: The above theorem is closely connected with the example (which we have frequently mentioned) of a dimension 1 function from the length scale (for segments, say), to the area scale (for rectangles, say). For, if F is this length scale and G the area scale, with domains A and B respectively, a fixed segment $\overline{a}$ determines a fixed length function $\overline{f}$, which has $\overline{a}$ as unit. Let $a \in A$, and let $\delta_1 : a \to (a , \overline{a})$. Let $b \in B$, and let $\delta_2 : b \to (a_1 , a_2)$, where a_1, a_2, are the sides of b (in an arbitrarily selected order). Then, for each $f \in F$, $\mu(f \times \overline{f})\,\delta_2$ is an area function for B. That is, there is an area function $g \in G$ which makes the following diagram commutative: (Commutativity of the top "triangle" is trivial; for the lower "triangle", this was proved in Section 3-5.)

$$\begin{tikzcd}
 & A \arrow[dl, "\delta_1"'] \arrow[dr, "f"] & \\
A \times \bar{A} \arrow[r, "f \times \bar{f}"] & R^+ \times R^+ \arrow[r, "\mu"] & R^+ \\
 & B \arrow[ul, "\delta_2"] \arrow[ur, "g"'] &
\end{tikzcd}$$

For fixed $\bar{a}$ (and hence $\bar{f}$) the set of all functions $\{\mu(f \times \bar{f})\}$ is $F \underline{\times} F$. Hence there are induced homogeneous functions δ_1^*, δ_2^*, both of dimension 1,

$$\begin{tikzcd}
F & F \underline{\times} F \arrow[l, two heads, "\delta_1^*"'] \arrow[r, two heads, "\delta_2^*"] & G .
\end{tikzcd}$$

Each function is 1-1 and hence $\delta_2^*(\delta_1^*)^{-1}$ is a homogeneous function of dimension 1 from F onto G. This is, of course, the homogeneous function η_σ of Section 3-5, with $\sigma = \bar{a}$.

Products of Powers of Ratio Scales. If F_1, F_2, ..., F_n are ratio scales, and α_1, α_2, ..., α_n are real numbers, we can form the product scale $\Pi F_i^{\alpha_i} = F_1^{\alpha_1} \underline{\times} F_2^{\alpha_2} \underline{\times} \dots \underline{\times} F_n^{\alpha_n}$. If p_i is a projection, $p_i : \Pi F_i^{\alpha_i} \to F_i^{\alpha_i}$, we can compose the injection p_i^{-1} with the "power" function $I_*^{\alpha_i}$, and obtain

$$\begin{tikzcd}
F_i \arrow[r, "I_*^{\alpha_i}", "(\alpha_i)"'] & F_i^{\alpha_i} & \Pi F_i^{\alpha_i} \arrow[l, two heads, "p_i"', "(1)"] ,
\end{tikzcd}$$

where the numbers in parentheses (on the diagram) denote the dimensions of the corresponding homogeneous functions. The composite function $p_i^{-1} I_*^{\alpha_i}$ has dimension α_i, and it is uniquely determined up to an automorphism of $\Pi F_i^{\alpha_i}$. This property is sometimes expressed in the statement "$\Pi F_i^{\alpha_i}$ has dimension α_i in F_i", but this only makes sense if it is interpreted to mean that the above homogeneous function has dimension α_i.

If $\alpha_i = \alpha$, $(i=1, 2, \ldots, n)$, then, as we have seen, $(\underline{\Pi F}_i)^{\alpha} = \underline{\Pi F}_i^{\alpha}$. In this case $\underline{\Pi F}_i$ can also be mapped homogeneously into $\underline{\Pi F}_i^{\alpha}$ by first injecting it into $\underline{\Pi F}_i$, and following this by the appropriate I_*^{α} function from $\underline{\Pi F}_i$ to $(\underline{\Pi F}_i)^{\alpha} = \underline{\Pi F}_i^{\alpha}$. This composite homogeneous function also has dimension α . In fact, for appropriate choices of the injections involved, we get a commutative diagram

$$\begin{tikzcd}
F_i \arrow[r] \arrow[d, dash] & F_i^{\alpha} \arrow[d] \\
\underline{\Pi F}_i \arrow[r] & (\underline{\Pi F}_i)^{\alpha} = \underline{\Pi F}_i^{\alpha}
\end{tikzcd}$$

<u>The Relationship of Powers and Products of Ratio Scales.</u> If F is a ratio scale, we have now defined ratio scales $F \underline{\times} F$, and F^2 . It is natural to ask whether there is any simple connection between these scales. Clearly they cannot be the same, because if the domain of F is A , then the domain of $F \underline{\times} F$ is $A \times A$, while the domain of F^2 is A .

In Theorem 4-4.6 above we saw one way of mapping a "factor" scale into a "product" scale. This method involved fixed choices of elements in all but one of the domains. But when we are considering the cartesian product of a ratio scale with itself, there is a simple and natural way to map the domain of the scale into the cartesian product of the domain with itself without making arbitrary choices: this important mapping (which has many uses in mathematics) is called the <u>diagonal</u> mapping. This is the function

$$\Delta : A \to A \times A$$

defined by

$$\Delta : a \to (a,a)$$

(Clearly the idea can be extended to the case of any finite number, n , of equal factors, but we shall keep things simple by examining only the case $n = 2$.)

If $\mu(f' \times f'')$ is any mapping in $F \underline{\times} F$, then $\mu(f' \times f'')\Delta$ maps A into R^{+}. Hence, by pre-composition, Δ determines a new ratio scale $\Delta^{*}(F \underline{\times} F)$, whose domain is A, as indicated in the diagram

$$\begin{tikzcd}
A \arrow[d, "\Delta"'] \arrow[rrd, dashed, "\mu(f' \times f'')\Delta"] & & \\
A \times A \arrow[r, "f' \times f''"'] & R^{+} \times R^{+} \arrow[r, "\mu"'] & R^{+}
\end{tikzcd}$$

It is natural to ask whether this scale has any simple connection with the scale F. In order to answer this question, we first prove a simple but useful theorem concerning the "representations" of a map $\mu(f' \times f'')$ from the scale $F \underline{\times} F$. (Recall that this map can be written in infinitely many different ways.)

Theorem 4-4.7.

(i) Given any function $\mu(f' \times f'')$ in $F \underline{\times} F$, there is a unique $f \in F$ such that

$$\mu(f' \times f'') = \mu(f \times f) .$$

(In other words, each function in the product has a unique "diagonal" representation.)

(ii) Different elements of $F \underline{\times} F$ have different representations in the form $\mu(f \times f)$.

(iii) The function $\bar{\Delta} : F \to F \underline{\times} F$ defined by $\bar{\Delta} : f \to \mu(f \times f)$ is a homogeneous function with dimension 2.

Proof.

(i) For each $(a_1, a_2) \in A \times A$, $[\mu(f' \times f'')](a_1, a_2) = [f'(a_1)][f''(a_2)]$ does not depend on the representation of the element $\mu(f' \times f'')$ of $F \underline{\times} F$. Hence if there is an $f \in F$ such that $\mu(f \times f) = \mu(f' \times f'')$, then for $(a, a) \in A \times A$, $[\mu(f \times f)](a, a) = [f(a)]^{2} = [f'(a)][f''(a)]$, so that $f(a) = [f'(a)]^{1/2}[f''(a)]^{1/2}$.

Define a function f on A by this equation. The functions f', f'', belong to F, hence, for some $k > 0$, $f'' = kf'$, so that $f = k^{1/2}f'$, and hence $f \in F$. (This is, of course, just a special case of the general result (cf. Exercise 2-4.16) that any homogeneous function of degree 1 in a set of ratio scale functions also belongs to the same scale; the function $f = \sqrt{f'f''}$ has degree 1.) The uniqueness of f is easily shown.

(ii) This is trivial.

(iii) From (i) and (ii), $\overline{\Delta}$ is 1-1 and onto. For any $f \in F$ and any $k \in R^+$, $\overline{\Delta}(kf) = \mu(kf \times kf) = k^2\mu(f \times f) = k^2\overline{\Delta}(f)$, so that $\overline{\Delta}$ is homogeneous, with dimension 2.

Remark: The above theorem clearly extends to the case where there are n equal factors in the cartesian product, in which case the function needed for the "diagonal" representation is simply the nth root of the product of the functions in any representation of an element of the cartesian product. In this case the corresponding homogeneous function $\overline{\Delta}$ has dimension n.

We are now in a position to answer the question above: "How is the scale $\Delta^*(F \underline{\times} F)$ related to the scale F?" If we use the unique "diagonal representation" for every element of $F \underline{\times} F$, and consider the following diagram, then the answer is almost obvious:

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "\Delta"'] & R^+ \arrow[dr, dashed, "?"] & \\
A \times A \arrow[r, "f \times f"'] & R^+ \times R^+ \arrow[r, "\mu"'] & R^+
\end{tikzcd}$$

The missing function, which makes the diagram commutative, is the function $I^2 : x \to x^2$; with this function, we have $I^2 f = \mu(f \times f)\Delta$, so that $\Delta^*(F \underline{\times} F) = F^2$. We incorporate this important result in the following theorem.

Theorem 4-4.8.

(i) $\Delta^*(F \underline{\times} F) = F^2$

(ii) $\Delta^* : F \underline{\times} F \to F^2$ is a homogeneous function with dimension 1.

(iii) The homogeneous functions $\bar{\Delta}$, Δ^*, and I_*^2 are related as in the following commutative diagram, in which the dimensions of the homogeneous scale functions are shown in parentheses after the names of the functions:

$$\begin{tikzcd}
F \arrow[r, leftrightarrow, "I_*^2(2)"] \arrow[d, leftrightarrow, "\bar{\Delta}(2)"'] & F^2 \\
F \underline{\times} F \arrow[ur, leftrightarrow, "\Delta^*(1)"'] &
\end{tikzcd}$$

Remark. In most scientific work (in which the dimensions of relationships between scales are often the main consideration) the notation $F \underline{\times} F$ is not used, and the notation F^2 is frequently used both for the scale which we have designated as F^2 (and whose domain is A) and also for the scale which we have designated as $F \underline{\times} F$ (and whose domain is $A \times A$). Perhaps conventions such as this are needed if notation is to be kept reasonably simple, but in our treatment it is important to be aware of the distinction.

Example. We return again to the question of the relationship of the length and area scales, involving the homogeneous functions $\bar{\Delta}$, Δ^*. Let F be the length scale (for segments, say) with domain A, let G_r be the area scale for rectangular regions (with domain B_r), and let G_s be the area scale for square regions (with domain B_s). Then the following diagram is commutative:

$$\begin{tikzcd}
B_r \arrow[drr, "\bar{\bar{\delta}}"'] & B_s \arrow[l, "j"'] \arrow[r, "\delta"] \arrow[dr, "\bar{\delta}"] & A \arrow[r, "f"] \arrow[d, "\Delta"] & R^+ \arrow[r, "I^2"] & R^+ \\
 & & A \times A \arrow[r, "f \times f"] & R^+ \times R^+ \arrow[ur, "\mu"'] &
\end{tikzcd}$$

(The mappings are the usual ones; $\bar{\bar{\delta}}$ maps a rectangular region onto an ordered pair of sides. We could, of course, use a corresponding function from $A \times A$ to B_r, defined by taking any element (pair of segments) in $A \times A$, and mapping this onto any rectangle with sides congruent to a_1, a_2.)

The corresponding homogeneous functions are shown in the following diagram, whose commutativity follows from the various results proved above. (The notation is standard: but recall that we use δ_1^*, δ_2^*, to indicate the different homogeneous functions induced by the same function δ.) The product of the dimensions along any two "function paths" joining two scales must be the same, because of commutativity: the composite of homogeneous functions is a homogeneous function whose dimension is the product of those of the separate functions; the dimension of a homogeneous function is uniquely determined by the function; and commutativity simply says that the composite functions corresponding to different paths are, in fact, the same function. This "equality of dimensions" is a useful checking device: it is a necessary, but not sufficient, condition for the commutativity of diagrams in which we have homogeneous functions connecting ratio scales.

$$\begin{tikzcd}
F \arrow[rr, leftrightarrow, "I^2_*(2)"] \arrow[d, leftrightarrow, "\overline{\Delta}(2)"'] \arrow[drr, leftrightarrow, "\delta_1^*(1)" near end] & & F^2 \arrow[d, leftrightarrow, "\delta_2^*(1)"] \\
F \underline{\times} F \arrow[urr, leftrightarrow, "\Delta^*(1)"' near end] \arrow[rr, leftrightarrow, "\overline{\delta}^*(1)"] \arrow[dr, leftrightarrow, "\overline{\overline{\delta}}^*(1)"] & & G_s \arrow[dl, leftrightarrow, "j^*(1)"] \\
& G_r &
\end{tikzcd}$$

Theorems 4-4.7 and 4-4.8 can be generalized to show the relationship of $F^\alpha \underline{\times} F^\beta$ and $F^{\alpha+\beta}$. We state the results in the following theorem, leaving the details (and the extension to the case of $n > 2$ factors) to you.

Theorem 4-4.9.

(i) $F^\alpha \underline{\times} F^{-\alpha}$ is a well-defined ratio scale.

(ii) The function $\overline{\Delta}_1 : F \to F^\alpha \underline{\times} F^\beta$ defined by $\overline{\Delta}_1 : f \to \mu(f^\alpha \times f^\beta)$ is a homogeneous function with dimension $\alpha + \beta$.

(iii) If $\alpha + \beta \neq 0$, then every function $\mu(f_1 \times f_2)$ in $F^\alpha \underline{\times} F^\beta$ can be uniquely expressed in the form $\mu(f^\alpha \times f^\beta)$, where $f = [f_1^\alpha f_2^\beta]^{\frac{1}{\alpha + \beta}}$

(iv) With Δ denoting, as usual, the diagonal map on the domain of F, $\Delta^*(F^\alpha \underline{\times} F^\beta) = F^{\alpha+\beta}$

(v) The following diagram (in which dimensions are shown in parentheses) is commutative:

$$\begin{tikzcd}
F \arrow[r, "I_*^{\alpha+\beta}(\alpha+\beta)"] \arrow[d, "\begin{matrix}\overline{\Delta}_1\\(\alpha+\beta)\end{matrix}"'] & F^{\alpha+\beta} \arrow[dl, leftrightarrow, "\Delta^*(1)"] \\
F^{\alpha} \underline{\times} F^{\beta} &
\end{tikzcd}$$

If $\alpha+\beta \neq 0$, the homogeneous functions $I_*^{\alpha+\beta}$ and $\overline{\Delta}_1$ have inverses, but if $\alpha+\beta = 0$ they are "constant" functions.

Remark: We point out the relevance of this theorem for the relationship of linear and angular measures. Assume that F is the length scale (on a domain A which includes at least segments and circular arcs), and that G is the angular measure scale for simple angles. The diagonal map $\overline{\Delta}$: $F \rightarrow F \underline{\times} F^{-1}$ ($\overline{\Delta} : f \rightarrow \mu(f \times f^{-1})$) is a homogeneous function with dimension zero. If $b \in B$ (the domain of G) is an angle, and (a_1, a_2) are the intercepted arc and the radius of some fixed circle with center at the vertex of the angle, then the function $\delta : b \rightarrow (a_1, a_2)$ induces a homogeneous function $\delta^* : F \underline{\times} F^{-1} \rightarrow G$ of dimension 1. The radian measure function is the function $\delta^*[\mu(f \times f^{-1})]$; this is the same for every f . Thus we have the (trivially) commutative diagram

$$\begin{tikzcd}
F \arrow[r, "\overline{\Delta}(0)"] \arrow[dr, "\rho(0)"'] & F \times F^{-1} \arrow[d, "\delta^*(1)"] \\
& G
\end{tikzcd}$$

in which ρ is defined to be the 0-dimensional composite function $\delta^* \overline{\Delta}$. This result is the source of the common statement "angular measure has dimension zero in length".

However, it is possible to map the length scale into the angular measure scale by means of homogeneous functions with non-zero dimension. For example, we may use a "fixed" circle (whose radius is congruent to a fixed segment, $\bar{a}$, which is the unit of a fixed length function $\bar{f}$). For any (simple) angle b, we then have $\delta(b) = (a, \bar{a}) \in A \times A$. Any function in $F \underline{\times} F^{-1}$ can be uniquely represented in the form $\mu(f \times \bar{f}^{-1})$, the composite function $\mu(f \times \bar{f}^{-1})\delta$ is an angular measure function for G, and the resulting $\delta^* : F \underline{\times} F^{-1} \to G$ has dimension 1. If we compose this with the injection i (from F to $F \underline{\times} F^{-1}$) which is determined by $\bar{f}$, we get

$$\begin{tikzcd}
F \arrow[r, leftrightarrow, "i(1)"] \arrow[dr, leftrightarrow, "\sigma(1)"'] & F \underline{\times} F^{-1} \arrow[d, leftrightarrow, "\delta^*(1)"] \\
& G
\end{tikzcd}$$

and the composite function $\sigma = \delta^* i$ has dimension 1. (Intuitively, σ is the function determined when angular measures are set up by "fixing" the circle, and using arc length as the angular measure.)

A Non-Commutative Diagram. We have encountered so many commutative diagrams, that you might be inclined to think that commutativity will hold just about any time that we can exhibit functional relationships on a diagram. This is very far from being the case, and the statement that a diagram is commutative is usually a nontrivial statement.

As an example of a non-commutative diagram, let us consider again the relationship of the length and area scales. Let F be the length scale (for segments will do) and G the area scale for rectangles. Then we have seen that there are homogeneous functions relating the length scale F, its "square" F^2, the product $F \underline{\times} F$, and the area scale G. We can indicate these in the diagram

$$\begin{tikzcd}
 & F^2 \arrow[dr, leftrightarrow, "\delta_1^*"] & \\
F \arrow[ur, leftrightarrow, "I_*^2"] \arrow[dr, leftrightarrow, "i"'] & & G \\
 & F \underline{\times} F \arrow[ur, leftrightarrow, "\delta_2^*"'] &
\end{tikzcd}$$

where δ_1 is the function which maps a square b onto its side a, δ_2 maps $b \rightarrow (a,a)$, and i is the injection determined by a fixed $\bar{f} \in F$. As all functions shown are 1-1 and hence have inverses, all composite functions are 1-1 homogeneous functions. But the diagram is distinctly non-commutative, as you can easily check. (Of course we deliberately complicated matters by introducing F^2; as we saw earlier the composite function $\delta_1^* I_*^2$ is equal to $\delta_2^* \bar{\Delta}$, where $\bar{\Delta} : f \rightarrow (f,f)$. In this form the non-commutativity would be more obvious, because $\bar{\Delta}$ and i are clearly not the same function.) One method of checking the possibility of commutativity is to put in the dimensions: a necessary condition for commutativity is that the products of dimensions along different "function paths" should be the same.

Our purpose in introducing this simple and overworked example again, was not just in order to warn you that all diagrams are not automatically commutative. Let us redraw the diagram as follows, simplifying by using the fact that $\delta_2^* \bar{\Delta} = \delta_1^* I_*^2$, and showing the dimensions:

$$\begin{tikzcd}
 & F \underline{\times} F & G \arrow[l, "\delta_2^*(1)"'] \arrow[dd, leftrightarrow, "\rho"] \\
F \arrow[ur, leftrightarrow, "\bar{\Delta}(2)"] \arrow[dr, leftrightarrow, "i(1)"'] & & \\
 & F \underline{\times} F \arrow[r, leftrightarrow, "\delta_2^*(1)"'] & G
\end{tikzcd}$$

If we now ask whether there is a function ρ from G to G which makes the diagram commutative, the answer is obviously "yes" : Because all functions shown are 1-1, we can simply define ρ to be the composite function $\delta_2^* \, i(\overline{\Delta})^{-1}(\delta_2^*)^{-1}$, and the diagram is trivially commutative. Moreover, as all of the original functions are homogeneous functions, so is the composite function ρ , and its dimension is $\frac{1}{2}$. (ρ^{-1} has, of course, dimension 2.) Thus we have made a very important discovery, which emphasizes the impossibility of assigning a dimension to a ratio scale itself: namely, that there are 1-1 homogeneous functions, with dimension $\neq 1$, from a scale to itself.

Once we have seen this example, we can easily construct others. In fact, it is a simple matter to show that every ratio scale can be mapped 1-1 onto itself by a homogeneous function of any dimension $\alpha \neq 0$. If F is a ratio scale, the set of all such functions includes, of course, the automorphisms of $(F , +)$, which all have dimension 1.

To see the general case, let $\overline{f}$ be any fixed element of F , let $\alpha \neq 0$, and let $c > 0$. Define $\rho : F \to F$ by $\begin{cases} \rho(\overline{f}) = c\overline{f} \\ \rho(k\overline{f}) = k^{\alpha}\rho(\overline{f}) = ck^{\alpha}\overline{f} \, . \end{cases}$

Theorem 4-4.10.

(i) For every choice of c and $\overline{f}$, ρ is a homogeneous function with dimension α .

(ii) If $\alpha \neq 1$, then every such ρ has a "fixed point"; i.e., a function f' such that $\rho(f') = f'$.

Proof.

(i) This is a direct consequence of the relevant definitions.

(ii) If $f' = k\overline{f}$ is a fixed point, then $k\overline{f} = f' = \rho(f') = \rho(k\overline{f}) = ck^{\alpha}\overline{f}$, whence $k^{1-\alpha} = c$, and $k = c^{\frac{1}{1-\alpha}}$. With this k , $f' = k\overline{f}$ is easily shown to be a fixed point.

Remarks:

1. In the case of our length/area example, it is easy to find the fixed point--in fact you can almost guess what it must be. (Recall that

$\alpha = \frac{1}{2}$ or 2, so that there will be a fixed point.) If $\overline{f}$ is the fixed length function (corresponding, in the original form of the example, to a fixed segment a) then you can easily verify that the fixed point of the homogeneous function ρ described above (see last diagram) on the area scale, is $\delta_2^*[\mu(\overline{f} \times \overline{f})]$.

2. Back in the previous section we pointed out that a homogeneous function was not determined by its value on a single scale function--we needed two values in order to determine it completely. What we have been discussing above is clearly directly related to this earlier observation: the homogeneous functions (see last diagram) $\delta_2^* \overline{\Delta}$ and $\delta_2^* i$ have the same value on $\overline{f}$, but they are otherwise completely different.

3. The above length/area example, and Theorem 4-4.10, are relevant to the sometimes confusing question of "dimensional constants". We shall have more to say about this question in a later section.

The Use of Standard Homogeneous Functions. At the end of the previous section we gave a very simple example (using Pythagoras' Theorem) which indicated that functions which are homogeneous in the standard sense (i.e., which are not necessarily multihomogeneous) can sometimes be used in the construction of ratio scales. Moreover, in the measurement of "secondary quantities" it is quite conceivable that several domain elements of one of the "primary quantities" might be associated (together) with a single domain element of the secondary quantity, and that we might actually encounter a situation where such a homogeneous function is used.

Assume that we have a domain mapping $B \to \prod A_i$, where the A_i are not all distinct. Clearly we can bring together those "factors" A_i for which $A_i = A_0$, and get

$$\prod A_i = (\prod_{A_i=A_0} A_i) \times (\prod_{A_j \neq A_0} A_j).$$

This suggests that we might consider separately those ratio scales on $\prod_{A_i=A_0} A_i$ which are "generated" by the use of functions which are homogeneous in the ordinary sense. To keep things reasonably simple, let us consider the case of two equal "factors"; i.e., let F be a ratio scale with domain A, and let $h : R^+ \times R^+ \to R^+$ be a homogeneous function of degree α, with $\alpha \neq 0$.

That is $h(cx_1, cx_2) = c^{\alpha}h(x_1, x_2)$ for all $c > 0$, and all $(x_1, x_2) \in R^+ \times R^+$.

We observe first, that every such homogeneous function can be uniquely written as $h = I^{\alpha}\bar{h}$, where $I^{\alpha} : x \to x^{\alpha}$, and $\bar{h}$ is homogeneous of degree 1. (Define $\bar{h}$ by $\bar{h} : x \to [h(x)]^{1/\alpha}$, prove that $\bar{h}$ is homogeneous of degree 1, and show that $h = I^{\alpha}\bar{h}$.)

Now let $f \in F$, and consider the set $H = \{h(f \times f) : f \in F\}$. It is quite straightforward to verify that H is a ratio scale, and that if $\bar{H}$ is similarly defined using $\bar{h}$ instead of h, then $\bar{H}$ is also a ratio scale, and $\bar{H}^{\alpha} = H$. The domain of each of these scales is $A \times A$, so it is natural to look for relationships between the scales $\bar{H}^{\alpha}$ and other scales (such as $F \underline{\times} F$) which also have domain $A \times A$. We might, for example, ask whether $\bar{H}^2$ and $F \underline{\times} F$ are the same scale. If $h : (x_1, x_2) \to x_1 x_2$ (which is homogeneous of degree 2) then, clearly, $H = \bar{H}^2 = F \underline{\times} F$; but if we take $h : (x_1, x_2) \to x_1^2 + x_2^2$ (which is also homogeneous of degree 2) and compare values of $h(f \times f)$ and $\mu(f \times f)$ on non-diagonal domain elements in $A \times A$, we can quickly dispose of the question asked: in general $F \underline{\times} F$ and $\bar{H}^2$ are not the same scale. On the other hand, using the same example, we find that, for every $a \in A$, $h(f \times f)(a,a) = 2[f(a)]^2 = 2\mu(f \times f)(a,a)$. In fact, if we restrict ourselves to the set of diagonal elements only, we find that

(i) the sets of restricted mappings from $F \underline{\times} F$ and $\bar{H}^2$ each form ratio scales on the diagonal of $A \times A$ as domain;

(ii) these restricted scales are, in fact, the same scale. (Of course the restrictions of $h(f \times f)$ and $\mu(f \times f)$ to the diagonal of $A \times A$ are not the same function, but the fact that they differ by the constant factor "2" means that they belong to the same scale.)

In other words, for this particular homogeneous function h, there is a ratio scale on the diagonal of $A \times A$, such that $F \underline{\times} F$ and $\bar{H}^2$ (which have the same domain $A \times A$) are each extensions of this "diagonal" scale.

This situation holds generally. To see this, let h be any homogeneous function of degree 2 on $R^+ \times R^+$, and let F be any ratio scale with domain

A . Then for every diagonal element $(a,a) \in A \times A$, $h(f \times f)(a,a) = h(f(a), f(a)) = [f(a)]^2\, h(1,1)$ (from the homogeneity of h); while $\mu(f \times f)(a,a) = [f(a)]^2$. Thus the restricted functions $h(f \times f)$ and $\mu(f \times f)$, on the diagonal of $A \times A$, differ by the constant factor $h(1,1)$, and hence they belong to the same scale, whose domain is the diagonal of $A \times A$.

The diagonal mapping $\Delta : a \rightarrow (a,a)$ on A , relates the scale $F \underline{\times} F$ to F^2 , and it also relates the scale $\bar{H}^2$ to F^2 . (More specifically, $\Delta^*(F \underline{\times} F) = F^2$; $\Delta^*(\bar{H}^2) = F^2$.) Moreover the functions $\bar{\Delta}_1 : f \rightarrow \mu(f \times f)$, and $\bar{\Delta}_2 : f \rightarrow \bar{h}^2(f \times f)$ are homogeneous, and the following diagram is commutative:

$$\begin{tikzcd}
(F \underline{\times} F)^{1/2} \arrow[r] \arrow[d, "\Delta^*"'] & F \underline{\times} F \arrow[r] \arrow[d, "\Delta^*"] & (F \underline{\times} F)^{\alpha/2} \arrow[d, "\Delta^*"] \\
F \arrow[r] \arrow[ur, "\bar{\Delta}_1"] \arrow[d, "\bar{h}(1,1)"'] \arrow[ddr, "\bar{\Delta}_2"] & F^2 \arrow[r] \arrow[d, "{[\bar{h}(1,1)]^2}"] & F^\alpha \arrow[d, "{[\bar{h}(1,1)]^\alpha}"] \\
F \arrow[r] & F^2 \arrow[r] & F^\alpha \\
\bar{H} \arrow[r] \arrow[u, "\Delta^*"] & \bar{H}^2 \arrow[r] \arrow[u, "\Delta^*"'] & \bar{H}^\alpha \arrow[u, "\Delta^*"']
\end{tikzcd}$$

To avoid cluttering the diagram, we have not distinguished in notation between the various functions induced by Δ , nor have we labelled the horizontal "power" functions. The upper rectangles of the diagram represent the (established) commutativity of the precomposition (δ) functions (with $\delta = \Delta$) and the power functions. The commutativity of the lower rectangles is similar. And the commutativity of the middle rectangles follows from the fact that the horizontal I_*^α functions are homogeneous functions of the relevant dimensions: e.g., $I_*^2[\bar{h}(1,1)f] = [\bar{h}(1,1)]^2 I_*^2(f)$. Thus the only commutativity that really needs to be checked is the relation

$$\Delta^* \bar{\Delta}_2 = [\bar{h}(1,1)]^2 I_*^2 ,$$

which you can verify directly.

These results can be generalized to the situation where there are any finite number of "primary quantities", and where the "secondary quantity" is determined from a finite number of domain elements for each "primary quantity".

The Duality of Units and Functions. If F is a ratio scale, with domain A, and $f \in F$, then, as we have seen earlier, the relation defined by: $a_1 \sim a_2$ if and only if $f(a_1) = f(a_2)$, is an equivalence relation on A . As you may easily verify, this relation depends only on F , and not on f . Let $\tilde{A}$ denote the resulting set of equivalence classes. Then each $f \in F$ establishes a 1-1 correspondence of $\tilde{A}$ and R^+ if and only if F is a complete scale. Assume that F is complete, and define an addition, and a "scalar" multiplication by positive real numbers, for elements of $\tilde{A}$, by

(i) $\tilde{a}_1 + \tilde{a}_2 = \tilde{a}_3$, where $\tilde{a}_3$ is the unique class such that $f(\tilde{a}_1) + f(\tilde{a}_2) = f(\tilde{a}_3)$ for each $f \in F$;

(ii) $c\tilde{a}_1 = \tilde{a}_2$, where $\tilde{a}_2$ is the unique class such that $f(\tilde{a}_2) = cf(\tilde{a}_1)$ for each $f \in F$.

It is a trivial matter to verify that this addition is associative and commutative, that the scalar multiplication is doubly distributive (i.e., $(c_1 + c_2)\tilde{a} = c_1\tilde{a} + c_2\tilde{a}$, and $c(\tilde{a}_1 + \tilde{a}_2) = c\tilde{a}_1 + c\tilde{a}_2$) , and that it is associative in the sense that $(c_1c_2)\tilde{a} = c_1(c_2\tilde{a})$. An order can be defined in $\tilde{A}$ in the obvious way. In other words $\tilde{A}$ is an ordered abelian semigroup under addition; and, with the scalar multiplication, it has the linear structure of an R^+ - semimodule. Thus for any complete ratio scale, the equivalence classes have a structure just like that which we were able to define directly (i.e., before establishing the ratio scale of length functions) on the set of congruence classes of geometric segments. If the ratio scale is not complete, we have a corresponding, but "incomplete", structure on the set of equivalence classes of the domain.

In much elementary work dealing with measurement, measure functions are established and named in terms of their "units". Sometimes the unit is thought of as a single element of the domain, and sometimes as the equivalence class which contains that element; (i.e., which is uniquely determined by that element). We shall continue to regard a unit as an equivalence class of domain elements, so that the <u>unit of the function</u> $f \in F$ is $\{a : a \in A , f(a) = 1\}$, provided that this set is not empty. This establishes a natural correspondence between functions and units, for those functions which have units. The following result is immediate:

<u>Theorem 4-4.11</u>. The natural correspondence of functions and units is a 1-1 correspondence if and only if the ratio scale is complete.

We remind you that, when F is a complete scale, the structures of F and $\tilde{A}$ are similar (both are ordered R^+ - semimodules) but the correspondence of functions and units is not a structure preserving function. In particular, if $\tilde{a} \longleftrightarrow f$, then $k\tilde{a} \longleftrightarrow \frac{1}{k} f$; if $\tilde{a}_1 \longleftrightarrow f_1$, and $\tilde{a}_2 \longleftrightarrow f_2$, then $\tilde{a}_1 < \tilde{a}_2$ if and only if $f_2 < f_1$; and $\tilde{a}_1 + \tilde{a}_2 \longleftrightarrow \frac{f_1 f_2}{f_1+f_2}$. As we pointed out earlier, if you are familiar with the notion of dual space in linear algebra, you will recognize that (with a simple generalization of the notion of duality as applied to vector spaces) the R^+ - semimodule $(F , + , R^+)$ is dual to $(\tilde{A} , + , R^+)$; and there is a natural 1-1 correspondence from $(\tilde{A} , + , R^+)$ to the dual space of $(F , + , R^+)$.

In the case of complete ratio scales, much of what we have developed concerning the operations on, and the relationships between ratio scales can be developed dually in terms of the structure of the corresponding sets of units. The main advantage in working with functions rather than with units, is that we can take advantage of the well-developed mathematical theory of function composition, and the algebra of real-valued functions. A second advantage is that we can deal easily with incomplete, as well as with complete scales: a ratio scale is an ordered R^+ - semimodule whether or not the "domain" is closed under its operations of addition and scalar multiplication. We have no intention of interpreting all of our previous work in terms of the language of "units", and the structure of the domain, but it is instructive to look at a few particular situations.

Before doing this, we make an important observation. As we have seen, if F is a ratio scale with domain A, we can use the ratio scale F to obtain an equivalence relation on A°, and to obtain an operation of addition and an operation of scalar multiplication for the set $\tilde{A}$ of equivalence classes. ($\tilde{A}$ need not be closed under either operation.) If G is another ratio scale which also has domain A, then G also determines an equivalence relation on $\tilde{A}$, and appropriate operations. It is not necessary that these agree with those which F determines. In the particular case where $G = F^{\alpha}$, and $\alpha \neq 0$, it is easy to prove that the equivalence relations determined by F and G are the same, but the addition and the scalar multiplication are not the same unless $\alpha = 1$.

Units and Homogeneous Functions. In the discussion which follows, all ratio scales are assumed to be complete, and hence the set of domain classes is closed under addition and scalar multiplication. Let F and G be complete ratio scales, with domains A and B, and sets of units (i.e., equivalence classes of domain elements) $\tilde{A}$, $\tilde{B}$ respectively. Let γ be a homogeneous function, $\gamma : F \to G$. That is, there exists $\alpha \in R$, such that for every $f \in F$ and every $k \in R^{+}$, $\gamma(kf) = k^{\alpha}\gamma(f)$. Clearly, using the 1-1 correspondence of functions and units, γ determines a "corresponding" mapping of units. Denote this by $\bar{\gamma} : \tilde{A} \to \tilde{B}$. We can examine the relationship of $\bar{\gamma}$ to the structures of $\tilde{A}$ and $\tilde{B}$. If $f \longleftrightarrow \tilde{a}$, then $kf \longleftrightarrow \frac{1}{k}\tilde{a}$; and if $\gamma(f) \longleftrightarrow \tilde{b} = \bar{\gamma}(\tilde{a})$, then $\gamma(kf) = k^{\alpha}\gamma(f) \longleftrightarrow \frac{1}{k^{\alpha}}\tilde{b} = \bar{\gamma}(\frac{1}{k}\tilde{a})$. That is,

$$\bar{\gamma}(\tfrac{1}{k}\tilde{a}) = (\tfrac{1}{k})^{\alpha}\,\bar{\gamma}(\tilde{a}).$$

But the set of all numbers $\frac{1}{k}$ (for all $k \in R^{+}$) is just R^{+}. Therefore $\bar{\gamma}$ is also homogeneous of degree α; i.e., it satisfies a similar homogeneity condition to γ. The converse is easily shown, hence

Theorem 4-4.12.

(i) Let $\bar{\gamma} : \tilde{A} \to \tilde{B}$ be a unit function, and let $\gamma : F \to G$ be the dual function on the corresponding ratio scales. Then $\bar{\gamma}$ is homogeneous if and only if γ is homogeneous.

(ii) The dimension of a homogeneous scale mapping is the same as the dimension of the associated (dual) homogeneous unit mapping.

We can apply this result to the study of those homogeneous functions on ratio scales which are induced by domain mappings. Let F and G be complete scales with domains A and B respectively, and let $\delta : B \to A$.

Then, without any conditions on δ, we always have an induced ratio scale $\delta^*(F)$ (not necessarily complete) and a homogeneous function of dimension 1, $\delta^* : F \to \delta^*(F)$. If δ induces a homogeneous unit mapping from $\tilde{B}$ to $\tilde{A}$, then it is natural to look for a relationship between $\delta^*(F)$ and G. Clearly, if δ is to induce such a mapping, then δ must map equivalent (as determined by G) elements in B into equivalent (as determined by F) elements in A; (i.e., $b \sim b'$ implies $\delta(b) \sim \delta(b')$). Assume this property, and that δ induces a homogeneous unit mapping $\bar{\delta} : \tilde{B} \to \tilde{A}$ of degree α. Then $\bar{\delta}$ induces a homogeneous scale function δ_* ($\delta_* : G \to F$) of dimension α. It follows that the composite function $\delta^*\delta_* : G \to \delta^*(F)$ is homogeneous with dimension α. Moreover the ratio scales G and $\delta^*(F)$ have the same domain B, so it is not unreasonable to suspect a relationship of $\delta^*(F)$ and G^α. It is a simple matter to verify directly that if $g \in G$, and $\delta_*(g) = f \in F$, then for every $b \in B$, $g^\alpha(b) = [g(b)]^\alpha = (f\delta)(b)$. That is g^α and $f\delta$ are the same function. Hence the following diagram is commutative:

$$\begin{tikzcd}
\tilde{B} \arrow[r, "g"] \arrow[d, "\bar{\delta}"'] & R^+ \arrow[d, "I^\alpha"] \\
\tilde{A} \arrow[r, "f = \delta_*(g)"'] & R^+
\end{tikzcd}$$

It follows that $g^\alpha = \delta^*(f)$. Hence we get

Theorem 4-4.13. With the above notation, if $\delta : B \to A$ induces a homogeneous unit mapping $\bar{\delta} : \tilde{B} \to \tilde{A}$, of dimension α, then $\delta^*(F) = G^\alpha$, and the following diagram is commutative.

$$\begin{tikzcd}
G \arrow[r, "I^\alpha_*(\alpha)"] \arrow[d, "\delta_*(\alpha)"'] & G^\alpha \\
F \arrow[ur, leftrightarrow, "\delta^*(1)"'] &
\end{tikzcd}$$

Example 1. Probably the simplest example of this situation occurs in the relationship of the length scale $(F$, with domain $A)$ for segments and the area scale $(G$, with domain $B)$ for square regions. In this case δ is the function which maps a square b onto its side a. Clearly this induces a mapping $\overline{\delta} : \tilde{B} \to \tilde{A}$ with the property that $\overline{\delta}(k\tilde{b}) = \sqrt{k}\,\overline{\delta}(\tilde{b})$., so that $\overline{\delta}$ and δ_* are homogeneous with dimension $\frac{1}{2}$, and we have $\delta^*(F) = G^{1/2}$. From the general properties of δ^*, this means that $\delta^*(F^2) = [\delta^*(F)]^2 = G$, as was shown earlier. The homogeneous function δ_* is, of course, the inverse of the "standard" function $\eta : F \to G$, (with dimension 2) which we introduced in Section 3-5.

Example 2. Theorem 4-4.13 is directly related to Theorem 4-4.9. For, if F is any ratio scale with domain A, then $F^\alpha \underline{\times} F^\beta$ has domain $A \times A$, and (as you should verify) the diagonal mapping $\Delta : a \to (a, a)$ on the domains induces a homogeneous unit mapping $\overline{\Delta}$, of degree $\alpha + \beta$, from the units of F to th units of $F^\alpha \underline{\times} F^\beta$. (I.e., $\overline{\Delta}(k\tilde{a}) = k^{\alpha+\beta}\,\overline{\Delta}(\tilde{a})$.) If we call the corresponding (dual) homogeneous scale mapping Δ_* $(\Delta_* : F \to F^\alpha \underline{\times} F^\beta)$ then Theorem 4-4.13 tells us (cf. Theorem 4-4.9(iv)) that $\Delta^*(F^\alpha \underline{\times} F^\beta) = F^{\alpha+\beta}$, and that the following diagram is commutative:

$$\begin{tikzcd}
F \arrow[r, "I_*^{\alpha+\beta}(\alpha+\beta)"] \arrow[d, "\Delta_*(\alpha+\beta)"'] & F^{\alpha+\beta} \arrow[dl, leftrightarrow, "\Delta^*(1)"] \\
F^\alpha \underline{\times} F^\beta &
\end{tikzcd}$$

You can easily verify directly that $\Delta_*(f) = \mu(f^\alpha \times f^\beta)$, and therefore the homogeneous function Δ_* is the same as the function which we denoted by $\overline{\Delta}_1$ in Theorem 4-4.9.

Example 3. (Cf. Theorem 4-4.6) Another type of domain mapping which we made use of earlier, was the mapping

$$\delta : \overline{a}_1 \to (a_1, \overline{a}_2, \ldots, \overline{a}_n)$$

where F_i $(i = 1, 2, \ldots, n)$ is a ratio scale with domain A_i, and the element $\bar{a}_j$ $(j \neq 1)$ is a fixed domain element in A_j. Again you may easily verify that if ΠA_i is regarded as the domain of the scale $\underline{\Pi} F_i^{\alpha_i}$, then δ induces a homogeneous unit mapping, of dimension α_1, from the units of F to the units of $\underline{\Pi} F_i^{\alpha_i}$. It follows that if δ_* denotes the corresponding homogeneous scale function of dimension α_1, and if δ^* is, as usual, the homogeneous scale function induced by the pre-composition mapping δ, then (from Theorem 4-4.13) $\delta^*(\underline{\Pi} F_i^{\alpha_i}) = F_1^{\alpha_1}$, and the following diagram is commutative:

$$\begin{tikzcd}
F_1 \arrow[r, "I_*^{\alpha_1}(\alpha_1)"] \arrow[d, "\delta_*(\alpha_1)"'] & F_1^{\alpha_1} \\
\underline{\Pi} F_i^{\alpha_i} \arrow[ur, <->, "\delta^*(1)"'] &
\end{tikzcd}$$

Secondary Quantities Whose Measurement Involves Several Elements in the Same Domain. In the measurement of secondary quantities, it seemed desirable to include the possibility that more than one domain element (of a particular primary scale) might be associated with each domain element of the secondary scale; and, in that case, we saw that we could use a function on ΠR_i^+ which was merely homogeneous in the ordinary sense; (i.e., not necessarily multihomogeneous). A question which you might well have asked was whether this sort of generality was ever needed, or whether the same secondary scale might not be obtained from a different domain mapping, in which only one domain element from each primary domain is used, with the consequence that the corresponding function on ΠR_i^+ must be multihomogeneous; (i.e., some constant multiple of a product of powers). The following example shows that with fairly general assumptions, this is possible, at least in principle.

Example. Consider the simple case of a complete ratio scale F, with domain A, and a homogeneous function $h : R^+ \times R^+ \to R^+$ of degree α $(\alpha \neq 0)$, such that the scale $H = \{h(f \times f) : f \in F\}$ (with domain $A \times A$) is complete. Then the corresponding scale $\bar{H} = \{\bar{h}(f \times f) : f \in F\}$ (where $\bar{h} = h^{1/\alpha}$) is also complete, and $H = \bar{H}^{\alpha}$.

Let $f_0 \in F$, and define $\delta : A \times A \to A$ such that $f_0(\delta(a_1 , a_2)) = \bar{h}(f_0 \times f_0)(a_1 , a_2)$ for every $(a_1 , a_2) \in A \times A$. There is, of course, a high degree of arbitrariness in the choice of $\delta(a_1 , a_2)$ but each set of choices makes the following diagram commutative:

$$\begin{tikzcd}
A \times A \arrow[r, "f_0 \times f_0"] \arrow[d, "\delta"'] & R^+ \times R^+ \arrow[d, "\bar{h}"] \\
A \arrow[r, "f_0"] & R^+
\end{tikzcd}$$

It is easy to prove

(i) this diagram remains commutative if f_0 is replaced by any $f \in F$;

(ii) δ induces a homogeneous unit function $\bar{\delta}$ (of degree 1) on the set of units of H, and a corresponding homogeneous scale function $\delta_* : \bar{H} \to F$;

(iii) $\delta^*(F) = \bar{H}$;

(iv) $\delta^* = (\delta_*)^{-1}$;

(v) $H = \bar{H}^{\alpha} = (\delta^* F)^{\alpha} = \delta^*(F^{\alpha})$.

It follows that if B is the domain of the secondary quantity, and if

$$\delta_1 : b \to (a_1 , a_2)$$

$$\delta_2 : b \to a = \delta(a_1 , a_2)$$

then the secondary quantities (F , B , δ_1 , h) and $(F , B , \delta_2 , I^{\alpha})$ give the same secondary scale, and each has the same dimension (α) in the primary scale F. Thus, in a certain sense, these secondary quantities are "equivalent". This does not mean that it is always practically possible to dispense with multiple domain elements and ordinary homogeneous functions h, but, in principle, it is possible to do so.

As a rather trivial (but concrete) example of this, consider the ratio scale on the domain of rectangles which takes the pair of adjacent sides a_1, a_2, measures their "lengths" $x_1 = f(a_1)$, $x_2 = f(a_2)$, under each length function f, and uses the homogeneous function $h : (x_1, x_2) \to x_1^2 + x_2^2$. (Intuitively, the functions in this scale assign to each rectangle the sum of the "areas" of the squares on its sides.) In this case the domain mapping $\delta : A \times A \to A$, which maps the pair of sides (a_1, a_2) onto the diagonal a of the rectangle, determines a homogeneous unit function; and the fact that, for every $f \in F$, $f^2(a) = [f(a)]^2 = [f(a_1)]^2 + [f(a_2)]^2$ is, of course, Pythagoras' Theorem.

Links With Other Parts of Mathematics. In the present section we have explored the construction of "new" ratio scales by the formation of "powers"; by the formation of "products"; and by pre-composition with domain mappings to yield, roughly speaking, "similar scales but with different domains". In developing these constructions our approach was quite intuitive in the sense that we followed closely the sequence of steps which was suggested by the physical notion of "secondary quantity". We now plan to show you that these constructions are closely related to some of the basic constructions of modern linear and multilinear algebra, and to natural generalizations of these ideas to homogeneous and multihomogeneous functions.

We have already pointed out that a ratio scale has a "linear structure": that is, it has an addition operation, and a scalar product operation (by positive real numbers); and these operations are related to one another like the corresponding operations in vector spaces and ring modules, the main difference being that the ratio scale is only a semigroup under its addition operation. We called this type of structure an R^+-semimodule.

We also saw that the domain has (or can be given) a corresponding "dual" structure on the set of "units" (i.e., the set of equivalence classes of domain elements under the equivalence relation which the scale determines). If the ratio scale is complete, this dual structure is also "complete" in the sense that the system of "units" is closed under the operations of addition and scalar multiplication of "units"; that is, the "unit" structure is also an R^+-semimodule.

The product operation which we defined for ratio scales is naturally related to the so-called "tensor product" of linear spaces. This product operation on the set of all linear spaces over the same scalar domain (e.g., a ring, or a field) is connected with older ideas concerning tensors; it has become very important in modern algebra and topology. If F and G are R^+-semimodules (e.g., ratio scales) it is a simple matter to generalize the standard tensor product construction (see, for example, [7], 3rd edition, 1965) so as to give a tensor product operation for the R^+-semimodules F and G, leading to the tensor product of F and G. We denote this by the usual $F \otimes G$.

There are several equivalent ways of defining this tensor product. For example, the definition given in [7] for vector spaces can be directly generalized as follows:

> If F and G are R^+-semimodules, denote by $B(F, G; R^+)$ the set of all bilinear functions from $F \times_0 G$ to R^+. It is easy to show that this set is also an R^+-semimodule under the natural operations of addition of functions and scalar multiplication of functions by positive real numbers. The tensor product, $F \otimes G$, is now defined to be the R^+-semimodule which is the dual space of $B(F, G; R^+)$, in the usual sense of linear algebra. [That is, $F \otimes G$ is the space of linear functions (or functionals) from $B(F, G; R^+)$ to R^+.]

The standard procedure for defining a natural mapping of a vector space into the dual of its dual, can be generalized to define a natural mapping from $F \times G$ (cartesian product) to $F \otimes G$: If $(f, g) \in F \times G$, define

$$\otimes: F \times G \to F \otimes G$$

by,

$$\otimes: (f, g) \to f \otimes g$$

where

$$f \otimes g : B(F, G; R^+) \to R^+$$

is the linear functional on $B(F, G; R^+)$, which is defined by

$$f \otimes g : b \to b(f, g),$$

for each bilinear function $b \in B(F, G; R^+)$.

It can be proved that $\otimes$ is also a bilinear function. We refer to $f \otimes g$ as the element of the tensor product which is determined by f and g. The function $\otimes$ is, of course, not 1-1.

The tensor product can be defined in other ways, leading to "tensor products" which are naturally isomorphic to the one which we have defined. One common procedure defines the tensor product of F and G to be the R^{+}-semimodule generated by the symbols $f \otimes g$ $(f \in F, g \in G)$, subject to the relations,

$$(f_1 + f_2) \otimes g = f_1 \otimes g + f_2 \otimes g ;$$
$$f \otimes (g_1 + g_2) = f \otimes g_1 + f \otimes g_2 ;$$
$$(kf) \otimes g = f \otimes kg .$$

The tensor product of F and G is naturally isomorphic to the product scale $F \underline{\times} G$ which we have defined for ratio scales, under the 1-1 correspondence

$$j : f \otimes g \longleftrightarrow \mu(f \times g) ,$$

where $f \otimes g$ denotes the element of the tensor product which is naturally determined by f and g. The reason for this is quite simple, and relates to the motivation for the tensor product construction in relation to bilinear (and multilinear) functions:

The function

$$\rho : F \times G \to F \underline{\times} G$$

defined by

$$\rho : (f, g) \to \mu(f \times g)$$

is easily seen to be bilinear; the tensor product $F \otimes G$ is a "universal space" for all bilinear functions on $F \times G$ in the sense that any such bilinear function (such as ρ) can be "factored" through $F \otimes G$, as the composite of the bilinear function $\otimes$, and a linear function. Thus there exists a linear function

$$j : F \otimes G \to F \underline{\times} G$$

which makes the following diagram commutative:

$$\begin{tikzcd}
 & F \otimes G \arrow[d, "j"] \\
F \times G \arrow[ur, "\otimes"] \arrow[r, "\rho"'] & F \underline{\times} G
\end{tikzcd}$$

The function j is easily verified to be 1-1, so that j is a linear isomorphism of the R^+-semimodules $F \otimes G$ and $F \underline{\times} G$. The essential distinction between $F \otimes G$ and $F \underline{\times} G$ is that $F \otimes G$ is the result of a formal construction on F and G, which depends only on the fact that F and G are R^+-semimodules and does not depend on the fact that the elements of F and G are functions; on the other hand, in the construction of $F \underline{\times} G$ we made use of the fact that the elements of F and G were functions, to describe related functions as elements of $F \underline{\times} G$. Thus, while $F \underline{\times} G$ is, essentially, the tensor product of F and G, we have in addition an interpretation for the elements of $F \underline{\times} G$.

If F and G are complete ratio scales we can establish the isomorphism of $F \otimes G$ and $F \underline{\times} G$ in another way. In this case F and G determine associated domain structures $(\tilde{A}, +, R^+)$, $(\tilde{B}, +, R^+)$ which are R^+-semimodules, and F and G are, of course, the respective sets of linear functionals on $\tilde{A}$ and $\tilde{B}$. That is, $F = \tilde{A}^*$, and $G = \tilde{B}^*$, where "*" denotes, as usual, the dual space. In this case the product scale $F \underline{\times} G$ can be shown (directly) to be the space (an R^+-semimodule) of all bilinear functions from $\tilde{A} \times \tilde{B}$ to R^+. Hence, from the definition of tensor product, $\tilde{A} \otimes \tilde{B} = (F \underline{\times} G)^*$. It follows from the basic duality theorem of linear algebra, that

$$(\tilde{A} \otimes \tilde{B})^* = (F \underline{\times} G)^{**} \approx F \underline{\times} G$$

But it is a standard theorem (easily proved) in tensor product theory, that there is a natural isomorphism

$$(\tilde{A} \otimes \tilde{B})^* \approx \tilde{A}^* \otimes \tilde{B}^* .$$

Hence

$$F \underline{\times} G \approx \tilde{A}^* \otimes \tilde{B}^* = F \otimes G .$$

The tensor product operation and the product operation for ratio scales can, of course, be iterated, and all of the above results can be correspondingly extended in relation to multilinear functions.

If, F and G are complete ratio scales with domains A and B, then the product scale $F \underline{\times} G$ determines a "unit" structure in its domain $A \times B$. It follows from the above discussion that this structure is, in essence, the tensor product of the structures determined in $\tilde{A}$ and $\tilde{B}$. Because of this duality, we can interpret the familiar L, T, A, V, M, etc., used by scientists either as representing the domain structure, or as representing the corresponding ratio scale structure; and, in either case, a "product" such as LT^{-1} is naturally isomorphic to the corresponding tensor product $L \otimes T^{-1}$.

This should prompt you to ask: What about the "T^{-1}"? Does this have an interpretation as a "power" of T, whether T is regarded as representing either a domain structure (or the corresponding scale structure) or in general, for any R^+-semimodule? The answer to this question is "yes". We can discuss the whole question of "powers" in relation to homogeneous and multihomogeneous functions (and corresponding "generalized tensor products") much as we discussed the relationship of tensor products to linear and multilinear functions.

As a starting point for such a discussion, we point out that if F is a complete ratio scale with domain A, and with a corresponding "domain structure" $(\bar{A}, +, R^+)$, then, for any $\alpha \in R$, the power scale F^α which we defined earlier, is (if $\bar{A}$ is regarded as the domain) just the set of all homogeneous functions of degree α from $\bar{A}$ to R^+. In general, the set of all homogeneous functions of fixed degree on any R^+-semimodule is easily shown to be an R^+-semimodule, whose operations are, of course, ordinary functional addition and scalar multiplication. The fact that we were able to define the concept of "homogeneous function" (of any degree) for ratio scales and their domains, depended strongly on properties of R^+; however, whenever we have a linear space F on which a suitable concept of homogeneous function can be defined, the set $H_\alpha(F)$ of homogeneous functions of degree α is a linear space (under the usual operations of functional addition and scalar multiplication). If F is a ratio scale, the dual space $\{H_\alpha(F)]^*$ is naturally isomorphic to the scale which we have already designated as F^α. It is therefore convenient to designate $[H_\alpha(F)\}^*$ by $\underline{F}^\alpha$, and refer to these spaces as "generalized powers".

Generalized powers and tensor products can, of course, be combined. Moreover it is relatively easy to define a sort of "generalized tensor product": For R^+-semimodules F_1 and F_2, we define $F_1 \underset{(\alpha_1, \alpha_2)}{\otimes} F_2$ to be the dual space of the space of bihomogeneous functions of degree (α_1, α_2) from $F_1 \times F_2$ to R^+. If F_1 and F_2 are ratio scales, there is a natural isomorphism of $F_1 \underset{(\alpha_1, \alpha_2)}{\otimes} F_2$ and $\underline{F}_1^{\alpha_1} \otimes \underline{F}_2^{\alpha_2}$. If F_1 and F_2 are complete ratio scales with associated domain structures

$\mathfrak{X}_1$ and $\mathfrak{X}_2$, then $F_1 \underset{(\alpha_1, \alpha_2)}{\otimes} F_2$ is naturally isomorphic to the "power-product" $F_1^{\alpha_1} \underline{\times} F_2^{\alpha_2}$; and this is just the space of bihomogeneous functions of degree (α_1, α_2) on $\mathfrak{X}_1 \times \mathfrak{X}_2$.

There are many interesting relationships among the various operations which we have sketched above. In the case of ratio scales the situation is especially simple, because of the fact that we are dealing with the simplest possible kind of R^+-semimodule, one which is "1-dimensional" in a linear sense. Thus some of the results stated above cannot be directly generalized.

We do not intend to pursue these ideas any further in this book. Our only reason for even mentioning them was to show you that the constructions which we appeared to invent in response to the motivation of measurement ideas are, in reality, just special cases of some very general types of construction on linear spaces. And these constructions (whether applied to the ratio scales or to the corresponding domain structures) seem to provide the best abstract framework in which to interpret the frequently mysterious symbolic operations of the physical sciences. For example, the manipulation of "units" which usually accompanies calculations of the values of "derived" quantities when the "units" are changed in the "fundamental" quantities, can be interpreted in terms of the algebraic properties of tensor products and their generalizations, and in terms of the conventions (isomorphisms) by means of which we normally relate the tensor products to the derived scales. Thus (to take a very simple example) assume that the area scale is related in the usual way to the tensor product of the length scale with itself, so that (with "ft" and "in" denoting domain classes in the length scale, and "ft^2", "in^2", domain classes in the area scale), $\text{ft}^2 \longleftrightarrow \text{ft} \otimes \text{ft}$; $\text{in}^2 \longleftrightarrow \text{in} \otimes \text{in}$; etc. Then

$$\begin{aligned} 10\ \text{ft}^2 &\longleftrightarrow 10\ (\text{ft} \otimes \text{ft}) \\ &= 10\ [(12\ \text{in}) \otimes (12\ \text{in})] \\ &= 1440\ (\text{in} \otimes \text{in}) \\ &\longleftrightarrow 1440\ \text{in}^2 . \end{aligned}$$

This is a reasonable place to end this very long section. By combining the various operations on ratio scales we can generate an enormous variety of scales. Which particular scales are significant in relation to empirical measurement, and what are the significant relationships between these scales, will, in general, be suggested by empirical evidence. This is not a question that can be answered in the abstract.

4-5 Formulas, Dimensions, and Related Topics.

~~In~~ Section 4-3 we saw how the notion of a "secondary quantity" implied the existence of a set of similar "measure" functions on the domain of the "secondary quantity", and that this set of functions determined a unique ratio scale. We also saw that, with the additional assumption of a monotonicity (or, alternately, continuity) condition, the secondary scale had to be related in a rather specific way to the primary scales involved in its description. This led, in Section 4-4 to the study of those ratio scales which could be derived from other ratio scales by the procedures which were suggested by the relationship of "primary" and "secondary" quantities, and to the study of homogeneous functions between ratio scales. In spite of the fact that only a few fairly simple basic ideas were involved, it became quite complicated to keep detailed track of the various scales and homogeneous functions, and it is fairly clear that if we are to make any practical use of these ideas in more complicated situations, we should try to find a way of grasping the essentials, without getting bogged down in details. Fortunately this is not too difficult from a mathematical point of view, but of course, in scientific situations, the assumptions made will have to be justified by empirical evidence.

In the following discussion it will be useful to make use of simple physical examples, as well as examples from mathematics. We therefore need to introduce some additional ideas concerning physical "quantities". In particular, we shall want to consider the notions of mass, and time-interval, which, together with length, are generally taken as the "primary" quantities from which the various "secondary" quantities involved in classical (i.e., Newtonian) mechanics (area, volume, density, force, velocity, etc.) are derived. You will find these ideas discussed in [2] (and, of course, in many other books on mechanics and physics). From our point of view, we have to make certain basic assumptions. These are, in effect, that each of these "quantities" has a domain; that, corresponding to each "quantity", there is a set of functions from the relevant domain to the positive reals; and that each set of functions is a ratio scale. These functions are established by measurement operations, and different operations might lead to the same ratio scale. It is sometimes argued that identical scales should be considered to be different if they are obtained by different measurement procedures, but this would create considerable difficulties from the mathematical point of view: a ratio scale for empirical measurement would then have to be defined not merely as a set of functions, but as a set of functions together with a clearly defined empirical process (a set of operational procedures) which "generates" the function; and such a process, even if it could be completely and unambiguously defined, is not an idea that is readily susceptible to mathematical treatment. Our definition of secondary quantity in Section 4-3 takes care of some of these problems, but of course a secondary quantity, as we have defined it, is not a ratio scale: it merely determines one.

Derived Scales From a Less-Detailed Standpoint. We recall the earlier discussion of secondary, or derived scales. Assume that $F_1 , F_2 , \ldots , F_n$ are different ratio scales, with domains $A_1 , A_2 , \ldots , A_n$ respectively. Let B be a set of "objects", and suppose that to each element of B there is associated in a unique way a collection of objects from the domains A_i. (It is not necessary that there be exactly one object from each domain; i.e., we do not insist that the associated set of objects belong to the simple cartesian product ΠA_i, of the domains A_i.) Let $f_i \in F_i$ $(i=1, 2, \ldots, n)$. Let the domain elements in the A_i be "measured" by the respective f_i, and let the values be combined in some well-defined way to give a positive real number. If this is done for each $b \in B$, we obtain a function (g, say) from B to R^+. Clearly this function depends on a number of things, including

(i) the rule δ (which could be expressed as a function on B with values in a suitable domain) by which a finite number of domain elements $a_i , a_i' , \ldots$, from each A_i $(i=1, 2, \ldots, n)$ are associated with each $b \in B$;

(ii) the choice of $f_i \in F_i$, $(i=1, 2, \ldots, n)$;

(iii) the rule φ for combining the measurements $f_i(a_i)$, $f_i(a_i')$, ..., $(i=1, 2, \ldots, n)$ to give a function $\varphi : \Pi R^+ \to R^+$. (We do not want to be more specific here concerning this function.)

Under suitable assumptions, which we examined in detail in the preceding sections, all such functions g are similar, and hence they determine a unique ratio scale G, which is related in a specific way (depending on δ and φ) to the scales F_i. In this case the function Φ on the cartesian product of the sets F_i (not on the product scale ΠF_i)

$$\Phi : (F_1 \times F_2 \times \ldots \times F_n) \to G$$

defined by $\Phi : (f_1, f_2, \ldots, f_n) \to g$, is multihomogeneous. That is, if $k_1, k_2, \ldots, k_n$ are any positive real numbers, then there exist unique real numbers α_i, such that for all choices of $k_1, k_2, \ldots, k_n$,

$$\Phi : (k_1 f_1, k_2 f_2, \ldots, k_n f_n) \to k_1^{\alpha_1} \cdot k_2^{\alpha_2} \ldots k_n^{\alpha_n} g \, .$$

If Φ is multihomogeneous, the ratio scale G is said to have dimension α_i in F_i $(i=1, 2, \ldots, n)$, <u>with respect to the rules δ and φ</u>. If δ is single valued in each A_i (i.e., $\delta(b) \in \Pi A_i$) then G is, of course, the scale $\delta^* \Pi F_i^{\alpha_i}$. In any event, for each choice of functions $\overline{f}_j$ in F_j $(j \neq i)$, δ and φ (and hence Φ) determine a homogeneous function (Φ_i, say) of dimension α_i, from the scale F_i to the scale G.

We emphasize the fact that the "dimensionality" of G is not an absolute property of G, but depends on the rules δ and φ used in the derivation of G. In other words, the dimensionality is determined by the secondary quantity, and not by the secondary scale alone. It is important in the establishment of scientific systems of measurement, to adopt clear and unambiguous conventions regarding the rules δ and φ. When this is done it is possible to refer to the "dimensions of G", with the understanding that we mean the dimensions of G with respect to specific F_i, and fixed rules δ and φ.

When the primary scales F_i are fixed, it is sometimes convenient to regard a secondary quantity as being specified by its secondary scale (G say), and by the multihomogeneous function $\Phi : \Pi F_i \to G$ on the cartesian product ΠF_i. Thus we shall sometimes refer to a "secondary quantity $\{G, \Phi\}$".

If $G_0 (\subseteq G)$ denotes the set of similar functions actually obtained from the definition of a secondary quantity (i.e., G_0 is the range of Φ) then we can distinguish several possibilities:

(i) G_0 is a complete ratio scale. (This implies not all $\alpha_i = 0$.)

(ii) G_0 is an incomplete ratio scale. (That is, $g \in G_0$ implies that $kg \in G_0$ for every $k \in R^+$, but g is not onto R^+ . This implies not all $\alpha_i = 0$.)

(iii) Each $g \in G_0$ is the same function. (This implies that each $\alpha_i = 0$. Such a secondary quantity is often called dimensionless.)

We have already seen numerous examples of (i), in the genesis of "derived" scales for area and volume. An example of (ii) occurred when we were considering the "measurement" of simple angles by using the arc length of a fixed circle. We saw an example of (iii) in our treatment of radian measure. Case (ii) with each function a constant function, is connected with the question of "dimensional constants". From our point of view a dimensional constant is best treated as the simplest possible kind of secondary quantity: one whose associated functions form a ratio scale in which each function has a constant value (i.e., all domain elements are equivalent, so that, in effect, there is only one equivalence class, b_0 say). Such a dimensional constant is usually specified by giving its value $g(b_0)$ for a particular choice of functions $\{f_i\}$, and its dimension α_i in each F_i : Not all α_i are zero; and the value for any other choice of functions $(\{k_i f_i\}$ say) is easily calculated as $\prod k_i^{\alpha_i} g(b_0)$. If the set of secondary functions $(G_0)'$ contains only a single constant function, we might regard this dimensionless secondary quantity (or its unique value) as an absolute constant. From this point of view an absolute constant might be regarded as a dimensional constant with dimension zero in each primary quantity. (E.g., the well-known secondary quantity: "length of circumference divided by length of diameter", whose domain is the set of all circles, and whose single constant value is the number π .)

In a particular context (such as in mechanics) a system of measurement consists of a choice of a particular set of functions, one from each of the primary scales involved (e.g., foot-pound-second (FPS); or centimeter-gram-second (CGS)) together with a complete description of each of the secondary scales involved, as the scale derived from a particular secondary quantity. Thus for each secondary scale in the system we have a specific multihomogeneous function Φ from the cartesian product of the primary scales to the secondary scale, and a corresponding homogeneous function from each primary scale to each secondary scale. The selection of a particular function in each primary scale determines a particular secondary function in each secondary scale. Sometimes a system of measurement is regarded as being merely the resulting set of primary and secondary functions, (or some similar set in which the secondary functions are selected as some particular constant multiple of the "natural" selections) but, by itself, this is not enough if we are interested in dimensional questions: the homogeneous functions (from each primary scale to each secondary scale) which are determined by the definitions of the secondary quantities, are not determined by their values on only one element of the primary scale. If, in addition to specifying a particular function from each secondary scale, we specify the dimension of each secondary quantity in each primary quantity, then the homogeneous scale functions are fully determined, and we may regard this situation as giving an adequate description of the associated "system of measurement".

Example. As a simple example involving a dimensional constant, let us consider the classical experiment of Galileo, in which he determined the "law" for falling bodies. As a result of the measurement of the distance fallen (from rest) and the elapsed time of free fall, Galileo discovered that the number which "measured" distance fallen appeared to be proportional to the square of the number which "measured" the elapsed time. If B is the set of "experiments", then we can associate with each experiment b, a pair of elements a_1, a_2, in the domains A_1, A_2, of the length scale and the time-interval scale respectively. Galileo discovered that, for every experiment b (with different objects dropped, but with the same functions f_1, f_2, measuring length and time-intervals) the function $g: B \to R^+$ defined by $g(b) = \dfrac{f_1(a_1)}{[f_2(a_2)]^2}$ had the same value, c, say. Let Φ be the associated function $\Phi : (f_1, f_2) \to g$, and let $G = \{\Phi(f_1, f_2)\}$.

Without carrying out any more experiments, we can use the assumptions that length measurement and time-interval measurement functions form ratio scales, and we see that if F_1, F_2, are the ratio scales for length and time-interval measurement, and if $f_1' = k_1 f_1 \in F_1$, $f_2' = k_2 f_2 \in F_2$, then

$$\Phi(f_1', f_2')(b) = g'(b) = \frac{f_1'(a_1)}{[f_2'(a_2)]^2} = \frac{k_1 f_1(a)}{k_2^2[f_2(a)]^2} = k_1 k_2^{-2} c \quad \text{for every } b \in B .$$

Thus each g function is a constant function on the domain of experiments B, but the constant depends on the choices of f_1, f_2. The set G of all such functions, g, is a ratio scale, and we say that the corresponding secondary quantity is a <u>dimensional</u> <u>constant</u>, with dimension 1 in length and dimension -2 in time. We observe that c is not an "absolute" constant, but varies with the choice of f_1 and f_2; it is only constant in a particular system of measurement.

The dimensional constant of this simple example is, of course, directly related to the so-called "acceleration due to gravity", and the set of functions G is, in a sense, a ratio scale for the measurement of this acceleration at a particular place. Moreover the whole of our discussion involves a number of oversimplifications, which can be taken care of in a more refined approach to this simple situation. In such an approach the domain of G has to be reinterpreted and extended; and the functions on the extended domain are no longer constant functions, but they still make up a ratio scale.

You might have been concerned that the domain of G, in the above example, was rather vaguely suggested to be a "set of experiments". If we wished to be more precise, we could have taken the corresponding set of pairs (a_1, a_2) (a distance and a time-interval) as the domain. This subset of $A_1 \times A_2$ is a function; (or, if you prefer, the subset determines a function: this depends on which definition of function we are using). The relationship of a_1 and a_2 is a 1-1 correspondence, hence the set of pairs (a_2, a_1) is also a function. Thus we might think of the domain of G as the functional relationship between distance fallen and elapsed time, a functional relationship which exists independently of any question of measurement. If $f_1 \in F_1$, $f_2 \in F_2$, then the set of ordered pairs $\{(f_2(a_2), f_1(a_1))\}$ is also a function, whose domain and range are contained in R^+. This function, which depends on f_1 and f_2, can be expressed by the "formula" $s = ct^2$, where $s = f_1(a_1)$ is the number which "measures" the distance fallen from rest in the time interval a_2 (whose measure under f_2 is $t = f_2(a_2)$) and c depends on f_1 and f_2. [If you are familiar with calculus methods in mechanics, you will know that the measure of the acceleration (in appropriate "units") is the second derivative of s with respect to t, and is therefore $2c$.]

Dimensional Constants and Homogeneous Functions on Ratio Scales. In the previous section we mentioned that a dimensional constant may be associated with a homogeneous function (with dimension $\neq 1$) from a ratio scale to itself. We can easily tie this in with the above example, as follows: Let γ be any homogeneous function of degree 1 from F_1 to F_2 (e.g., define γ ("foot") = "second", and extend in the only possible way so as to give γ dimension 1). Then the experiment of Galileo may be considered as defining another homogeneous function from F_1 to F_2 as follows: If $f_1 \in F_1$, let $\gamma' : f_1 \to f_2$, where $f_2 \in F_2$ is the unique function in F_2 such that for $(a_1, a_2) \in B$, $f_1(a_1) = [f_2(a_2)]^2$. The result of Galileo's experiment shows that γ' does not depend on the particular element $(a_1, a_2) \in B$, which we use in its definition. Moreover, for each such pair of functions $(f_1, f_2 = \gamma'(f_1))$ the corresponding value of the dimensional constant is 1. If $(a_1, a_2) \in B$, and f_1, f_2 are the functions which have a_1, a_2 as units, then $f_2 = \gamma'(f_1)$; that is γ' corresponds to the unit mapping $a_1 \to a_2$, for those functions which have units within the domain of the experiment. That this unit mapping is homogeneous is of course, part of Galileo's discovery. It is easily verified that γ' is a homogeneous function of degree $\frac{1}{2}$ from F_1 to F_2, so the composite function $\gamma' \gamma^{-1}$ is a homogeneous function from F_2 to F_2, with degree $\frac{1}{2}$, and $\gamma^{-1} \gamma'$ is a homogeneous function from F_1 to F_1, with degree $\frac{1}{2}$.

With this example in mind, we can now take another look at the relationship of length and area for polygonal regions. Let f be a length function, let g be the area function whose unit is the square region whose side is the unit of f (in the terminology defined in Section 3-5, $g = \eta(f)$), and let g_σ be the area function whose unit is (determined by) the rectangular region whose sides are congruent to the unit of f, and to a fixed segment σ, respectively. It is easy to prove (in the context of the area theory developed in Chapter 3) that for every polygonal region b,

$$g(b) = cg_\sigma(b)$$

where $c = f(\sigma)$. That is, c depends only on f. We can put this in the form of a dimensional constant by defining $\Phi(f) : B \to R^+$ by

$$[\Phi(f)](b) = \frac{g(b)}{g_\sigma(b)} = f(\sigma) .$$

For each choice of f, the corresponding function $\Phi(f)$ is a constant function on the set B of polygonal regions, and the resulting dimensional constant has dimension 1 in length.

The example of Section 3-7, concerning the measurement of volumes by fluid displacement in a fixed cylinder, can be similarly regarded. In this case the relevant dimensional constant has dimension 2 in length. It determines a ratio scale whose constant values are, in effect, the areas of the constant cross-section under the relevant area functions.

There are many facets to the simple question of dimensional constants, and they may appear in different ways in what is essentially the same context. For example, whenever we have a fixed set of primary scales, a dimensional constant $\{H , \Phi\}$, and another secondary quantity $\{G , \Phi'\}$, we can derive a third secondary quantity whose dimensions are the respective sums of those of H and G, and whose domain corresponds naturally to that of G. More specifically, let $i=1, 2, \ldots, n$, let $\{F_i\}$ be the fixed primary scales, let $\Phi : \Pi F_i \to H$ have dimension α_i in F_i, and let $\Phi' : \Pi F_i \to G$ have dimension α'_i in F_i. Let B be the domain of G, and let σ be an element in the domain of the dimensional constant. (As the functions in H are constant functions, all domain elements are equivalent.) The product $H \underline{\times} G$ of the scales H and G has domain elements (σ , b), and there is a natural 1-1 correspondence $((\sigma , b) \longleftrightarrow b)$ of the domains of $H \underline{\times} G$ and G. Moreover we can combine the functions Φ and Φ' to obtain a function $\Phi'' : F_i \to H \underline{\times} G$, defined by $\Phi'' : \Pi f_i \to \mu([\Phi(\Pi f_i)] \times [\Phi'(\Pi f_i)])$ whose dimension in F_i is $\alpha_i + \alpha'_i$ $(i=1, 2, \ldots, n)$. That is, $\{H \underline{\times} G , \Phi''\}$ is a secondary quantity whose ratio scale is virtually the same as that of G, (because the domains correspond 1-1, and $\Phi(\Pi f_i)$ is constant for each choice of $\{f_i\}$) but whose dimensions are quite different. You can easily apply this to our overworked area example, where H is the length scale on the domain $\{\sigma\}$ whose single element is the segment σ, F is the length scale for segments, G is the area scale for polygonal regions, $\Phi : f \to f \mid \{\sigma\}$ for every $f \in F$, and Φ' is the homogeneous function $\eta_\sigma : F \to G$ (of dimension 1) which is determined by σ as in Section 3-5. If Φ'' is combined with the homogeneous function (of dimension 1) induced by the natural domain correspondence $(\sigma , b) \longleftrightarrow b$, we obtain the homogeneous function $\eta : F \to G$, of dimension 2, by which the length and area scales are conventionally related.

Products of Secondary Quantities. The process used in the above example for deriving a third secondary quantity $\{H \underline{\times} G , \Phi''\}$ as the product of the secondary quantities H and G can be applied whether or not H is a dimensional constant, but, of course, in the general case the domain of $\{H \underline{\times} G , \Phi''\}$ does not have such a simple relationship to the domain of G. We shall not go into this in detail. The important property from a dimensional point of view is that for each i, the dimension of $\{H \underline{\times} G , \Phi''\}$ in F_i, is the sum of the corresponding dimensions of $\{H , \Phi\}$ and $\{G , \Phi'\}$.

Formulas. The word "formula" is used in so many ways in mathematics and science that there is no point in trying to give a general definition of the word. What we are interested in here, are formulas which are statements about numbers (generally positive real numbers) some or all of which are the values of measure functions. (We include, naturally, the numbers obtained from empirical measurements in this description.) Moreover we are not generally interested in isolated statements (e.g., "Bill's height in inches is 73"; or "the length of this box in inches is 4.37 times its weight in pounds". Our concern is rather with statements which themselves have a "domain of validity" (which, in most cases, is indicated by an accompanying verbal description) which contains more than an isolated element.

Let us consider in detail the simple and well-known formula for the area of a triangle. Often this is expressed as

> "$A = \frac{1}{2}bh$, where A denotes the area of a triangle, and b, h, are its base and height."

Sometimes this is amplified by such phrases as: "the formula for the area of any triangle is - - -"; and "- - - when the length and area are measured in corresponding units"; but it is rare to find specified all of the conditions under which the formula holds. Provided that all of these conditions are clearly understood, there is no harm in taking such notational shortcuts, but it might be worthwhile to look at this formula in detail, to see just what it involves.

First of all, we might ask: what is such a formula from a mathematical point of view? Clearly, at the simplest level, it is a symbolic statement, about numbers, accompanied by a verbal statement which gives the source of the numbers: the numbers are the values of certain measure functions. Which measure functions? One is an area function, and one is a length function, and these are related in a specific way. Is it a statement about one measure function and one area function? No; any length function may be used, but then the area function must be suitably related to the length function; in fact this relationship involves the particular "standard" homogeneous function η of dimension 2 from the length scale (for segments, say) to the area scale (for polygonal regions, say), which we defined in Chapter 3. (That is, η maps a given length function onto that area function whose unit is the square region whose sides are congruent to the unit of the given length function.) What are the domain elements whose values under a chosen length function and the related area function are the numbers in the formula? The domain element for the area function is any triangular region; and the domain elements for the length function are any side of the triangle as "base", and the corresponding "altitude". (We use "altitude" here to denote the relevant segment, and not its measure under a particular length function; both uses are of course well established.)

We could collect all of the above answers together, to give a complete statement of the result which is conveyed by the common formula $A = \frac{1}{2}$, but we shall not bother to do this. Instead, we suggest another line of thought: if the formula is to be valid for the whole domain of all triangular regions and related pairs of segments, using any pairs of related length and area functions, is it not possible to express the same result as a relation between functions? The answer is, of course, "yes", but the formula does not express such a relationship as it stands.

In order to give the corresponding functional relationship we introduce some additional terminology. Let B be the set of all triangular regions, and A the set of all segments. (We are going to drop the suggestive form "$A = \frac{1}{2}bh$" of the formula, in favor of "$y = \frac{1}{2}x_1x_2$", so we won't need the letter "A" to denote a number any more.) Let δ be a function from B to $A \times A$ which associates with each $b \in B$ an ordered pair (a_1, a_2) consisting of any side $a_1 \in A$, and the corresponding altitude $a_2 \in A$. Let η be (as above) the standard homogeneous function of dimension 2 from the set F of all length functions for A, to the set G of all area functions for B. Let $f \in F$, let $\eta(f) = g \in G$, and let $f(a_1) = x_1$, $f(a_2) = x_2$, $g(b) = y$. Then, in this terminology, the usual formula which relates the area of a triangle with the lengths of its base and altitude, is

$$y = \frac{1}{2}x_1x_2 .$$

That is, if $\delta(b) = (a_1, a_2)$, then, for every $b \in B$, $[\eta(f)](b) = \frac{1}{2} f(a_1) f(a_2)$. It follows that (with the notation of the previous section for functions in a product of ratio scales)

$$\eta(f) = \frac{1}{2} \mu(f \times f)\delta$$

for every $f \in F$.

The domain of the function which appears on each side of this equation is the set of all triangular regions. This functional relationship, which is the essential content of the so-called area formula for a triangle, can be pictured in the commutative diagram

$$\begin{tikzcd}
B \arrow[r, "\eta(f)"] \arrow[d, "\delta"'] & R^+ & \\
A \times A \arrow[r, "f \times f"] & R^+ \times R^+ \arrow[r, "\mu"] & R^+ \arrow[ul, "\frac{1}{2}"']
\end{tikzcd}$$

The validity of the formula, in functional form, is equivalent to a statement that this diagram is commutative. This relationship is equivalent to

$$\eta(f) = \frac{1}{2} \delta^*(\mu(f \times f)) = \frac{1}{2} \delta^* \bar{\Delta}(f) \; ;$$

where $\bar{\Delta}$ is the "diagonal" homogeneous function $f \to \mu(f \times f)$. This relationship holds for every f in the length scale F , and hence it can be reduced to

$$\eta = \frac{1}{2} \delta^* \bar{\Delta} \, .$$

In other words, the well known simple area "formula" for a triangular region is equivalent to the statement that the homogeneous functions η and $\frac{1}{2} \delta^* . \bar{\Delta}$, from the length scale to the area scale for triangular regions, are just the same function. This can be pictured in the following commutative diagram, in which all of the functions shown are homogeneous functions on the indicated ratio scales, with dimensions as shown:

$$\begin{tikzcd}
 & G \\
F \arrow[ur, leftrightarrow, "\eta(2)"] \arrow[dr, leftrightarrow, "\overline{\Delta}(2)"'] & \\
 & F \underline{\times} F \arrow[uu, leftrightarrow, "\frac{1}{2}\delta^*(1)"']
\end{tikzcd}$$

The fact that η has dimension 2 was shown in Chapter 3, and the dimensionality of the diagonal function $\overline{\Delta}$, and of the function δ^* (and hence of $\frac{1}{2}\delta^*$, which differs from δ^* by a "constant" homogeneous function of dimension 1) was thoroughly discussed in the last section. We check easily that the dimensions satisfy the necessary condition for commutativity: namely, that $\dim \eta = (\dim \overline{\Delta})(\dim(\frac{1}{2}\delta^*)) = 2$.

This "checking of dimensions" of formulas is a useful device, provided that it is clearly understood what the formula means. This includes an understanding of the implicit, as well as the explicit, statements which accompany the formula. If we revert to the numerical form of the formula

$$y = \frac{1}{2}x_1x_2$$

it is often stated that "y has dimension 2 in length, x_1, x_2 each have dimension 1 in length, and $\frac{1}{2}$ is dimensionless, hence the equation satisfies the necessary condition that the sum of the dimensions on each side is the same". From our point of view this sort of thing must be considered merely as an abbreviation of something like what we have spelled out above in considerable detail, at least up to the point where we had $\eta(f) = \frac{1}{2}\mu(f \times f)\delta$; each side of this equation, considered as a function of f, has dimension 2 in f.

The detail which we have gone through above in order to exhibit this particular area formula as equivalent to a commutativity statement concerning homogeneous functions on ratio scales, will not always be possible or desirable. Moreover as indicated earlier in this section, this amount of detail is not necessary if we are merely concerned with the checking of dimensions. For if we look at the statement $y = \frac{1}{2}x_1x_2$, we can consider directly the way in which these numbers change when a different length function $f' = kf$, and the related area function $\eta(f') = k^2\eta(f)$, are used. [This relationship, remember is not just an absolute fact: it is a consequence of the definition of the homogeneous function η, and was proved in Chapter 3; and the fact that we are using η to relate F and G is one of the assumptions under which the formula is valid.] With these related changes in f and $\eta(f)$, and with the same domain elements b, (a_1, a_2), we get $\eta(f')(b) = k^2\,\eta(f)(b) = k^2y$, and $\frac{1}{2}\, f'(a_1) \cdot f'(a_2) = \frac{1}{2}\,(kf)(a_1) \cdot (kf)(a_2) = k^2\,\frac{1}{2}\,x_1x_2$. Thus the functions represented by each side of the equation (namely, $\eta(f)$ and $\frac{1}{2}\,\mu(f \times f)\delta)$ each have dimension 2 in f. Thus this necessary condition for the validity of the formula, is satisfied.

A final comment on this simple example: What is the status of this "formula" anyway? In our treatment, it is quite definitely a theorem, (proved in Chapter 3) which really contains two parts: one which relates certain length and area functions; and another which asserts that the product $f(a_1) \cdot f(a_2)$ is the same, no matter which side of b is taken as a_1. In elementary work, the formula $y = \frac{1}{2}\,x_1x_2$ is often taken as the definition of the area function $\eta(f)$ for triangular regions, and this definition is motivated by informal arguments concerning congruence and additivity (which actually come quite close to our formal definition of an area function) with the question of invariance, with respect to choice of base, quietly ignored. In scientific work the formula is sometimes taken as having empirical justification; but this presupposes an independent method of measuring area. The invariance of the product $f(a_1) \cdot f(a_2)$, for different choices of base and corresponding altitude, could certainly be tested empirically, as it is a simple (and not completely obvious) statement about certain length measurements on triangles. There would seem to be advantages in having students do this empirical checking at the time that the area of a triangle is first discussed, as this result is usually given long before there is any possibility of providing a formal invariance proof.

Dimensional Methods. Dimensional considerations (as in the above example are frequently used for checking the possible validity of formulas whose terms represent numbers derived from measurement, where the relevant measure functions belong to ratio scales. Typically, such a formula is asserted to be valid over a specified domain or domains, and for any choices of measure functions from the different ratio scales involved, provided that the secondary scales involved are derived from secondary quantities, and therefore have a specific dimensional relationship to each primary quantity. A measure function may be selected arbitrarily from each primary scale, but the secondary measure functions used must then be those which are determined by the definitions of the secondary quantities. Each such secondary quantity has a definite dimension in each primary quantity. For a formula to be valid under such conditions, it follows that, corresponding to the primary quantity F_i , a scale change (within the scale) from a function f_i to the function $f'_i = k_i f_i$, will multiply the numbers on each side of the equation by $k_i^{\alpha_i}$ and $k_i^{\alpha'_i}$, where α_i and α_i are the respective dimensions of the two sides of the formula, in F_i . It follows that α_i and α'_i must be equal, if the formula is to be valid under the conditions specified. Dimensional agreement of the two sides of such a formula is, of course, a necessary, but not a sufficient condition for the correctness of the formula. (In the example above concerning the triangle area formula, the dimension of the right hand side of the equation in "length" was not dependent on the factor "$\frac{1}{2}$" ; any other constant would have left the dimensionality unchanged, but the formula would then have been incorrect.)

As a further simple example of a formula whose "dimensionality" we can check, we give a formula (which used to be a well-known theorem of elementary trigonometry) relating the area of a triangle and the lengths of its sides.

Let b be a triangular region, with sides a_1 , a_2 , a_3 . Let f be a length function, and $\eta(f)$ the "naturally" related area function. Let $f(a_1) = x_1$, $f(a_2) = x_2$, $f(a_3) = x_3$, and let $\eta(f)(b) = y$. Let $s = \frac{1}{2}(x_1 + x_2 + x_3)$ (that is, s is half of the perimeter, as measured by f) . Then the numbers s , x_1 , x_2 , x_3 , y , are related by the following formula (which is proved in many trigonometry books):

$$y = [s(s - x_1)(s - x_2)(s - x_3)]^{1/2} .$$

It is a relatively simple matter to put each side of this equation in a form where it represents a "derived" function (derived from a length function f) on the domain of all triangular regions, and to verify directly that the right hand side has the same dimension in f (namely, 2) as the left hand side, $\eta(f)$. Thus the necessary dimensional condition is satisfied.

This example is, of course, rather trivial, but the method is not. In the physical sciences such simple dimensional tests provide a useful check on the possible validity of formulas.

Deriving Formulas From Dimensional Considerations. As we have seen above, dimensional considerations can sometimes be used to test the validity of a formula which relates numbers obtained as the result of specific, and interrelated, measurement operations: the dimensional argument cannot show that the formula is correct, but it might show that it is incorrect. Another use of dimensional arguments is in the derivation of formulas. Again, such a derivation cannot generally be complete, but sometimes dimensional arguments can be used to give some limitation on the possible form of a relationship between "measurements", a relationship which is, in fact, usually one between certain "primary" and "derived" measure functions; (i.e., between homogeneous functions on ratio scales). Virtually all of the significant examples of this method (which is usually known as "dimensional analysis" or the "method of similitude") belong to science and engineering, and any useful discussion of them requires a knowledge of these areas far beyond anything that we have presupposed for this book. In spite of this limitation, we can give you some idea of what dimensional analysis is about, and we can point out some of the difficulties by means of a very simple example.

Roughly speaking, in the study of a particular physical situation, or phenomenon, dimensional analysis concerns the derivation of possible relationships between certain "variables", representing numbers obtained as a result of the measurement of the quantities involved in the phenomenon. The measure functions involved all belong to ratio scales. Some are regarded as "primary", and some as "secondary". These terms have no absolute significance, but are regarded as being fixed for the particular phenomenon under consideration. Moreover all of the "secondary" measures are regarded as having specific relationships to the "primary" measures involved, so that, in particular, each "secondary measure" has a quite definite dimension in each "primary" measure. The type of relationship sought is one which holds for all instances of the phenomenon (i.e., over a certain non-trivial domain), and which takes the same form if a different set of "primary" measure functions are used, along with the corresponding "secondary" measure functions. Thus the relation sought is, in effect, a relationship between functions, and not just one between numbers; and the requirement that its form should be invariant under "change of units", implies that (assuming that the desired relationship is expressed as the equality of two dimensionally homogeneous functions of the primary quantities involved) each side of the equation must have the same dimension in each of the "primary" quantities.

Why we should be interested in physical relationships which have this particular type of invariance is, of course, not a question that can be answered mathematically. Appeal is frequently made to some sort of physical "principle", which you will find expressed in some scientific texts by such statements as

> "If an observable is to have qualitative usefulness it must have a significance which is independent of the choice of measurement units."
>
> "The laws of physics do not depend on the chosen system of units."
>
> "Every legitimate physical equation must satisfy the principle of dimensional homogeneity."

The status of these statements need not concern us here. As far as the mathematical side of dimensional analysis is concerned, certain assumptions have to be made. These include:

(i) The phenomenon under consideration is one to which the methods of dimensional analysis are applicable.

(ii) All of the relevant "quantities" (e.g., length, mass, time, etc., including any relevant dimensional constants) are known. Some of these are distinguished as "primary" and the remainder are "secondary". The "primary" scales are ratio scales, and all of the other "quantities" transform homogeneously with respect to change of units in the "primary" scales. Thus specific dimensional relationships (in terms of the "primary" quantities) are assumed for all of the "secondary" quantities and all of the dimensional constants involved.

At this point a mathematical discussion is possible. If the desired relationship is to be expressed as an equation, then each side of the equation will represent a function of the various measure functions and dimensional constants involved, and each side must have the same dimension (separately for each of the "primary" quantities) if the equation is to remain valid when arbitrary changes of function are made for the "primary" scales, with consequential changes in the "secondary" scales and dimensional constants. The mathematical analysis of this situation sets definite limits to the kind of relationships which can occur. In many cases, each side of the equation will itself represent a function which belongs to a ratio scale, and the earlier discussion of the possible form of such a secondary scale will apply. (In effect, it must be simply related to a product of powers of the scales (primary and secondary) which are involved.) The dimensions of each side of the equation must be the same in each primary quantity. These powers are then equated, separately for each primary quantity, and elementary linear algebra can be used to obtain the possible "forms" of the equation sought. You will find examples of this procedure in [17]. At best the relationship may be determined up to an unknown (absolute) constant (which may usually be determined from other information) but, more generally, the "dimensional analysis" will only determine the relationship to within one or more arbitrary functions of certain zero dimensional (or dimensionless) products of the quantities involved. The general solution is known as the Π-Theorem. Proofs of this theorem can be found in most books devoted to dimensional analysis, including [3], [17], and [18]. The mathematics involved is mainly linear algebra (specifically, the theory of simultaneous linear equations), and is not particularly difficult. But no amount of valid mathematical argument can give a result which is physically significant, unless the proper physical assumptions are made before the mathematical argument begins. And, in general, the decision as to what are the proper assumptions to make in any particular physical situation demands a considerable amount of genuine experience.

The books referred to above contain many examples of the use of dimensional analysis, and some discussion of the sort of difficulties which even an experienced scientist can encounter. The following example illustrates a few of these difficulties in a particularly simple form.

Example. A student of science, who had just learned some of the elementary facts about measurement and dimensions, is engaged in painting his home. He commences painting with a can of paint containing $\frac{1}{8}$ cubic feet of paint. (We use the "cubic foot" volume function to avoid unnecessary and irrelevant complications.) When he has used up the can of paint he finds (by measurement and calculation) that he has covered 500 square feet, and that he has 300 square feet still to paint. He wishes to know how much paint to buy in order to complete the job. From his knowledge of dimensional methods, he reasons as follows:

> The area covered depends on the volume of paint used. Hence, if the area covered by volume y cubic feet is x square feet, then $x = \varphi(y)$. The relationship of x and y will have the form $x = cy^{\alpha}$, for some real numbers $c > 0$, and α. From dimensional

considerations, if the length function (feet) is changed by a factor of k, the area function changes by a factor of k^2, and the volume function by a factor of k^3. Hence

$$k^2 x = c(k^3 y)^\alpha$$

from which $k^2 = (k^3)^\alpha$, so that $\alpha = \frac{2}{3}$. The constant c is now determined from the relation $500 = c\,(\frac{1}{8})^{2/3} = \frac{c}{4}$, whence $c = 2000$. Hence the remaining 300 square feet can be painted with y cubic feet, where $300 = 2000\,y^{2/3}$, so that $y = (\frac{3}{20})^{3/2}$.

Of course this is complete nonsense. "Obviously" the solution calls for a simple proportionality: If y cubic feet is required, then

$$y : \frac{1}{8} = 300 : 500$$

so that $y = \frac{3}{40}$.

How should we "explain" the mistake which our student has made? We might say that, by experience (or even "common sense") this type of problem is always solved by a simple proportionality, and we might even back this up by pointing out that the method used by our student would lead us to conclude that, in order to cover twice the area covered by a volume v of paint, we would require a volume $2^{2/3}v$, whereas, "clearly", twice as much paint is required to cover twice the area. Our student is likely to agree with this, but still feel unhappy concerning the reasons for the failure of his method. A sophisticated answer would be to point out that the constant c in his "formula" $x = cy^\alpha$ is a dimensional constant, related to an assumed constant thickness of the paint film; and that this dimensional constant is "clearly" of dimension -1 in length. If our student now reworks the problem with this assumption, then α will be 1, and his answer will be "correct".

There are two worthwhile comments which can be made concerning this simple example:

(i) Although we would not normally think of using dimensional analysis on such a problem, this should not prevent us from getting a correct answer by the use of dimensional methods. The knowledge, or intuition, that enables a scientist to see at once that a dimensional constant is involved, can only be gained from experience, or by writing down carefully all of the assumptions of the problem.

(ii) From a mathematical point of view, there is an incompleteness about the whole approach to the problem, whether using dimensional methods or using the method of proportion. This incompleteness results from the use of methods which depend on certain assumptions, without making these assumptions explicit. If our student asks us why the method of proportion "works" we should be able to analyze the problem to discover the relevant implicit assumptions. In this particular example these are

(a) that the paint is applied in a film of uniform thickness;

(b) that the shape of the surface is irrelevant, so that we can assume that the volume of the paint after application can be calculated (as area of base × height) as if it were a rectangular prism of very small height;

(c) that the volume of paint in the can is the same as the volume of paint on the surface after application. (Actually it is sufficient to assume that the two volumes are proportional, with a constant (not dimensional!) factor of proportionality; e.g., a "shrinkage" factor.)

From these assumptions we may solve the problem directly, by first calculating the assumed constant thickness from the fact that the volume of paint on 500 square feet of surface is $\frac{1}{8}$ cubic foot. If we "set up", without actually completing, this calculation, we will see the justification for the use of proportionality methods.

There is an almost endless list of simple proportionality problems which are related to questions of measurement. The mathematics of these problems is usually completely trivial, and the fact that so many students have trouble with them is possibly due to the fact that their intuitive grasp of the consequences of the implicit assumptions is not satisfactory. It is not always possible to improve this intuition by actual physical experience (in fact many proportionality problems are only nominally "real" in the sense that the language suggests a "real" situation) so it might be worthwhile occasionally to examine such a problem critically, and extract the underlying assumptions to the point where a complete mathematical treatment is possible. Such a treatment will usually disclose why a proportionality argument is valid. The difference in approach of the two methods is roughly the same as that between the solution (or partial solution) of a problem (in mechanics, say) by the use of dimensional methods, and the solution from the differential "equations of motion". Most mathematicians seem to prefer the latter approach, but the former can yield useful information in situations which are too complex to permit a complete mathematical formulation or solution. This is particularly true of many engineering applications, in which valuable information is obtained through the use of "scale models", a procedure which involves the use of dimensional methods. Examples of this are discussed in the books given earlier as references. In addition, some very simple ideas about such "scaling" are contained in [20] and [21]. The latter book, which was first published in 1638, has very great historical interest, in spite of its many deficiencies.

REFERENCES

1. Churchman, C. W., and Ratoosh, P. (Eds.), Measurement: Definitions and Theories, New York, John Wiley and Sons, 1959.

2. Bridgman, P. W., The Logic of Modern Physics, New York, The MacMillan Company, 1938.

3. Focken, C., Dimensional Methods, London, Arnold, 1953.

4. Margenau, H., Nature of Physical Reality, New York, McGraw-Hill, 1950.

5. Eddington, A. S., The Philosophy of Physical Science, New York, The MacMillan Company, 1939.

6. Henkin, L., et al, Retracing Elementary Mathematics, New York, The MacMillan Company, 1962.

7. Birkhoff, G., and MacLane, S., A Survey of Modern Algebra, New York, The MacMillan Company, 1965.

8. Von Neumann, J., and Morgenstern, D., Theory of Games and Economic Behavior, Princeton, Princeton University Press, 1944.

9. Stevens, S. S. (Ed.), Handbook of Experimental Psychology, New York, John Wiley and Sons, 1951.

10. Cohen, L. W., and Ehrlich, G., The Structure of The Real Number System, Princeton, Van Nostrand, 1963.

11. Halmos, P., Measure Theory, Princeton, Van Nostrand, 1950.

12. Niven, I., Mathematics of Choice: How to Count Without Counting, New York, Random House, 1965.

13. Polya, G., Mathematical Methods in Science, Stanford, School Mathematics Study Group, 1963.

14. Moise, E. E., Elementary Geometry From an Advanced Standpoint, Reading, Addison-Wesley, 1963.

15. Moise, E. E., and Downs, F., Geometry, Reading, Addison-Wesley, 1964.

16. Natanson, I., Theory of Functions of a Real Variable, Vol. I, New York, F. Ungar, 1955.

17. Bridgman, P. W., Dimensional Analysis, New Haven, Yale University Press, 1963.

18. Birkhoff, G., Hydrodynamics, Princeton University Press, 1960.

19. Kestelman, H., Modern Theories of Integration, New York, Dover, 1960.

REFERENCES (cont'd)

20. Physical Sciences Study Committee, Physics, Boston, D. C. Heath and Company, 1960.

21. Galileo, G., Two New Sciences, New York, Dover. (First published in 1638.)

INDEX

abelian group, 32
absolute constant, 404
affine group, 37
angle,
  definition of, 210
  generalized, 236
  idea of, 209
angle measurement, 195
  empirical, 199
angular arc, 218
angular coordinate functions, 197
angular coordinates, 229
angular distance, 197, 229
angular measure, 353
  domain extension, 232
angular measure function, 213, 222
angular measure
  in mathematics, 208
  of a rotation. 239
angular measurement
  in synthetic geometry, 208
angularity, 196
archimedean postulate, 132
archimedean property, 215
arc length, 219
  as angle measure, 228
area, 243
area, from an empirical standpoint, 244
area function, 250, 322
  for polygonal regions, 268
area functions,
  for developable surfaces, 304
  existence of, 261
  extension of domain, 295
  monotonicity, 283
  naming of, 256, 287
  for polygonal regions, 267, 282
  for polyhedral surfaces, 303
  for rectangular regions, 248
area of surfaces, 302
arithmetic progression, 107
automorphism (of groups), 34
automorphisms of (R+,+), 65

betweeness postulates, 128
betweenness,
  for rays, 210
bihomogeneous function, 30
  degree of, 30
bilinear function, 30
binary relation, 10
Bolyai's Theorem, 283
Boolean algebra,
  of sets, 182
Boolean ring,
  of sets, 182
Borel, 297
Borel sets, 190
broken segment, 174
broken segments, 166
Bridgman, 317, 321, 334

Cantor-Dedekind postulate, 143, 225, 226
cardinal measure, 83
cartesian plane, 9
cartesian product, 9
Cavalieri's Principle, 312
characteristic function, 302
closed interval, 185
closed set in R, 186
co-halfplanar angles, 214
co-halfplanar rays, 210
commutative diagram, 16
commutative group, 32
completeness.
  of real number system, 58
completeness postulates, 143
congruence,
  for angles, 211
  of segments, 120
congruence postulates, 129
constant function, 22
coordinate system, 118
coordinate systems, 155
coordinates,
  transformation of, 157
correspondence,
  1-1, 13
curve,
  definition of, 172
  length, 165
  simple, 173
curves,
  rectifiable, 170, 171, 176

Dedekind cut, 56
degree,
  of homogeneous function, 336
Dehn, 312
developable surfaces, 303
diagonal mapping, 373
diagram chasing, 327
diagrams,
  commutative, 16
dimension,
  of homogeneous function, 336
  of secondary quantity, 340
dimensional analysis, 414
dimensional constant, 402
dimensional constants, 405
dimensional methods, 411
directed angles, 232, 233, 198
directed rotations, 232
directed segments, 162

- direction on a line, 164
- distance function, 118
- dual space, 147
- duality,
  - of functions and units, 147; 386

- endomorphism, 18
  - (of groups), 34
- endomorphisms of $(R^+,+)$, 65
- endomorphisms of $(R^+,\cdot)$, 67
- equivalence classes, 11
- equivalence relations, 11
- Erlanger program, 42
- exponential functions, 69.
- existence of area functions, 261

- formulas, 407
- formulas and dimensions, 399
- function, 8
  - 1-1, 13
  - argument of, 8
  - as set of ordered pairs, 13
  - bihomogeneous, 30
  - bilinear, 30
  - decreasing, 19
  - domain of, 8
  - graph of, 13
  - increasing, 19
  - inverse of, 13
  - isotone, 19
  - linear, 26
  - monotone, 69
  - polynomial, 26
  - range of, 8
  - strongly monotone, 19
  - value of, 8
- functions, 12
  - algebra of real valued, 22
  - composition of, 15
  - exponential, 69
  - homogeneous, 28, 109
  - linear, 28
  - logarithmic, 69,
  - multihomogeneous, 30, 347
  - multilinear, 30
  - power, 69
  - power of, 76
  - similar, 25, 26

- geometric plane rotation, 237
- geometric ray rotations, 237
- graph, 13
  - of function, 13
- group, 32
  - abelian, 32
  - affine, 37
  - commutative, 32
  - isotonic, 36
- groups,
  - permutation, 34

- harmonic progression, 108
- homeomorphism, 172
- homogeneous functions, 28, 109
  - degree of, 29
  - properties of, 356
  - on ratio scales, 336
- homomorphism, 18
  - (of groups), 34

- image, 8
- image space, 8
- induction,
  - axiom of, 53
- integers,
  - construction of, 54
- isomorphic geometries, 123
- isomorphism, 18
  - (of groups), 34
  - of sets, 13
- isotone function, 19
- isotonic group, 36

- Jordan measure, 297
  - properties of, 298

- Lebesgue, 297
- Lebesgue inner measure, 188
- Lebesgue measure,
  - for plane regions, 296, 300
- Lebesgue measurable set, 189
- Lebesgue outer measure, 188
- length,
  - in cartesian geometry, 153
  - of a curve, 176
  - of curves, 165
  - empirical measurement of, 91
  - in mathematical systems, 116
  - measurement of, 51
  - in metric geometry, 118
  - in synthetic geometry, 127
- length function, 94, 119, 322
  - for broken segments, 168
  - for segments, 129
- length functions, 278
  - domain extension, 113, 165, 179
  - existence of, 132
  - range of, 143
  - similarity of, 141
  - units of, 99
- linear function, 26
- linear functions, 28
- linear functional, 146
- logarithmic functions, 69

magnitude,
    relative, 321
mapping, 8
Martinot-Lagarde, 334
measure dimension, 316
measure function
    for rotations, 238
measure functions, 8
    affine-invariant, 45
    classification of, 42, 48
    secondary, 324
measurement and dimension, 315
measurement,
    idea of, 6
    of angles, 195
    of secondary quantities, 338
    system of, 402
    of volume, 308
measures,
    similarity-invariant, 44
measure theory, 179
monotone endomorphisms of R^+, 334
monotone function, 19
monotone ray rotations, 241
monotone uniform functions, 334
multihomogeneous functions, 30, 347
multilinear functions, 30

natural numbers, 52
    initial segment, 54
    trichotomy, 54
non-measurable set, 191
normal rectangular region, 284
number line, 9
numerosity,
    function, 84
    measurement of, 51, 77, 245

open interval, 185
open set in R, 186
operation,
    associative, 14
    binary, 14
    commutative, 14
order,
    archimedean, 55
    dense, 55
order relation, 11
    strict total, 12
    weak total, 12
ordered field, 55
ordering,
    partial, 11
orientation, 233
    of a line, 164
    of directed angle, 235
    of plane, 234
    of segments, 162

partial ordering, 11
partition, 11
Peano axioms, 52
permutation groups, 34
physical quantity, 322
plane curve, 178
plane ray rotations, 241
plane rotations, 242
point transformations, 159
polar coordinates, 231
polygonal region, 248
    triangulation of, 249
polyhedral regions,
    volume functions for, 311
polyhedral surface, 303
polynomial function, 26
power function, 335
power functions, 67, 69
    properties of, 71
primary quantity, 322
primary scale,
    of measure functions, 322
product,
    of ratio scales
products of secondary quantities, 407

radian measure, 207, 218, 353
radian measure function,
    range of, 222
ratio operation, 100, 102
    additive, 103
    cancellation property, 102
ratio operations, 148
ratio scale, 44
    as semimodule over R^+, 387
    complete, 319
    extension of, 363
ratio scales, 148, 317
    and duality, 393
    generalized tensor products of, 397
    generation of, 355
    mathematical theory of, 354
    powers of, 359
    products of, 366
    relationships between, 320
    relationship of powers and
        products, 373
    tensor products and duality, 396
rational numbers,
    construction of, 55
ray rotation, 239
    monotone, 241
rays,
    co-halfplanar, 210
real numbers, 56, 58
real number systems, 52
rectangle, 250
rectangular region, 250

rectifiable curves, 170, 171, 176
relation,
  antisymmetric, 11
  binary, 10
  domain of, 10
  irreflexive, 12
  order, 11
  range of, 10
  reflexive, 10
  symmetric, 10
  transitive, 10
relations, 9
  equivalence, 11
rigid motion, 123, 158
rotation,
  center of, 238
  of the plane, 242
  of a ray, 238
rotations,
  for measure function, 238
  of the plane, 241
Russell, Bertrand, 5

scalar multiplication, 145, 148
  (of functions), 24
scale,
  interval, 46
  models, 157
  nominal, 46
  ordinal, 46
  ratio, 44
secondary measure function, 324
secondary quantity, 322
  definition of, 342
  dimension of, 347
  dimensionless, 402
  general definition of, 352
  simple, 342
secondary quantities
  measurement of, 338
secondary scale, 323
segment,
  broken, 166, 174
  directed, 162
sensed rotations, 198
semigroup, 39
  of segment classes, 131
semimodule over R^+, 146
similar functions, 25, 26, 96, 315, 355
similarity group, 38
similitude, 26
simple curve, 173
similarity transformation, 25, 26
space curve, 178
subgroup, 32
surd field, 126
surface area, 302
synthetic geometry,
  length in, 127
symmetric group, 34

tensor products,
  of R^+-semimodules, 394
  of ratio scales, 394
topological transformation, 172
triangle, 248
triangular region, 248
triangulated polygonal region, 273
triangulation,
  refinement of, 273
trichotomy,
  law of, 12
  for natural numbers, 54

uniform function,
  for ratio scales, 332
uniform functions,
  monotone, 334
unit, of a measure function, 89
unit-free statement, 107, 112

value space, 8
vector measures, 162, 165
volume,
  measurement of, 308
volume functions
  for polyhedral regions, 311
  for rectangular parallelepipeds, 310