DOCUMENT RESUME

ED 234 819 IR 050 453

AUTHOR Eckels, Diane Cole
TITLE Introductory Data Collection and Analysis.
INSTITUTION Medical Library Association, Chicago, Ill.
PUB DATE 78
NOTE 71p.; Medical Library Association Courses for Continuing Education: CE 41.
PUB TYPE Guides - Classroom Use - Materials (For Learner) (051) -- Reference Materials - Bibliographies (131)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Course Content; *Data Collection; Library Administration; *Library Research; Professional Continuing Education; *Questionnaires; Research Methodology; *Statistical Analysis; Statistics
IDENTIFIERS *Library Statistics

ABSTRACT

Designed for persons with no prior knowledge of statistics, this continuing education course syllabus presents basic information on methods of data collection and analysis in libraries. It is noted that emphasis is placed on concepts rather than mathematical formulas and on reasons for using particular techniques. Topics covered include problem definition; study design; the advantages and disadvantages of using direct observation, historical records, published surveys, interviews, and questionnaires for data collection; the design, administration, and evaluation of questionnaires; random sampling; the tabulation and graphical representation of descriptive statistics; statistical estimation; statistical decisions; variance tests (the F test and analysis of variance); the Chi-Square test; correlation; and regression. Examples are provided of the use of data collection and analysis techniques in libraries. Also provided are a series of 12 problems to be completed, a list of important equations, an 11-item bibliography, and a core list of journals that report the results of research in library science. (ESR)

MEDICAL LIBRARY ASSOCIATION
COURSES FOR CONTINUING EDUCATION

CE 41

Introductory Data Collection and Analysis

Diane Cole Eckels
Houston Academy of Medicine
Texas Medical Center Library
Houston, Texas

This syllabus is not intended to stand alone. It is only one part of an integrated instructional package involving a qualfied instructor, the instructional environment, supplementary materials and program evaluation. CEU's may be granted only by the Medical Library Association in accordance with it's Continuing Education Program.

Revised edition

1978

OT 41: Introductory Data Collection and Analysis

This course is intended as an introduction to methods of data collection and analysis in libraries. No prior knowledge of statistics is assumed.

The topics included should provide background both for persons interested in conducting simple quantitative studies themselves and for those who are interested in better understanding quantitative articles in the literature.

Emphasis will be given to the questionnaire and direct sampling methods of data collection. Data analysis methods discussed will include descriptive statistics, regression, correlation and tests of significance. The use of these techniques in a library setting will be discussed with examples familiar to the participants.

Concepts will be stressed rather than mathematical formulas although examples will be worked in class. Problem formulation, reasons for using particular techniques, advantages and disadvantages of the various methods, and the interpretation and presentation of results will be emphasized.

Course objectives

At the conclusion of the course, participants will be able to:

1. design and administer a simple questionnaire or direct sampling investigation using appropriate methodology;

2. critically evaluate questionnaires and direct sampling studies as reported in the library literature;

3. evaluate and improve existing data collection and analysis methods within their own institution.

INTRODUCTION

The course is intended to serve as an introduction to the collection and analysis of data as an aid to decision-making. Terminology and applications are the main focus, the intention being to provide you with a starting point for you to read and evaluate the literature of quantitative techniques as used in libraries.

If you want to learn more, choose really introductory books (some examples are given in the bibliography) and read from several different sources. It is advantageous to see the same concepts expressed in different ways.

A word of advice: if you are interested in conducting substantial studies, get someone with some experience to help you. There should be someone around your work place who routinely uses statistical techniques. This person will probably be willing to advise you in designing your study providing that you actually do the work.

As for actually solving mathematical problems, there are many different computer programs: such as Omnitab and SPSS

which can perform the actual calculations necessary in solving statistical problems. So don't get mired in the mathematics of the problems you solve. However, you must know what techniques to use, and how to interpret the results of the computer programs.

It is important that you read this Syllabus before you come to the session. Exercises in the Syllabus as well as additional example problems will be gone over in class. Class attendees will have the opportunity to discuss problems particular to their institutions as time permits. It will be helpful to bring a pocket calculator.

OUTLINE

I. Data Collection and Analysis

A. Problem Definition/Purpose of Study

B. Study Design

C. Data Collection

1. Data Sources

a. Direct Observation

b. Historical Records

c. Published Surveys

d. Interviews

e. Questionnaires

Non-Response Problem in Questionnaires

Other Guidelines for Questionnaires

2. Introduction to Random Sampling

D. Data Analysis

1. Descriptive Statistics

2. Statistical Estimation and Statistical Decision

3. Variance Tests

a. F Test

b. Analysis of Variance

4. Chi-Square Test

5. Correlation

6. Regression

E. Presentation of Results

II. Problems

III. Equations

IV. Recommended Readings

Data Collection and Analysis

A. Problem Definition/Purpose of Study

Collection and analysis of data (statistics) is merely a tool; as such, it can be used for good or evil. The real power of this tool depends upon its ability to shed light on a particular problem. Defining the problem or determining the goal of the study is the most crucial step in the entire process. It is all too easy to become overly concerned with numbers and forget the real problem.

There are certainly many ways to solve problems, for example, accepting the advice of a trusted person, reading the literature, or using quantitative methods. Numerical methods are not necessarily better than others. The fact that a technique is quantitative does not mean that it is "scientific". A good quantitative study is appropriately designed and addresses the real problem to be solved.

This course introduces some techniques which may help you glean from data information that will help solve a problem. One must

keep in mind that there is a lot of judgment involved in applying quantitative methods, because the right problem has to be addressed. A very real danger in defining a problem is "suboptimization". This means that you don't look far enough for the problem and therefore concentrate on only a portion of the problem or even on the wrong problem. If the problem that you have defined is not the real problem but only a symptom of it, all the sophisticated mathematical analyses in the world will not help you solve the problem that you wanted to attack. For example, you might define the problem as being how to encourage people to read more books, whereas the broader problem, and the one you should be studying is how can more people be educated. The latter problem would lead to investigation of methods of education, non-print media, books, computer retrieval systems and so forth instead of merely purchasing more books or initiating reading programs.

In summary, the first step in any quantitative study is to define the problem to be investigated, thus clarifying the objectives of the study.

Class participants should bring some examples of problems to be discussed.

B. Study Design

After the purpose of the study has been defined, the study is designed. The following questions should be answered before beginning a study:

Questions Related to Background and Problem Definition

1. Is the study necessary?
2. How much is known about the subject from the literature or other studies?
3. What information is to be determined from the study; i.e., what type of information is required to achieve the purpose of the study?

Questions Related to Data Collection and Analysis

4. Where and how can that information be obtained?
5. What is the target population?
6. Should there be a sample or a complete census?
7. If you sample, how important is it that the information be accurate, and how much time and money are available?
8. What type of data source is most appropriate?
9. What type of analysis techniques would be useful?

Once the problem is defined and steps 1 and 2 completed, the appropriate data to solve the problem must be isolated. Consider a fairly tightly constrained problem such as selection of a jobber for

journal purchase. What information is needed to solve the problem? First, determine what you think is most important to make a jobber effective. For example, you might choose:

1. Reliability
2. Cost
3. Time from order to receipt

If it is not possible to measure some aspect of effectiveness, a surrogate measure can be chosen. For example, reliability is difficult or impossible to measure, so the number of claims made might be substituted for lack of a better measure. Thus, in evaluating one jobber's performance against another's, the time, cost, and number of claims might be used for comparison.

Now steps 4 through 7 must be completed. Try to choose the most cost-beneficial method of getting the information. Additional accuracy in data often is obtained only at additional cost which may or may not be justified by the use to which you

put the data. For example, to address the problem; "how is the library meeting the needs of the users?" You might consider the following data collection methods:

1. Ask every user exiting the library a group of well-selected questions
2. Take a sample of users and ask them certain questions
3. Put up a suggestion box
4. Wait for comments by users.

The first approach requires quite a bit of time for carefully designing the questions, asking them, and analyzing the results as well as time from the users. The second method takes time for designing a sampling technique but less time than questioning all users. With a carefully designed sample and good questions, quite accurate results can be obtained at a lower cost than by the first method. The third method is cheap but will definitely result in biased results. For example, the suggestions will be based on the expectations of the users. Users will probably not make a suggestion unless they feel the library can make

a change. They might suggest longer operating hours but not better reference service because they have grown to accept mediocre service. The last approach will probably result in even more of a biased result than the third.

In summary, good study design means choosing the method which will get the most useful information in the most accurate and cheapest way for a particular problem, defining the population from which data is to be collected, and selecting a data analysis technique. The study design stage involves planning the entire study, in other words, through the data analysis stage.

Consider the problem which has been defined as:

How much card catalog space will be required in five years?

Go through steps 1 through 7 with this problem. Could the problem be defined in another way?

The class will have time to discuss some problems of special interest to them.

C. Data Collection

Data may be obtained either by a complete count (census) or a sampling of the group being investigated. In practical cases a census is often unwieldy because of the size of the group to be investigated; a correctly taken sample can yield accurate enough results at a much lower cost and in a shorter time than a census. In fact, sometimes a sample can be more accurate than a census because the larger amount of data collected by a census introduce a greater chance of inaccuracies in the data.

A list of data collection methods and some advantages and disadvantages of each follow. No method is particularly better than another, but the most appropriate source of data should be chosen for each problem.

1. Data Sources

Using either a census or a sample, there are several sources of data:

a. Direct Observation

Advantages

-- All assumptions are known

-- Data are usually consistent

-- Data are usually not subject to misinterpretation

Disadvantages

-- Time-consuming

-- May be difficult to obtain a random sample

b. Historical Records

Advantages

-- Simple to collect data if it already exists

-- Data are usually consistent if records are kept consistently

Disadvantage

-- The data desired may not be available in its complete form

c. Published Surveys

Advantages

-- The data are already compiled, saving time and expense

-- The responsibility for accuracy may be shifted

Disadvantages

-- The data obtained by the primary investigation cannot be verified

-- The statistical technique used may not be ascertainable and therefore the accuracy of the results may not be verifiable

-- Subjective compiling and interpretation may have influenced the result shown

-- The purpose of the study may have prejudiced the choice of material and technique adopted

-- A representative sample may not have been taken

d. Interviews

Advantages

-- A higher degree of accuracy is attained through the acquisition of data direct from the source

-- Data are often obtained that cannot be obtained through a questionnaire alone

-- There is opportunity personally to check information acquired

-- The "no response" proportion is usually minimized

Disadvantages

-- Only relatively small samples can be gathered because of the cost

-- The subjective factor is involved in recording by interviewer

e. Questionnaires

Advantages

-- A large area may be easily and

quickly covered

-- The method of collecting data is relatively inexpensive

Disadvantages

-- Frequently questions cannot be answered without supplementary explanations

-- In many cases the results are unreliable due to a large "non response"

-- Questions may be misinterpreted or answered incorrectly

For the following problems which would be the most appropriate data source and why?

1. What amount of space will be required for housing materials in five years?
2. How satisfied are users with the library's service?
3. What type of materials are used by practicing physicians?
4. What should the operating budget of a nursing library be?

Because questionnaires have been so frequently used in library studies, guidelines for conducting a questionnaire and a discussion of the problem of non-response follow:

Non-Response Problem in Questionnaires

The effectiveness of the questionnaire method depends not only on the size of the original sample but also on the percentage of returns. The return rate is very important in questionnaires because of potential bias in the results due to the non-respondents being different from the respondents in some important way(s). An attempt should be made to discover who does not return the questionnaire and every effort made to encourage response in order to minimize problems with non-respondent's bias.

Some ways of encouraging response are:

--incentives

--follow-ups

--include stamped self-addressed envelope

--guarantee anonymity

--make questionnaire attractive and easy to complete

--communicate the importance of the questionnaire

--establish reasonable deadlines

--offer the results of the survey to interested respondents

Other Guidelines for Questionnaires

-- Be sure that all the questions are necessary; there is a temptation to ask too much.

-- Be sure that each question is understandable and precise; ambiguity discourages response and accuracy.

-- Define terms carefully to avoid misinterpretation.

-- Pretest the questionnaire with a similar group; a pretest can indicate where there are problems with ambiguous questions.

-- Do not ask people to answer questions they cannot. Questions should be easy to answer; if they require looking up from various sources, respondents may make up answers.

-- Send the questionnaire to the person most likely to know the answer.

-- Sequence the questions in a logical order; the order of questions is important for a train of thought.

-- Precoded questions (those that provide for only certain, set answers) are easier to complete than open ended ones.

-- Pay attention to length and physical layout; long, unattractive questionnaires have lower response rates than short, attractive ones.

-- The questionnaire should be easy to analyze.

-- Questions should be objective, not "loaded".

-- How well does the following questionnaire follow the guidelines?

To: Reference Librarian, University Library

QUESTIONS

I. Books

Classification	No. of Titles	No. of Volumes
General Works		
Philosophy		
Social Science		
Pure Science		
Art, (Fine)		
History		
Travel		
Biography (medical)		
Fiction		
Medicine		

How are paperback acquired?

Used?

Name five periodicals subscribed to:

List other periodicals available.

Are some periodicals maintained in back files?

List five:

How are they kept?

Are visual aids available? _____

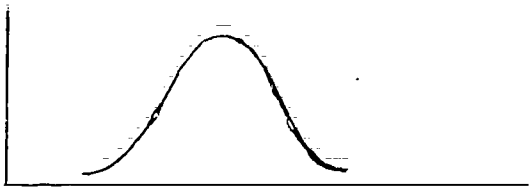
Check Below	Number Available
Motion Pictures	_____
Film Strips	_____
Slides	_____
Recordings	_____
Tapes	_____
Discs	_____
Transparencies	_____

SECTION D.1. SHOULD BE READ BEFORE THIS SECTION.

2. Introduction to Random Sampling

Certain statistical laws enable us to infer things about a population from a sample taken from that population. If the sample is chosen according to certain rules, the accuracy of the information collected from the sample can be determined. This theory makes it possible to do inferential statistics--tests of significance, hypothesis testing, and estimation of population parameters from a sample. The basic concept of sampling theory is closely related to the normal distribution.

A normal distribution is symmetrical (the curve looks identical on either side of the mean), bell-shaped, and asymptotic to the horizontal axis; i.e., never touches the horizontal axis.



A lot of phenomena in the world follow this normal distribution. Sampling theory is

based on the fact that the means of all possible samples of a population follow a normal distribution and the mean of all possible sample means equals the mean of the population. This permits us to infer information about the entire population from a small sample because of certain known properties of the normal distribution.

In order to properly utilize inferential statistics, the sample from which you want to infer something about the population must be representative of the population. One way to accomplish this objective is by simple random sampling (each member of the population has an equal chance of being selected). Two methods that are commonly used to select random samples are, (1) systematic sampling, (taking every "ith" item on a list) and (2) using a random number table, frequently found in the back of statistics books, to determine which items in a list to include in the sample. Both methods will be explained in class.

There are other correct methods of taking random samples which may be more cost-effective than simple random sampling, but they are outside the scope of this course.

Determining an appropriate sample size

The sample size is related to both confidence level (risk) and confidence interval (precision). The objective of determining the sample size is to achieve a satisfactory degree of certainty (95-99% for example) that the sample estimate of a particular characteristic is in error by no more than a specified amount.

The equation for determining the sample size is:

$$n = \frac{z^2 v}{e^2}$$

where n = sample size to be determined

v = variance of population, the number representing the population dispersion (how much the members of the population vary). To ascertain the size of the sample to achieve a specified precision, we must have a reasonable approximation of v in advance. A knowledge of the approximate

size of v may be available from past experience. If not, some pretesting and preliminary study may be necessary in order to have an approximate value for it.

If the population has little variance, the size of the sample can be small. For example, if every member of the population were exactly alike, you could just sample one member and you would then know about all the others.

e = number representing confidence interval (precision) required. The size of the sample required is tremendously affected by the size of e (error tolerated or precision required).

z = number representing confidence level (risk) required. The value selected for z determines the probability that the sample result will have an error no greater than e.

z and e are chosen by the investigator based on what the information will be used for and the amount of time and money available.

Problem 3 on page 55 is related to this section; other examples of sample size determination will be worked in class.

D. Data Analysis

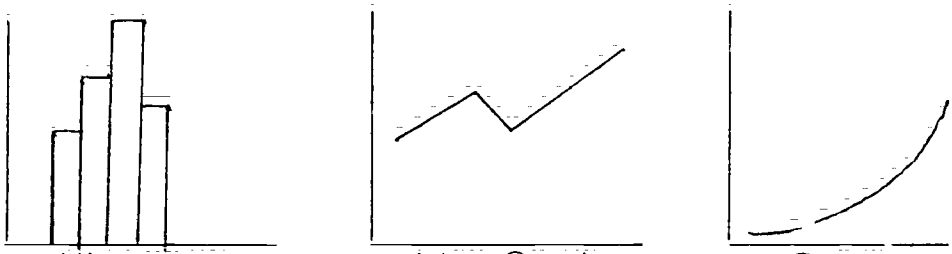
Quantitative studies may be descriptive such as describing the average salary in a sample or analytical, for example, predicting next year's salary increase for a group based on samples of data. In an analytical study, the planning of the analysis you conduct is essential. Do not gather a multitude of data and then try to decide what to do with it.

The type of analysis you perform with any set of data depends upon the quality of the data and should not be more sophisticated than the data nor inconsistent with the aim of the study: statistical manipulation and decimal points can give an aura of conviction which is unwarranted.

1. Descriptive Statistics

Descriptive statistics includes the tabulation, classification, graphical representation, and calculation of certain summary values which describe group characteristics.

Distributions are graphs which describe the occurrence of values. Histograms and line graphs are used to describe values that can be dissected into discrete parts; curves are used to describe continuous values (those which can take on any value).



Measures of Central Tendency

Measures of central tendency provide a single number which, in a sense, summarizes a group of data. The most common ones are:

- Mean--The total of all the values divided by the number of values--also called the average. Equation = $\bar{x} = \frac{\sum_{i}^{n} x_i}{n}$

- Median--If all values are arranged in ascending or descending order, the value which is the middle one is the median. This number is not as sensitive to extremely large or small numbers as is

the mean. Therefore the median may sometimes be more meaningful than the mean.

- Mode--The value which occurs most frequently. A distribution may have more than one mode.

Measures of Dispersion

Measures of dispersion are single values which describe the scatter or differences among values. Those which are most useful are:

- Range--Simplest of all measures of dispersion, the range can be the difference between the largest and smallest values or the largest and smallest numbers themselves.
- Variance--The arithmetic mean of the squared deviations from the mean.

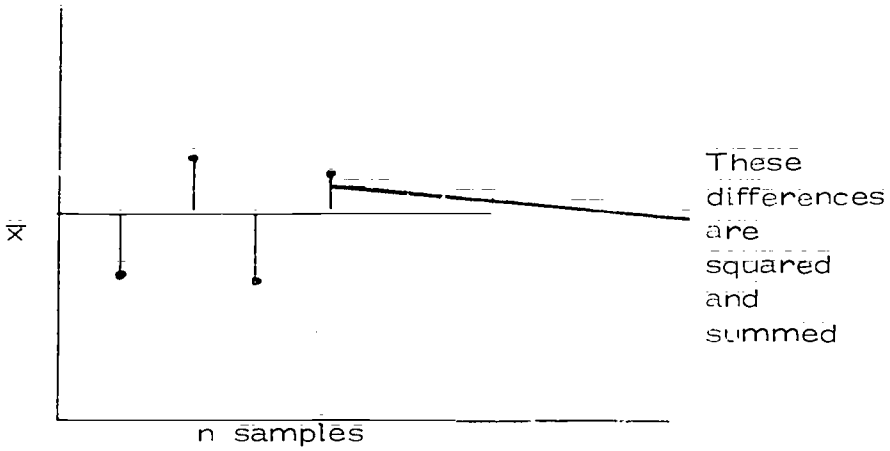
Equation =

$$V = \sum_{i}^{n} \frac{(x_i - \bar{x})^2}{n}$$

where $\bar{x}$ = the mean

$x_{i \text{ to } n}$ = each value

n = number of values



If the vertical distances shown above (the deviations from the mean) were just added and not squared before adding, the result would always be 0.

- Standard Deviation--The square root of the variance. This measure is in the same units as the original values.

$$s = \sqrt{v}$$

Two sets of numbers may have the same average but be quite different. The set with the larger variance and standard deviation would be more scattered.

Problems 1 and 2 on page 55 are related to this section.

2. Statistical Estimation and Statistical Decisions

The techniques of inferential statistics provide you with the ability to make sound conclusions about a population on the basis of a sample. Statistical inference may be used to make an estimate of a population value based on a limited sample (statistical estimates) or it may be used to test some hypothesis by means of information from an experiment or sample (statistical decisions).

Statistical Estimations

A population parameter can be estimated on the basis of a sample. For example, if the average price of a journal is unknown, an estimate can be obtained from a sample. The estimate of the average price of all journals is obtained from the sample mean. The type of estimate that consists of a single value is called a point estimate. An interval estimate states in terms of probability the

likelihood that the population mean will occur in a specified range.

An interval estimate of a sample mean is represented by:

$$\bar{x} \pm zs_{\bar{x}}$$

where $\bar{x}$ = mean calculated from sample

z = represents the level of confidence required and

$s_{\bar{x}}$ = standard deviation of the sample means; estimated by $\frac{s}{\sqrt{n}}$

Problem 4 on page 56 is related to this section.

Statistical Decisions

Procedures which enable us to decide whether to accept or reject hypotheses or to determine whether observed samples differ significantly from expected results are called tests of hypotheses, tests of significance, or rules of decision. A statistical hypothesis is a tentative statement about one or more parameters (values which describe the entire population) of a population or group of populations.

In the literature you will see statements such as:

- The difference between the sample means is significant at .01=P
- Sample means differed at the 0.05 level of significance.
- Null hypothesis was rejected at the 0.05 level.

This section will teach you what such statements mean. Hypotheses are usually made about populations and the data gathered from a sample or samples of the population are used to test the hypotheses. Hypotheses are stated in terms of differences between populations, such as the difference between the population of male librarians and the population of female librarians. The numbers obtained from the data from the populations are then subjected to tests of significance described later.

If the numbers from the two samples are different it may be only because of

sampling error. You must, therefore, determine if there is enough difference between them to say that the two samples come from different populations.

The null hypothesis states that there is no difference between the two groups. Rejecting the null means there is a real difference between the groups. However, an hypothesis can be rejected when it should have been accepted, and conceivably it can be accepted when it should have been rejected.

These two types of errors associated with rejection of the null are called "Type I" and "Type II". Type I error is rejecting a true hypothesis and Type II is accepting a false hypothesis. They are tabulated below.

	Accept	Reject
True	Correct Decision	I
False	II	Correct Decision

We usually specify the level of significance of a Type I error to be 0.05 or 0.01. Five per-cent (.05) means that there are five chances out of one hundred that we might

make a mistake and say there is a real difference between the two groups when there is not (reject the null hypothesis when it should not be).

Not being able to reject the null hypothesis does not mean that it can necessarily be accepted. That could result in a Type II error. Do not accept the null unless the probability of a Type II error is known. Just do not reject the null.

A t-test can be used to test an hypothesis about the mean of a population. For example, the binding company says they can return items in an average of 12 days. We take a sample of ten and the mean is 16. Could the population mean really be 12?

$H_o: \mu = 12$

$n = 10$

$\bar{x} = 16$ degrees of freedom = $n-1 = 9$

$s = 4$

$$t = \frac{\bar{x}-\mu}{s}\sqrt{n-1}$$

$$t = \frac{16-12}{4}\sqrt{9}$$

$$t = 3$$

t at .05 significance -2.26 to 2.26

t at .01 significance -3.25 to 3.25

3 > 2.26 Reject null at .05, but cannot reject at .01

The parameters of two samples can be compared using a t-test.

Equation:

$$t = \frac{\bar{x}_1-\bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}\sqrt{1/n_1+1/n_2}}$$

where $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{n_1 s_1^2+n_2 s_2^2}{n_1+n_2-2}}$

degrees of freedom = n_1+n_2-2

Problem 10 on page 59 is related to this section.

3. Variance Tests

As stated previously, standard deviations and variances reflect the amount of scatter in collected data.

a. F Test

The F test is used to determine whether one sample has more scatter than another or if the difference is really due to sampling error. Large samples do not require as large a F ratio for significance as small samples because there is less sampling error in larger samples.

$$F \text{ ratio} = \frac{\text{Larger variance}}{\text{Smaller variance}}$$

Example

Medical students' circulation of materials:

$$n = 41$$

$$s^2 = v = 33.5$$

Residents' circulation of materials:

$$n = 50$$

$$s^2 = v = 74.1$$

$$F \text{ ratio} = \frac{74.1}{33.5} = 2.21$$

The F ratio is looked up in an F table in most statistics books using also the number in each of the samples. If the F ratio is large enough the null hypothesis can be rejected and there is a significant difference in the scatter. In this case the F ratio for sample sizes 41 and 50 is around 1.6 which is less than 2.21 so the null can be rejected. Problem 6 is related to this section.

b. Analysis of Variance (ANOVA)

This method actually compares means of samples rather than variance, but the variance measures are used to compare the means. To compare the means of two samples to see if they differ significantly a t test could be used. To compare several samples ANOVA would be used. The purpose is to show that the variance estimate between the samples is significantly larger than the variance estimate within the samples.

ANOVA indicates if there is an overall difference among the means; it does not indicate which groups produced the difference. ANOVA can be used to decide whether samples were drawn from the same population.

Example of the use of ANOVA

Three methods of supervisory counseling on absenteeism are being analyzed. The null hypothesis says there is no difference in the effectiveness of these methods. You would expect some variability among absence rates of the three samples just due to sampling error. Is the variability among absence rates of the samples large enough to reject the null and say there is a difference in the effectiveness of the methods?

Procedure:

1. Calculate within-groups variance for each of the three samples, then combine the three estimates to obtain one estimate.

2. Calculate the between-groups variance estimate using the mean of each of the samples and compute a variance estimate using these three means and the size of the samples.

An F ratio is computed between the two variance estimates using the between groups variance as the numerator and the within-groups variance as the denominator. If the means of the groups really explain the variation (not from the same population) the numerator has much more variance than the denominator.

$$F = \frac{\text{Between-groups variance}}{\text{Within-groups variance}}$$

A computer program can be used to calculate the F ratio for ANOVA. The results printed will usually be the F ratio and the level of significance at which the groups differed. You must then decide if the level is sufficiently high to reject the null hypothesis.

Another example of the use of ANOVA might be when designing an automated

circulation system to determine the type of record format one might wish to see if there is a difference in the average length of titles between journals, monographs and audiovisuals.

Often it is desirable to test hypotheses concerning two variables. In the previous example of the record length, the question was whether or not there were significant differences between types of materials. To carry this a step further assume there were two languages. The investigator might be interested in testing to see if there was a significant difference between the two languages in the average length of the titles. This is called two-way analysis of variance.

4. Chi Square

The chi square test determines whether there is a statistically significant difference between the frequency of events among groups. (It does not reveal how much difference there is and if the test fails it does not really prove that the groups are the same. This test is used when comparing frequencies of events among several groups and can be used with qualitative data, e.g., hair color.

Characteristics of the Chi Square Method

-- Non-parametric method, i.e., it makes no assumptions about the distribution of the item sampled. The assumptions of non-parametric methods are usually easier to satisfy than parametric.

-- Deals with frequencies rather than scores.

-- Can be used with several groups and attributes.

The Chi Square Test can be used:

- To test the significance of a statistic; do the groups differ significantly? As in all sampling situations, there is a possibility that the difference in the numbers may be due to sampling error. A significant difference means that the observed frequencies differ enough from the expected frequencies (those we would expect to occur by chance) to justify rejection of the null hypothesis of no difference in the groups.

- To test the goodness of fit. How will the data gathered fit a theoretical distribution; for example, the normal? If the data fits a standard distribution about which certain properties are known, those properties can be used for prediction.

Definition of Chi Square

Chi Square is a statistic χ^2 which supplies a measure of the discrepancy existing between observed and expected frequencies.

If $\chi^2 = 0$ the observed and theoretical (or expected) frequencies agree exactly; the larger the value of χ^2, the greater the discrepancy. At some point the null hypothesis of no significant difference can be rejected.

There are Chi Square tables on the back of most most statistics books. You need to know the number of categories studied and the level of risk to use the table. If Chi Square is larger in your study than in the table the null hypothesis of no difference can be rejected.

Equation for Chi Square:

$$\chi^2 = \sum_i^n \frac{(f_{oi} - f_{ei})^2}{f_{ei}}$$

Example 1

In 200 people using the library (randomly selected) 115 doctors and 85 nurses are found. The difference may be due to sampling error. Test the hypothesis that there really are an equal number of doctors and nurses using the library.

Observed frequencies: $fo_1 = 115$, $fo_2 = 85$

Expected frequencies: $fe_1 = 100$, $fe_2 = 100$

$$\chi^2 = \frac{(fo_1 - fe_1)^2}{fe_1} + \frac{(fo_2 - fe_2)^2}{fe_2}$$

$$= \frac{(115-100)^2}{100} + \frac{(85-100)^2}{100} = 4.5$$

Number of categories = 2, d.f = 2-1 = 1

$\chi^2_{(.95)}$ for 1 degree of freedom = 3.84, since $4.5 > 3.84$, reject null hypothesis at 0.05 level of significance.

$\chi^2_{(.99)}$ for 1 degree of freedom = 6.63, since $4.5 < 6.63$, cannot reject null hypothesis at 0.01 level.

There probably is a significant difference and there probably are not an equal number of doctors and nurses using the library; however, the difference could be due to sampling error.

Example 2

Is there a difference in the frequency of reading foreign language journals between clinicians and researchers? A random sample of 200 scientists was taken using the alphabetical campus directory and systematic sampling. The results were:

80 researchers, 40 of which read foreign language journals

120 clinicians, 50 of which read foreign language journals

	Foreign	No Foreign	Total
Clinician	50	70	120
Researcher	40	40	80
Total	90	110	200

f_o table (observed frequencies)

If there were no difference between the groups as to proportion who read foreign language journals:

	Foreign	No Foreign	Total
Clinician	54	66	120
Researcher	36	44	80
Total	90	110	200

f_e table (expected frequency table)

Differences between the two tables:

	Foreign	No Foreign
Clinician	+4	-4
Researcher	-4	+4

$f_o - f_e$ table

$$\frac{(+4)^2}{54} + \frac{(-4)^2}{36} + \frac{(-4)^2}{66} + \frac{(+4)^2}{44} = \chi^2$$

$$1.34 = \chi^2$$

What does the answer mean?

Since $\chi^2(.95) = 3.84$ for 1 d.f. and $3.84 > 1.34$ it cannot be said that there is a significant difference in the proportion of clinicians and researchers who read foreign language journals. The difference was likely due to sampling error.

Problems 5 and 7 on page 56 are related to this section.

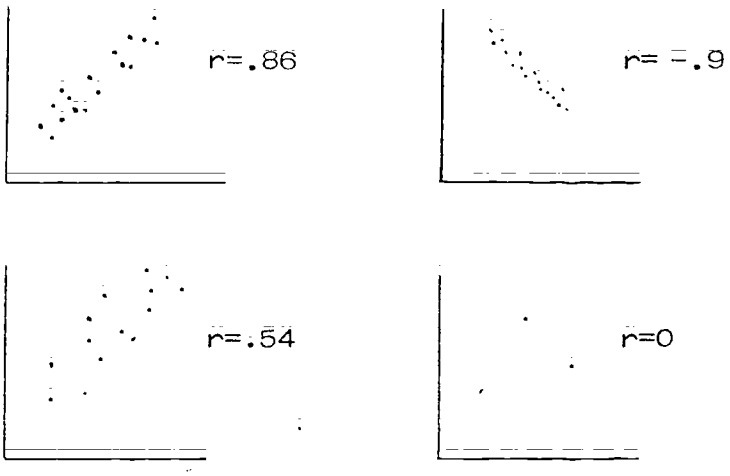
Can you think of some other library problems for which Chi Square would be an appropriate data analysis technique?

5. Correlation

Correlation indicates how related two variables are. One variable can be predicted from another if they are highly related. Correlation does not indicate a causal relationship. A third undiscovered variable may cause the effect or the relationship may be purely accidental (spurious correlation).

The statistic which is the measure of the correlation (closeness of the relationship between two variables) is the coefficient of correlation (r). The coefficient of correlation is useful in judging the relative strengths of association and indicating significant relationships between two variables. A coefficient of 1 is perfect, 0 is none.

Scatter Diagrams



Equation for coefficient of correlation:

$$r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}$$

Example:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

X	Y	$x=X-\bar{X}$	$y=Y-\bar{Y}$	x^2	xy	y^2
1	1	-6	-4	36	24	16
3	2	-4	-3	16	12	9
4	4	-3	-1	9	3	1
6	4	-1	-1	1	1	1
8	5	1	0	1	0	0
9	7	2	2	4	4	4
11	8	4	3	16	12	9
14	9	7	4	49	28	16
$\Sigma X=56$ $\bar{X}=7$	$\Sigma Y=40$ $\bar{Y}=5$			$\Sigma x^2=132$	$\Sigma xy=84$	$\Sigma y^2=56$

$$r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} = \frac{84}{\sqrt{(132)(56)}} = .977$$

Because data are taken from a sample, r is effected by sampling error. Therefore, after the correlation is done, the null hypothesis must be tested to see that r is significantly larger than zero and did not occur just as a result of sampling error. This is done with a t test.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

If you are doing the calculations by hand, you must determine from a table of values of the t statistic (usually found in the back of statistics books), whether the t calculated is large enough to reflect significance.

Computer programs are commonly available to conduct correlation analyses. The input required is the values for X and Y: the programs usually print the value of the correlation coefficient as well as its level of significance.

A significant result should be treated with caution since correlations can be due to coincidence or due to a third variable. Remember, no cause and effect can be associated with correlation.

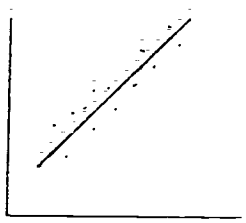
Examples of items which might be related and thus which might be subject to correlation analysis:

1. Books checked out and grade point
2. Operating expeditures and circulation
3. Book budget and circulation
4. Use of materials in house and circulation

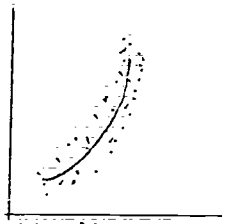
What are some other examples?

Regression

If a relationship is found to exist between variables, you might want to express this relationship in mathematical form with an equation connecting the variables.



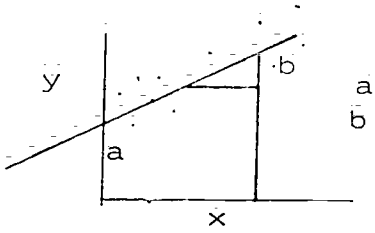
Linear



Nonlinear

The simplest type of approximating curve is a straight line, whose equation can be written:

$$y = a + bx$$



a = y intercept
b = slope, amount of change in y for one unit change in x

Using the least squares method, solving for a and b:

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

The method of least squares finds the coefficients a and b which make the sum of the squares of the distances from the line the least possible. This technique can also be used with more than two variables and is called multiple regression. Again, computer programs are available to calculate solutions.

After the regression line has been found, you need to know how well the line fits the data. If all the points of the scatter lie on the line, it is perfect linear correlation. The quantity r is called the coefficient of correlation.

$$r = \sqrt{\frac{\text{explained variation}}{\text{total variation}}}$$

r varies from -1 to +1 and is a measure of the linear correlation between two variables.

The standard error of the estimate indicates the accuracy of the estimate. It is a measure of the scatter about the regression line.

Baumol has used regression to predict total operating expenses of a library from number of personnel and other variables. Can you think of other examples?

Regression is a powerful tool for prediction, but it does not mean that a value outside the previously examined sphere will fit the line.

F. Presentation of Results

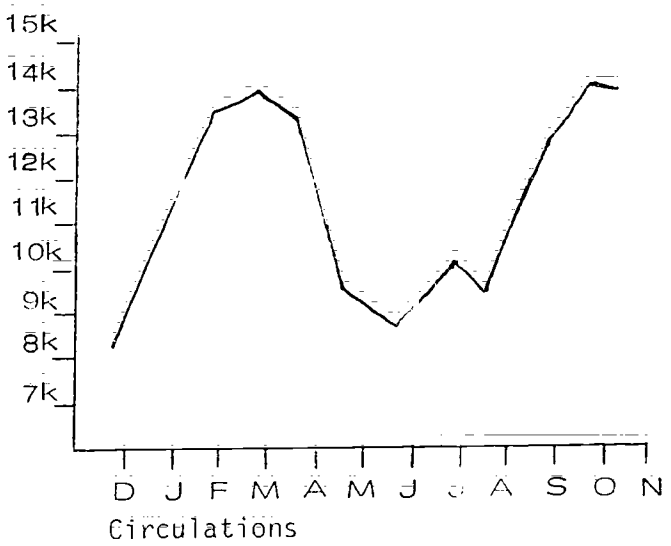
The final stage of a study is the fair interpretation and clear presentation of the results. The interpretation should include some consideration of the level of risk and the precision of the results. It is also essential that the reader be made aware of any distortions in the data due to sampling bias, as well as the limitations of the analysis procedures used. For example, it should be made clear that correlation does not indicate cause and effect.

The well-analyzed and presented study acknowledges its weaknesses, indicates what results are doubted, and suggests what remains to be done. In writing the report, emphasis should not be placed only on results which confirm the investigator's opinions: all significant findings, supporting or not, should be reported.

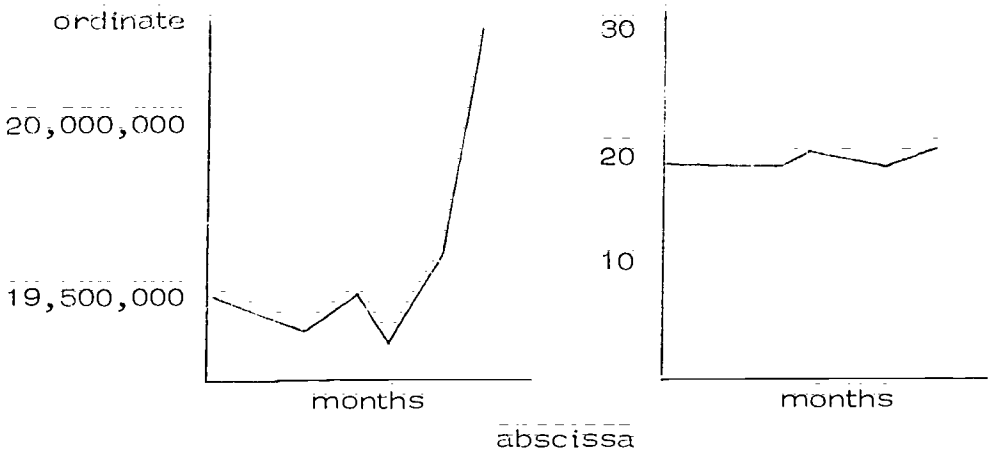
The presentation of results should highlight the important issues, not obfuscate them with data. Vast amounts of statistical data in raw form only confuse the issue. The data should be presented in an organized way to help communicate the facts to the reader. The organization can take two basic forms: statistical tables and figures. A statistical table is a presentation of numbers in a logical arrangement, with some brief explanation to show what they are. A figure is a chart, graph, map, or some other illustration designed to present statistical data in picture form.

Figures - The making and understanding of figures requires neither artistic nor mathematical skill--only common sense and the ability to use a ruler and compass are needed.

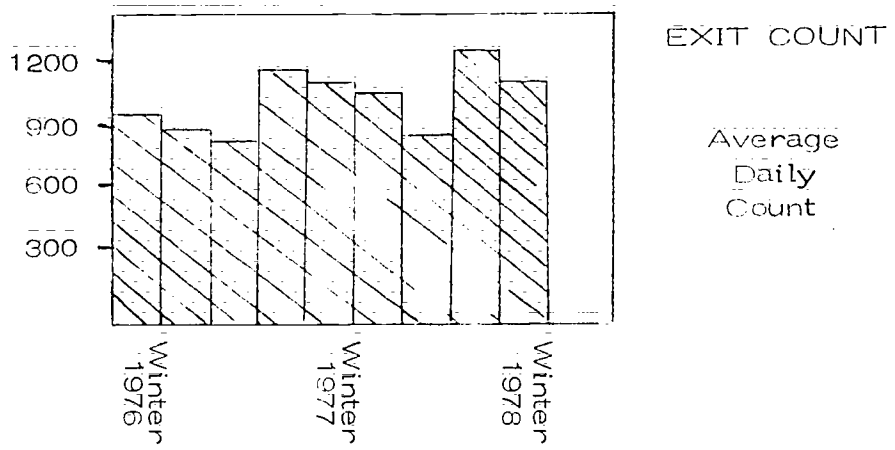
Line graphs are particularly good for indicating trends over periods.



However, truncation or changing the proportion between the ordinate and the abscissa can distort the true picture if this is not made clear to the reader.



Column or Bar charts are useful when a number of different groups are compared.



The reader should be presented with comparable figures. Often a percentage or ratio figure may be more informative in a column chart than only the raw figures.

Problems

1. Calculate the mean, median and mode for the two sets of data below:

Group 1	Group 2
2 5 9	2 3
2 6 6	9 1
3 7	10

2. Calculate the standard deviation and variance for the numbers in problem 1. Which set of data is the more scattered?

3. You wish to find the average time to answer a reference question. A preliminary sample indicates that the variance in questions is 9 minutes. You wish to have no more than one minute on either side of the mean with a level of risk of only five per cent (95% Confidence Level), $z = 1.96$. Using the equation for sample size, calculate what size a random sample should be taken.

4. The standard deviation of a group of data on the width of books is 2 inches. The mean of the sample was 5 inches. At a 95% confidence level, what is the confidence interval? There were 25 books in the sample.

5. A personnel director is interested in trying to determine if the season of the year has any effect on the number of employees who resign. His records give the following information:

Season	Number of resignations
Winter	10
Spring	22
Summer	19
Fall	9
	60

Test at a significance level of 0.05 to determine if there is a significant deviation between the observed distribution and a uniform distribution (equal for all seasons). H_o = null hypothesis, no difference, that is, the proportion of resignations is independent of the season of the year. A significance level of 0.05 requires

a χ^2 value of 7.815 (column in the chi square table headed 0.05 with three degrees of freedom) reject H_0 if $\chi^2 \geq 7.815$.

6: Make a significance test to determine if it can be assumed that the variance in width of books and journals is the same as the variance in width of journals.

Sample 1: Books	Sample 2: Journals
$n_1 = 16$	$n_2 = 21$
$s_1 = 1.5"$	$s_2 = 2.5"$

F ratio for $d.f._1 = 15$ and $d.f._2 = 20$ at 0.05 = 2.33.

7: The library speculated that if they offered a free MEDLINE search they would increase the percentage of questionnaires returned. To test this theory they sent questionnaires to a random sample of 30 persons on the list with the offer of a MEDLINE search. They sent another sample of 30 with no MEDLINE offer. The results are shown below:

MEDLINE	Returned	Not returned	Totals
Offered	22	8	30
Not offered	14	16	30
	36	24	60

The problem is to test at 0.05 level to see if there is a significant difference in the proportion of returns when the MEDLINE is offered. An 0.05 level required a Chi Square value of 3.841 (Column 0.05 with 1 degree of freedom).

8. You took some observations on the number of persons exiting the library and books used in the library each hour for eight hours. X = people exiting the library, Y = books used in the library.

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Using the equation for the correlation coefficient, see if these variables are highly correlated, is the t significantly greater than 0?

9. Find the least squares lines for the data in problem 8: $\bar{a}$ = ; $\bar{b}$ = . What is the equation for the line?

X	Y	X^2	XY	Y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\Sigma X = 56$	$\Sigma Y = 40$	$\Sigma X^2 = 524$	$\Sigma XY = 364$	$\Sigma Y^2 =$

10. Using the data in problem 6 and the fact that the sample mean of books = 6 and the sample mean of journals = 8 use a t-test to determine whether there is a significant difference between the means.

11. The amounts of money spent on foreign monographs and domestic monographs, foreign serials and domestic serials are given below. Using a compass construct a pie chart of the following data.

Foreign monographs	25,000	10%
Domestic monographs	50,000	20%
Foreign serials	75,000	30%
Domestic serials	100,000	40%
TOTAL	250,000	100%

12. Can you match the following problems to an appropriate data analysis technique?

Problems	Techniques
a. The average number of books checked out by each student last year was 3. Is it the same this year?	1. Chi Square
b. What's the average width of a journal volume?	2. Analysis of variance
c. Is the price of materials in chemistry and physics about the same?	3. t-test for the mean of a population
d. Is there a relationship between the number of in-house uses and circulations?	4. Regression
e. I want to estimate in-house uses from circulations.	5. Correlation
f. Do nurses use the library more frequently than physicians?	6. t-test for the means of two samples
	7. Estimation of a parameter

q. Is there a difference in the length of tenure of employees in Technical Services, Public Services and Administration Divisions?

Equations

Sample Size

$$n = \frac{z^2 v}{e^2}$$

Variance

$$v = \sum_{i}^{n} \frac{(x_i - \bar{x})^2}{n}$$

Standard Deviation

$$s = \sqrt{v}$$

Mean

$$\bar{x} = \sum_{i}^{n} \frac{x_i}{n}$$

Standard Error of the Mean

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Correlation Coefficient

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Regression Coefficients

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n\sum X^2 - (\sum X)^2}$$

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

RECOMMENDED READINGS

Baumol, W.J. and Marcus, M. Economics of Academic Libraries. (Washington, D.C.: American Council on Education, 1973).

Berdie, D.R., and Anderson, J.F. Questionnaires: Design and Use. (Metuchen, N.J.: Scarecrow, 1974).

Bookstein, A. "How to Sample Badly." Library Quarterly, 44, pp. 124-32, 1974.

Dougherty, R.M. and Heinritz, F.J. Scientific Management of Library Operations. (New York: Scarecrow, 1966).

Elzey, F. A First Reader in Statistics. (Monterey, Calif.: Brooks/Cole, 1974).

Hamburg, M.; Clelland, R.C.; Bommer, M.R.W.; Ramist, L.E., and Whitfield, R.M. Library Planning and Decision-Making Systems. (Cambridge, Mass.: MIT Press, 1974).

Huff, D. How to Lie with Statistics. (New York: Norton, 1954).

Lancaster, F.W. The Measurement and Evaluation of Library Services. (Washington, D.C.: Information Resources Press, 1977).

Line, M.B. Library Surveys. (Hampden, Conn.: Anchor Books, 1967).

Simpson, I.S. Basic Statistics for Libraries. (London: Clive Bengley, 1975).

Slonim, M.J. Sampling in a Nutshell. (New York: Simon & Schuster, 1960).

* Highly Recommended

Core List of Journals that Report Research in Library Science

1. ASLIB Proceedings
2. Bulletin of the Medical Library Association
3. College and Research Libraries
4. Information Storage and Retrieval
5. International Library Review
6. Journal of the American Society for Information Science
7. Journal of Documentation
8. Journal of Librarianship
9. Journal of Library Automation
10. Law Library Journal
11. Library Quarterly
12. Library Resources and Technical Services
13. Libri