DOCUMENT RESUME

ED 218 128 SE 038 238

AUTHOR Knapp, Thomas R.; And Others
TITLE Regression toward the Mean. Statistics. [and] Basic Descriptive Statistics. Descriptive Statistics. [and] Approximations in Probability Calculations. Applications of Statistics. Modules and Monographs in Undergraduate Mathematics and Its Applications Project. UMAP Units 406, 426, 443.
INSTITUTION Education Development Center, Inc., Newton, Mass.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE 80
GRANT SED-76-19615-A02
NOTE 89p.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Answer Keys; *College Mathematics; Higher Education; Instructional Materials; Learning Modules; *Mathematical Applications; *Mathematical Concepts; *Problem Solving; *Statistics

ABSTRACT

This document consists of three modules concerned with aspects of statistics. The first provides knowledge of the effect of imperfect correlation and random error on differences between means, and the reasons for the necessity of random allocation of objects to experimental and control conditions in scientific experimentation. The second unit shows how to: 1) Use frequency distributions and histograms to summarize data; 2) Calculate means, medians, and modes as measures of central location; 3) Decide which measures of central location may be most appropriate in a given instance; and 4) Calculate and interpret percentiles. The third module is designed to enable the student to: 1) discuss how approximation is pervasive in statistics; 2) compare "structural" and "mathematical" approximations to probability models; 3) describe and recognize a hypergeometric probability distribution and an experiment in which it holds; 4) recognize when hypergeometric probabilities can be approximated adequately by binomial, normal, or Poisson probabilities; 5) recognize when binomial probabilities can be approximated adequately by normal or Poisson probabilities; 6) recognize when the normal approximation to binomial probabilities requires the continuity correction to be adequate; and 7) calculate with a calculator or computer hypergeometric or binomial probabilities exactly or approximately. Exercises and tests, with answers, are provided in all three units. (MP)

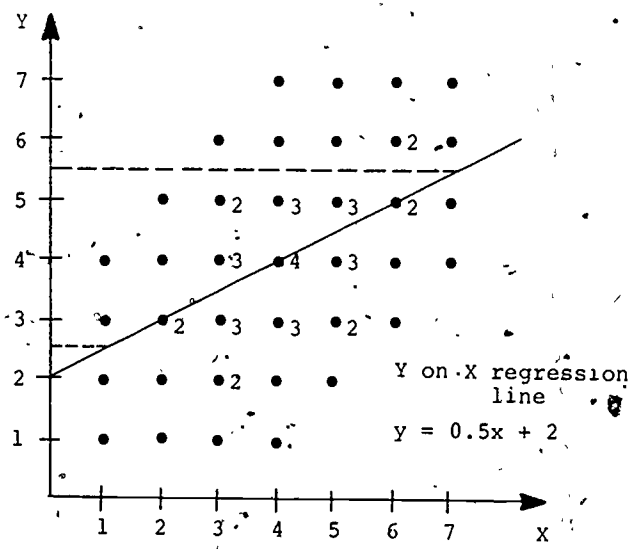
UNIT 406

umap

MODULES AND MONOGRAPHS IN UNDERGRADUATE MATHEMATICS AND ITS APPLICATIONS PROJECT

REGRESSION TOWARD THE MEAN

by Thomas R. Knapp



STATISTICS

edc/umap 55chapel st newton mass 02160

REGRESSION TOWARD THE MEAN

by

Thomas R. Knapp
Graduate School of Education and Human Development
University of Rochester
Rochester, NY 14627

TABLE OF CONTENTS

Intermodular Description Sheet: UMAP Unit 406

Title: REGRESSION TOWARD THE MEAN

Author: Thomas R. Knapp
Graduate School of Education and Human Development
University of Rochester
Rochester, NY 14627

Review Stage/Date: III 10/15/80

Classfication: STATISTICS

Prerequisite Skills:
1. Familiarity with basic descriptive statistics (mean, standard deviation, correlation coefficient).
2. Previous exposure to simple linear regression analysis.

Output Skills:
1. Knowledge of the effect of imperfect correlation and random error on differences between means.
2. Understand the necessity for random allocation of objects to experimental and control conditions in scientific experimentation.

The author would like to thank Joel R. Levin for the idea of using playing cards to demonstrate regression toward the mean.

The Project would like to thank Mattie E. Moss of Bennett College, Greensboro, North Carolina; Kenneth R. Driessel of AMOCO Research, Tulsa, Oklahoma, and Carol Stokes of Danville Area Community College, Danville, Illinois for their reviews, and all others who assisted in the production of this unit.

This module was prepared under the auspices of the UMAP Statistics Panel. Members of the Statistics Panel are: Thomas R. Knapp, Panel Chair, of the University of Rochester; Roger Carlson of the University of Missouri at Kansas City; J. Richard Elliot of Wilfred Laurier University; Earl Faulkner of Brigham Young University; Peter Holmes of the University of Sheffield; Peter Purdue of the University of Kentucky at Lexington; Judith Tanur of SUNY at Stony Brook; Maurice Tatsuoka of the University of Illinois at Champaign-Urbana; and Richard Walker of Mansfield State College.

This material was prepared with the partial support of National Science Foundation Grant No. SED76-19615 A02. Recommendations expressed are those of the author and do not necessarily reflect the views of the NSF or the copyright holder.

ABSTRACT

Regression toward the mean is a phenomenon that is a natural by-product of less than perfect correlation between two variables, but regression effects have often been mistaken for treatment effects in poorly-designed experiments. The purpose of this module is to explain, theoretically and empirically, this bothersome concept.

1. INTRODUCTION

Did you ever notice that the sons of very tall men are usually also tall but not quite as tall as their fathers? And that the sons of very short fathers tend to be not as short as their fathers? The famous anthropologist Francis Galton did, and he once believed that this would ultimately lead to the elimination of the very tall and the very short. Will it?

Probably not. As we shall see, this kind of "regression" is a statistical artifact of the imperfect correlation between any two variables (e.g., height of father and height of son). Unfortunately the lack of understanding of the principle continues to be a problem in scientific research.

2. WHAT IS REGRESSION TOWARD THE MEAN?

2.1 Definition

Regression toward the mean is the phenomenon whereby a high (low) set of observations on one variable is associated with a mean on another variable that is also high (low) but that is closer to the overall mean for that other variable. It is of no real scientific importance whatsoever; it is a necessary consequence of less than perfect correlation between two variables.

2.2 A Numerical and Graphical Illustration

Consider the scatterplot in Figure 1 for two variables X and Y that are on the same scale (the Pearson product-moment correlation coefficient for those data is 0.5), and pay special attention to the left-most array of four points (for X=1). The overall mean for variable X is 4, so those four observations are low relative to that mean. Note, however, that the mean for variable Y for those same observations is 2.5, which is closer to the overall mean for variable Y (also 4) than the 1 is to the mean of 4 for variable X. The reason for this is simply the shape of the scatterplot. Since there is not a perfect linear relationship between the two variables, the most extreme observations on X are not necessarily associated with the most extreme observations on Y. When the very lowest X measures are considered, the corresponding measures for Y have nowhere to go but up, so to speak.

This phenomenon also operates from the top down, as well as from the bottom up. Again referring to Figure 1, the right-most array of four points (for X=7) produces a mean for variable Y of 5.5, which is closer to the overall Y-mean of 4 than 7 is to the overall X-mean of 4.

For simplicity of illustration, the Y measures of Figure 1 were put on the same scale as the X measures. That is not necessary, however. The general shape of the scatterplot remains the same if either X or Y is transformed linearly.

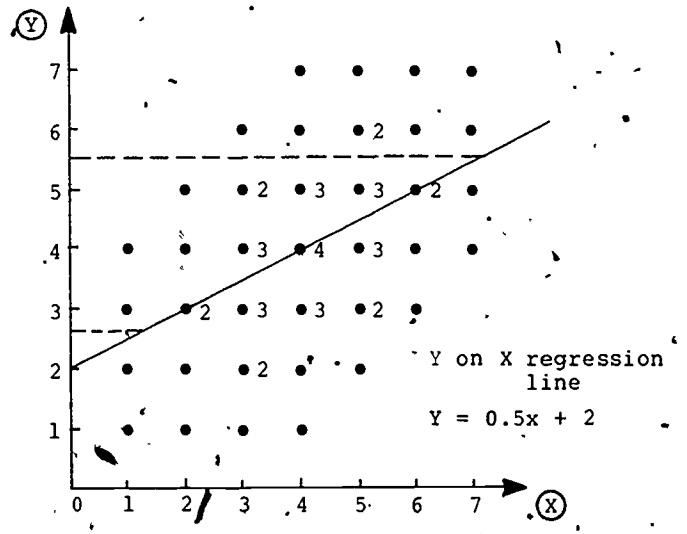


Figure 1. An illustration of regression toward the mean. (Adapted from Campbell, D.T. and Stanley, J.C., Experimental and quasi-experimental designs for research, Rand McNally, 1966, page 10. The numbers next to some of the points are the frequencies of those observations. The points without numbers represent single observations. The total number of observations is 58.)

2.3 Mathematical Explanation

A single illustration is not a sufficient explanation of a phenomenon. The following algebraic argument treats the general case.

Consider the equation of the regression line for Y on X, namely

(1) $Y = bX + a,$

where

$$b = r_{xy} \frac{S_y}{S_x} \quad (2)$$

and

$$a = M_y - bM_x. \quad (3)$$

(In these equations, M_x and M_y are the overall means, S_x and S_y are the overall standard deviations, and r_{xy} is the correlation between the two variables.) Substituting the values given by Eqs. (2) and (3) for b and a into Eq. (1), we have

$$Y = r_{xy} \frac{S_y}{S_x} X + M_y - r_{xy} \frac{S_y}{S_x} M_x. \quad (4)$$

Rearranging Eq. (4) algebraically leads to

$$Y = r_{xy} \frac{S_y}{S_x} (X - M_x) + M_y,$$

or

$$Y - M_y = r_{xy} \frac{S_y}{S_x} (X - M_x),$$

or

$$\frac{Y-M_y}{S_y} = r_{xy} \frac{X-M_x}{S_x}. \quad (5)$$

This is the so-called "standardized" form of the regression equation.

Now consider a set of observations for which X is k standard deviations from M_x. Then

$$\frac{Y-M_y}{S_y} = r_{xy} \frac{(M_x+kS_x)-M_x}{S_x} = kr_{xy}. \quad (6)$$

Since $|r_{xy}| \leq 1$, the value of Y on the regression line that "goes with" this extreme value of X (the Y-mean for the array) must be less than or equal to k standard deviations from M_y (equality holds only if $r_{xy} = \pm 1$). That's regression toward the mean, no matter what the values of k, r_{xy}, M_x, M_y, S_x, and S_y are.

3. SOME OTHER EXAMPLES

3.1 Reading Improvement

An educator gives a reading achievement test to a group of third-grade pupils, picks out the pupils who obtained the lowest scores on the test, gives them a two-month remedial reading program, tests them again, and observes that their scores are significantly higher. Is this evidence that the program has been successful? Not necessarily. It could be regression toward the mean; scores on the two tests probably do not correlate perfectly with one another.

3.2 Smoking and Lung Cancer

A physician examines several cancer patients, obtains a medical history of their cigarette smoking behavior, and discovers that those who smoked the most had only slightly more than an average amount of lung cancer. Does this mean that if you're going to smoke cigarettes you might as well smoke a lot? Perhaps; but there may be regression toward the mean here, too. Although there is a positive correlation between number of cigarettes smoked and amount of lung cancer, the correlation is far from perfect.

4. BUT WHAT IS IT THAT REGRESSES TOWARD WHICH MEAN?

This question can be best answered in the context of two technical, but simple, statistical concepts, namely expectation and conditionality. The expected value of a variable, say Y, is the mean value of that variable, usually written as E(Y). The conditional expected value of Y is the mean value given some constraint, say X, and is usually written as E(Y|X).

Regression toward the mean is concerned with the comparison between the quantities X - E(X) and E(Y|X) - E(Y). Referring to Figure 1 again, the (standardized) distance between any X and the mean of X is always greater than or equal to the distance between the mean of Y for that X and the overall Y mean. So it is E(Y|X) that regresses toward E(Y), relative to the discrepancy between X and E(X). If the correlation between X and Y is 0, i.e., if the scatterplot forms a "buckshot" pattern, the regression is maximal and E(Y|X) = E(Y). If the correlation is +1 or -1 there is no regression toward the mean, since the (standardized) distance between E(Y|X) and E(Y) is the same as the (standardized) distance between X and E(X).

5. AN EMPIRICAL DEMONSTRATION OF THE PHENOMENON

Take two decks of ordinary playing cards. Select the sevens, eights, and nines from one deck and call this reduced deck of 12 cards Deck A. Select the aces (ones) through nines from the other full deck and call this reduced deck of 36 cards Deck B. Pencil in the number -2 on each of the aces in Deck B; the number -1 on each of the twos and threes; the number 0 on each of the fours, fives, and sixes; the number +1 on each of the sevens and eights; and the number +2 on each of the nines (all in Deck B).

For each card in Deck A draw a card at random (with replacement) from Deck B. ("With replacement" means that you put the card back in the deck before you shuffle and

draw another one.) Add the 12 pairs of numbers (the actual denomination for the card in Deck A and the number -2, -1, 0, +1, or +2 drawn from Deck B). For example, paired with the seven of spades in Deck A you might have a -1 from Deck B. Adding these together you have 7 +(-1) = 6.

Now pick out the six largest sums (using any convenient randomizing procedure to resolve ties) and find their mean. (See Table 1 for an example of this step and all subsequent steps in the demonstration.) Set aside the six cards from Deck A that did not contribute to the largest sums. They will no longer be needed.

For the same six cards from Deck A that did contribute to the six largest sums, repeat the pairing, summing, and averaging process using six cards drawn at random from Deck B. Compare the two means. The second one should be lower. Do you know why? (Try to think of a reason before you read on.)

TABLE 1

One Set of Empirical Results

(regression toward the mean)

"First testing"

Deck A cards	Deck B cards	Sums
7	7 (+1)	8
7	9 (+2)	9
7	6 (0)	7
7	8 (+1)	8
8	8 (+1)	9
8	8 (+1)	9
8	7 (+1)	9
8	A (-2)	6
9	3 (-1)	8
9	5 (0)	9
9	6 (0)	9
9	6 (0)	9

mean of checked sums = 9.0

"Second testing"

Deck A cards	Deck B cards	Sums
7	A (-2)	5
8	2 (-1)	7
8	3 (-1)	7
9	6 (0)	9
9	8 (+1)	10
9	7 (+1)	10

48/6 = mean of 8.0

The sevens, eights, and nines originally chosen from the first full deck of cards are analogous to scores on a test that the 12 brightest of 36 students deserve to get. (The other 24 deserve to get one through six. Forget about the tens, jacks, queens, and kings.) The 12 sums are scores that they actually do get, scores that contain a random error component. (They all deserve high scores, but by chance some will "have a bad day" and obtain scores that are less than the ones they deserve, while others will "have a good day" and obtain scores that are greater than the ones they deserve.)

At the second "testing" the scores obtained by the "people" who had the six highest scores the first time would not be expected to correlate perfectly (because of the chance error components) with the first atypically high scores. Ergo, regression (downward) to the mean.

The moral to all of this is: if a group of people score very high on a test one time and get lower scores the next time, don't be surprised and don't get too concerned. The same implication holds at the low end of the scale: if a group of people score very low on a test one time and get higher scores the next time, don't get too elated. In both cases it could be wholly or partially regression toward the mean.

6. EXERCISES

1. Demonstrate for yourself that the implication just mentioned does hold at the low end of the scale by carrying out the demonstration described in Section 5 again. This time use the aces, twos, and threes from the first full deck of cards as Deck A, and pick out the six lowest sums.
2. Referring back to example 3.1, think of a reading improvement program being given to the "people" who obtain the six lowest scores at time 1, with the scores at time 2 as a measure of their performance at the end of the program. Do you see now why the "improvement" is a statistical necessity?

7. WHAT CAN BE DONE ABOUT IT

7.1 In Experimental Research

Whenever we're seriously interested in the effectiveness of a reading improvement program, a weight reduction plan, a headache remedy, etc., we should use two groups of people, randomly assigned to either receive (the experimental group) or not receive (the control group) the particular treatment in which we are interested. If all of the people happen to be recruited from extremely high or extremely low portions of some score distribution and are

given a pre-test before the experiment and a post-test after the experiment, the regression toward the mean effect will still take place, but it will be balanced across the two groups. For the reading program example, if people who get very low scores on the reading pre-test are randomly assigned to experimental (they get the program) and control (they don't) groups, both groups will do better on the post-test due to regression toward the mean, but if the program is really effective the members of the experimental group will score that much higher.

7.2 In Non-Experimental Research

The only thing that can be done in non-experimental research is to do the best we can in distinguishing between a legitimate finding and a regression effect. For the smoking and lung cancer example, the heights of sons vs. heights of fathers example, and similar studies, the extreme measures on one variable are usually associated with less extreme measures on the other variable for purely statistical reasons. (Selective mating has something to do with increasing the correlation between fathers' heights and sons' heights, but the regression effect provides a sufficient explanation for the reduction to "mediocrity" that Galton observed.)

Some people think that matching can take care of problems associated with regression toward the mean but, alas, it can't. In a well-known study by Helen Christiansen of the effect of high school graduation on economic adjustment during the early days of the depression, an original sample of 2127 people was reduced to 23 matched (on six background variables) pairs of graduates and non-graduates, with the graduates exhibiting better adjustment than the non-graduates. But the regression effect could very well account for the difference since the non-graduates who had been matched with the graduates on such things as mental ability and neighborhood status (both of which are positively correlated with economic adjustment) were well above average relative to their fellow non-graduates and would be expected to regress further (to their own population mean) than the graduates at the follow-up testing ten years later, thereby making the graduates appear to be better adjusted economically.

Note that it is not feasible to study the effect of high school graduation on economic adjustment experimentally, since it is socially unacceptable to assign some people to receive a high school education and to withhold it from others. However, there are better ways than the matched-pairs technique to control for confounding background variables, techniques that are also less subject to regression effects and do not result in the shrinkage of the research sample.

One final point: the regression effect works "backwards" as well as "forwards" statistically, even though it makes absolutely no sense scientifically. Very tall sons have fathers who are closer to average height than they are, which should convince you, if this module and your previous exposure to statistics have not already done so, that correlation per se does not necessarily imply causation.

8. REFERENCES

Campbell, D.T. and Stanley, J.C. Experimental and quasi-experimental designs for research. Rand McNally, 1966. (pp. 10-12 and 70-71).

Chapin, F.S. Experimental designs in sociological research revised edition. Greenwood, 1974. (pp. 99-124).

Freedman, D., Risani, R., and Purves, R. Statistics. Norton, 1978. (pp. 158-164).

9. END-OF-MODULE QUIZ

1. If a group of people who exhibited great test anxiety before counseling had greater test anxiety after counseling, is regression toward the mean a likely explanation? Why or why not?
2. If the regression equation for Y on X is $Y = 0.75X + 1.5$, $M_x = M_y = 6$, and $S_x = S_y = 2$, what is the mean on variable Y for ten observations for which X = 5? Does that make sense? Why or why not?
3. (Bonus question) In some experiments the people in the experimental group and the people in the control group are the same people, i.e., they receive both treatments. Is regression toward the mean a problem in such experiments? Why or why not?

10. ANSWERS

10.1 Answers to Exercises

1. It worked fine for me. The six lowest sums that I got were four 1's and two 2's, with a mean of 1.33. The corresponding sums the next time were 0, 1, 1, 3, 4, and 5, with a mean of 2.33, which is a point higher (and closer to the overall mean) than the first one.

2. It is artifactual because the six lowest "people" had bad luck the first time, and since luck plays no favorites they couldn't all have bad luck the second time; therefore, as a group they scored higher and would have done so with or without the program.

10.2 Answers to the Quiz

1. No, regression toward the mean is not a likely explanation, since they scored high the first time and higher, not lower, the second time. The regression effect is only relevant for high to lower and low to higher mean differences, i.e., an originally high group scores lower the second time or an originally low group scores higher the second time.

   The evidence suggests that the program was not only not effective, but harmful. However, since there was no control group (which would be treated in the same way as the experimental group except that they don't get the counseling) we cannot be sure that the counseling itself was ineffective. The disappointing results may be due to the counselor, the office in which the counseling took place, some other event that transpired during the counseling period, etc.

2. Substituting X = 5 in the regression equation, we obtain Y = 5.25. The 5.25 is closer to the mean of Y than the 5 is to the mean of X, so it indeed does make sense. X = 5 is not an extreme observation (it is only one-half of a standard deviation below the mean of X), but the regression effect actually works on all of the observations, not just the extreme ones, as Eq. (5) attests.

   The correlation coefficient for these data, by the way, is the same as the regression slope, b, i.e., 0.75, since

$$b = r_{xy} \frac{S_y}{S_x}$$

   and

$$S_y = S_x.$$

3. Yes, since pre-test and post-test scores still won't correlate perfectly. Things get a little more complicated, however, since you could have three or four, rather than two, testings to contend with: pre-testing before Treatment A, post-testing after Treatment A, pre-testing before Treatment B (which may be the same testing as the post-testing after Treatment A), and post-testing after Treatment B. The post-A scores should be closer to the mean than the pre-A scores, due to the regression effect, but since the experience of Treatment B is often not contemporaneous with the experience of Treatment A (the people usually can't be undergoing both treatments at once), the regression from pre-B to post-B may not be comparable.

STUDENT FORM 1

Request for Help

Return to:
EDC/UMAP
55 Chapel St.
Newton, MA 02160

Student: If you have trouble with a specific part of this unit, please fill out this form and take it to your instructor for assistance. The information you give will help the author to revise the unit.

Your Name_____ Unit No._____

Page_____
- ○ Upper
- ○ Middle
- ○ Lower

OR

Section_____
Paragraph_____

OR

Model Exam
Problem No._____
Text
Problem No._____

Description of Difficulty: (Please be specific)

Instructor: Please indicate your resolution of the difficulty in this box.

○ Corrected errors in materials. List corrections here:

○ Gave student better explanation, example, or procedure than in unit. Give brief outline of your addition here:

○ Assisted student in acquiring general learning and problem-solving skills (not using examples from this unit.)

Instructor's Signature_____

Please use reverse if necessary.

STUDENT FORM 2

Unit Questionnaire

Return to:
EDC/UMAP
55 Chapel St.
Newton, MA 02160

Name_____ Unit No._____ Date_____

Institution_____ Course No._____

Check the choice for each question that comes closest to your personal opinion.

1. How useful was the amount of detail in the unit?

____Not enough detail to understand the unit
____Unit would have been clearer with more detail
____Appropriate amount of detail
____Unit was occasionally too detailed, but this was not distracting
____Too much detail; I was often distracted

2. How helpful were the problem answers?

____Sample solutions were too brief; I could not do the intermediate steps
____Sufficient information was given to solve the problems
____Sample solutions were too detailed; I didn't need them

3. Except for fulfilling the prerequisites, how much did you use other sources (for example, instructor, friends, or other books) in order to understand the unit?

____A Lot ____Somewhat ____A Little ____Not at all

4. How long was this unit in comparison to the amount of time you generally spend on a lesson (lecture and homework assignment) in a typical math or science course?

Much	Somewhat	About	Somewhat	Much
____Longer	____Longer	____the Same	____Shorter	____Shorter

5. Were any of the following parts of the unit confusing or distracting? (Check as many as apply.)

____Prerequisites
____Statement of skills and concepts (objectives)
____Paragraph headings
____Examples
____Special Assistance Supplement (if present)
____Other, please explain_____

6. Were any of the following parts of the unit particularly helpful? (Check as many as apply.)

____Prerequisites
____Statement of skills and concepts (objectives)
____Examples
____Problems
____Paragraph headings
____Table of Contents
____Special Assistance Supplement (if present)
____Other, please explain_____

Please describe anything in the unit that you did not particularly like.

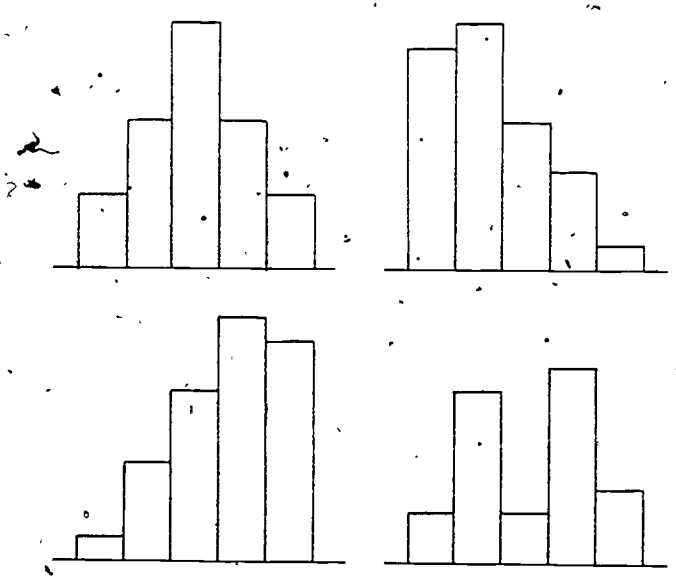
Please describe anything that you found particularly helpful. (Please use the back of this sheet if you need more space.)

umap

UNIT 428

BASIC DESCRIPTIVE STATISTICS

by Richard Walker



DESCRIPTIVE STATISTICS

BASIC DESCRIPTIVE STATISTICS

by

Richard Walker
Department of Mathematics
Mansfield State College
Mansfield, PA 16933

TABLE OF CONTENTS

Intermodular Description Sheet: UMAP Unit 428

Title: BASIC DESCRIPTIVE STATISTICS

Author: Richard Walker
Department of Mathematics
Mansfield State College
Mansfield, PA 16933

Review Stage/Date: III 7/30/80

Classification: DESCRIPTIVE STAT

Prerequisite Skills:
1. Be able to calculate with decimals and evaluate simple formulas.

Output Skills:
1. Use frequency distributions and histograms to summarize data.
2. Calculate means, medians, and modes as measures of central location.
3. Decide which measure of central location may be most appropriate in a given instance.
4. Calculate and interpret percentiles.

MODULES AND MONOGRAPHS IN UNDERGRADUATE
MATHEMATICS AND ITS APPLICATIONS PROJECT (UMAP)

The goal of UMAP is to develop, through a community of users and developers, a system of instructional modules and monographs in undergraduate mathematics and its applications which may be used to supplement existing courses and from which complete courses may eventually be built.

The Project is guided by a National Steering Committee of mathematicians, scientists, and educators. UMAP is funded by a grant from the National Science Foundation to Education Development Center, Inc., a publicly supported, nonprofit corporation engaged in educational research in the U.S. and abroad.

PROJECT STAFF

Ross L. Finney	Director
Solomon Garfunkel	Consortium Director
Felicia DeMay	Associate Director
Barbara Kelczewski	Coordinator for Materials Production
Paula M. Santillo	Assistant to the Directors
Donna DiDuca	Project Secretary
Janet Webber	Word Processor
Zachary Zevitas	Staff Assistant

NATIONAL STEERING COMMITTEE

W.T. Martin (Chair)	M.I.T.
Steven J. Brams	New York University
Llayron Clarkson	Texas Southern University
Ernest J. Henley	University of Houston
William Hogan	Harvard University
Donald A. Larson	SUNY at Buffalo
William F. Lucas	Cornell University
R. Duncan Luce	Harvard University
George Miller	Nassau Community College
Walter E. Sears	University of Michigan Press
George Springer	Indiana University
Arnold A. Strassenburg	SUNY at Stony Brook
Alfred B. Willcox	Mathematical Association of America

This material was developed under the auspices of the UMAP Statistics Panel whose members are Thomas Knapp (Chair) of Rochester University, Roger Carlson of the University of Missouri ar Kansas City, Earl Faulkner of Brigham Young University, Peter Purdue of the University of Kentucky, Judith Tanur of SUNY at Stony Brook, Richard Walker of Mansfield State College, and Douglas A. Zahn of Florida State University. The Project would like to thank the members of the Statistics Panel for their reviews, and all others who assisted in the production of this unit.

This material was prepared with the partial support of National Science Foundation Grant No. SED76-19615 A02. Recommendations expressed are those of the author and do not necessarily reflect the views of the NSF or the copyright holder.

BASIC DESCRIPTIVE STATISTICS

1. THE NEED TO SUMMARIZE DATA - AN EXAMPLE

There is a quantitative side to almost every academic field. The geologist measures the hardness of various rock specimens. The psychologist measures reaction times to a certain stimulus. The educator measures learning as it is reflected in scores on achievement tests. The economist records income. The list could be extended for many pages.

After a set of data has been collected the next task is to decide how to best present it so that it is available to others in a quick and useful way. The methods used to do this belong to a branch of study called *descriptive statistics*. Included in descriptive statistics are the methods of collection, organization and description of numerical information. The topics covered in this module are all from the fields of descriptive statistics.

Suppose we have collected the data below.

HEIGHTS OF ONE-HUNDRED-EIGHTY
17 YEAR-OLD FEMALES IN CENTIMETERS (cm)
(hypothetical)

162	157	160	160	162	160	158	148	160	170	160	152
152	162	159	149	166	167	174	159	153	154	164	163
170	161	166	162	158	168	164	164	159	160	165	166
149	160	174	170	167	145	155	154	180	159	154	161
165	167	172	152	171	164	156	156	165	156	156	147
147	157	162	158	170	157	164	161	158	153	148	158
165	159	161	167	157	148	146	169	161	166	151	158
173	161	168	160	164	157	155	170	157	163	156	157
157	160	168	167	166	177	150	154	153	167	149	158
160	156	150	168	168	158	177	157	164	151	160	161
157	168	152	159	168	165	154	157	166	171	160	174
160	160	161	157	153	176	147	167	160	157	158	154
159	160	160	164	145	155	162	154	163	155	160	154
161	149	163	166	162	159	163	162	158	164	160	162
169	158	168	158	162	161	159	163	163	170	165	176

Data in this form are called *raw data*. In this unorganized form the data can only be understood after a certain amount of time-consuming examination. If the data set included several thousand numbers the need to organize and summarize would be even greater.

2. METHODS OF SUMMARIZING DATA

In this section we will discuss two important methods of summarizing data: the frequency distribution and the histogram.

2.1 Frequency Distribution

The simplest way to organize data is by means of a frequency distribution with one value in each class. Such a distribution consists of a list of the values which appear in the data set, arranged in increasing order, and the frequencies which indicate the number of times the various values appear. Such a frequency distribution for the data on page 1 appears below.

HEIGHTS OF 17 YEAR-OLD FEMALES

HEIGHT (in cm)	TALLY	FREQUENCY
145	||	2
146	|	1
147	|||	3
148	|||	3
149	𝍸 ||	4
150	||	2
151	||	2
152	𝍸 ||	4
153	𝍸 ||	4
154	𝍸 |||	8
155	𝍸 ||	4
156	𝍸 |	6
157	𝍸 𝍸 |||	13
158	𝍸 𝍸 ||	12
159	𝍸 𝍸 ||||	9
160	𝍸 𝍸 𝍸 ||||	19
161	𝍸 𝍸	10
162	𝍸 𝍸	10
163	𝍸 ||	7
164	𝍸 ||||	9
165	𝍸 |	6

166	卌	7
167	卌	7
168	卌 |||	8
169		2
170	卌	6
171		2
172		1
173		1
174					3
175		0
176				2
177				2
178		0
179		0
180		1

TOTAL = 180

The tallies in the middle column above are included only as an indication of how the frequency distribution was obtained. It is not necessary, or even desirable, to include these tallies with a frequency distribution.

Already we have made significant progress in the process of summarizing the data. This frequency distribution allows us to "get a feeling" for the data much more quickly than was possible from the raw data. Furthermore, nothing has been lost. All of the information which was available from the raw data is available in this frequency distribution. This summary is, however, less than perfect. There are 37 different classes; it takes nearly a full page to present this frequency distribution; and even with the data in this form it takes some time to digest it.

The situation might have been worse. Each height in this data set has apparently been rounded to the nearest centimeter. If, instead, each height were rounded to the nearest tenth of a centimeter then there would have been many more classes and each class would have a very small frequency. In such a case the frequency distribution would represent only a small improvement over the raw data because it contains too much detailed information; there are too many different values.

In other cases it may happen that a frequency distribution of the type just given is a very effective summary. For example, the frequency distribution shown below gives a quick and accurate description of the number of games played in the World Series of Baseball.

NUMBER OF GAMES IN THE WORLD SERIES (1923-1978)

No. of Games	Frequency
4	11
5	10
6	11
7	24
	TOTAL = 56

Let us return to the set of data representing heights. We can condense the frequency distribution on page 2 by using intervals as our classes, rather than individual values. For example:

HEIGHTS OF 17 YEAR-OLD FEMALES

HEIGHT (in cm)	FREQUENCY
144.5--150.5	15
150.5--156.5	28
156.5--162.5	73
162.5--168.5	44
168.5--174.5	15
174.5--180.5	5

The first class contains all of the heights which fall between 144.5 cm. and 150.5 cm. The number 144.5 is called the *lower boundary* of the class and 150.5 is called the *upper boundary*. Note that the upper boundary of one class is the lower boundary of the next class. In this example the class boundaries have been chosen in such a way that no number from the data set is equal to a class boundary. Thus each number can be placed in one and only one class. By selecting class boundaries which contain one more significant digit than the data

it is always possible to choose these boundaries so that they are distinct from the data. This is desirable in order to avoid ambiguity.

The *midpoint*, or class mark, of each class interval may be found by adding the upper and lower class boundaries and dividing the sum by 2. In the frequency distribution given above the class marks are 147.5, 153.5, 159.5, 165.5, 171.5 and 177.5. The width of each class interval is called the *class width*. The class width may be found by subtracting the lower class boundary from the upper. Each class in the example has a class width of 6. It is desirable, but not necessary, to have all classes of the same width.

A frequency distribution which uses class intervals is called a *grouped* frequency distribution and the data in such a frequency distribution is called *grouped data*. The frequency distribution given on page 2 is sometimes called an *ungrouped* frequency distribution.

The grouped frequency distribution has been obtained at the cost of a certain loss of information. While the frequency distribution has been obtained from the raw data, the raw data cannot be recovered from the frequency distribution. For example, in the frequency distribution for heights we know that fifteen numbers lie between 144.5 and 150.5. But that is all we can tell. The exact values of these fifteen numbers cannot be determined from the frequency distribution.

Exercise 1. Forty students in a chemistry course did a laboratory experiment to determine the pH of a solution. The results are recorded below.

8.00	8.15	8.10	8.15	8.05
8.20	8.00	7.95	8.05	8.15
8.05	8.10	8.10	8.15	8.25
8.20	8.10	8.30	8.15	8.20
8.05	8.15	8.00	8.20	8.10
8.25	8.30	8.15	8.20	8.10
8.05	8.25	8.05	8.15	8.00
8.10	8.05	8.15	8.25	8.05

a. Construct a frequency distribution for these data in which each class consists of single value.

b. Construct a grouped frequency distribution for these data in which the boundaries of the first class are 7.895 and 7.995. Use classes of equal width.

Exercise 2. Thirty laboratory rats are run through a maze. The time required to complete the maze on the first run is recorded below for each rat. The times are in seconds.

10.8	23.2	11.6	13.1	16.1	17.7
17.5	15.9	42.9	16.0	56.2	14.1
38.3	15.7	15.3	19.8	14.8	39.7
16.9	29.8	14.0	21.3	13.3	11.8
14.4	18.3	34.6	13.9	20.3	10.7

Construct a frequency distribution for these data.

2.2 Histograms

A picture is worth a thousand words. If this is so then it makes sense to find a pictorial method of presenting data. The *histogram* is such a method. The histogram below is based on the frequency distribution for height data on page 4.

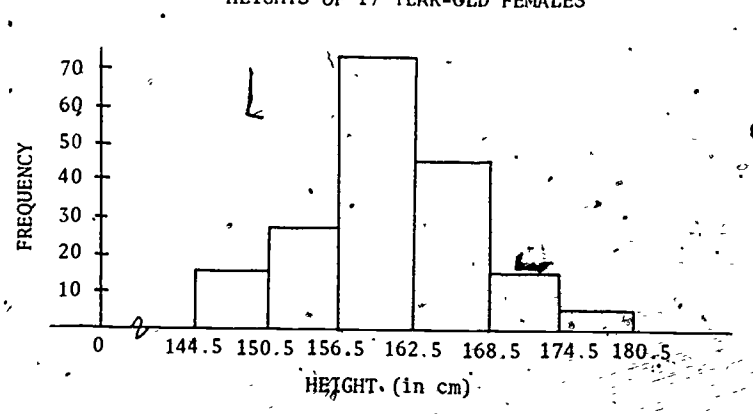


Figure 1. Histogram of height data.

On the horizontal axis in Figure 1 we see the class boundaries from the frequency distribution on page 4. On the vertical axis we see class frequencies. The *areas* of the rectangles in the histogram must be proportional to the frequencies of the classes which they represent. If, as in our example, all classes have the same class width then the area of each rectangle is proportional to its height. In this case the height of each rectangle may be thought of as representing the frequency of the corresponding class. The use of a vertical axis for frequencies is, in this case, desirable and recommended. However, should the frequency distribution contain classes of varying widths then a vertical axis for frequencies is impossible and must be avoided. (See the solution to Exercise 4, below, for an example of a histogram with unequal class widths.)

Exercise 3. Draw a histogram for the frequency distribution in Exercise 1, part b on page 6.

Exercise 4. Draw a histogram for the frequency distribution in Exercise 2 on page 6.

---

3. MEASURES OF LOCATION - ANOTHER METHOD OF SUMMARIZING DATA

In many cases an even more drastic summary of the data is required. For example, we might seek a single number that can be thought of as representative of the entire set of data. Such numbers are called *averages*, or *measures of location*, or *measures of central tendency*, or *measures of position*. We shall call them measures of location. This conveys the important idea that such measures tell us where the data are, or, equivalently, how large the data are. At the same time it avoids the word "average" to which some people are prone to give improper interpretations.

There are many measures of location. In this section we will discuss three of the most useful: the *mean*, the *median* and the *mode*. Each of these may be thought of as, in some sense, locating the center of the data.

3.1 The Arithmetic Mean

The most common measure of location, the one most people are thinking of when they say "the average of these numbers is such-and-such", is the *arithmetic mean*. Although there are other means than the arithmetic mean (for example: the geometric mean or the harmonic mean) when the word *mean* is used alone it is safe to assume that the arithmetic mean is the mean to which we are referring.

3.1.1 Computing the Mean from Raw Data

The arithmetic mean is the number obtained by adding all of the numbers together and dividing this sum by the number of numbers. For example, the mean of 6, 11, 7 and 5 is $(6 + 11 + 7 + 5)/4 = 29/4 = 7.25$.

If the variable x is used to represent the individual numbers in the data set, then $\bar{x}$ is used as a symbol for the mean. If the variable y were used to represent the individual numbers then $\bar{y}$ would be the mean of these numbers, and similarly for other variable names.

Let us use n to represent the number of numbers in a set of data. If we use x to represent the individual numbers then Σx will be used to represent the *sum* of the numbers. Then we have the following formula for the mean:

$$\bar{x} = \frac{\Sigma x}{n}$$

For example, if the data set consists of the numbers 6.2, 5.8, 2.9, 3.3 and 4.1 then $n = 5$, $\Sigma x = 22.3$ and

$$\bar{x} = \frac{22.3}{5} = 4.46.$$

For the data on page 1, $n = 180$, $\Sigma x = 28900$ and

$$\bar{x} = \frac{28900}{180} = 160\frac{5}{9} \doteq 160.6.$$

The symbol "$\doteq$" indicates approximate equality and is used here to indicate that the final answer has been rounded.

3.1.2 Computing the Mean from a Frequency Distribution

Sometimes the data are available to us only in the form of a frequency distribution. Thus it is necessary for us to have a method for calculating the mean from a frequency distribution. If the frequency distribution has only one value in each class, we use the following method:

a. Multiply each value by the corresponding frequency and add the products.
b. Add the frequencies to obtain n.
c. Divide the first number by the second.

This method is illustrated below using the World Series data from page 4.

NUMBER OF GAMES	FREQUENCY	
x	f	$x \cdot f$
4	11	44
5	10	50
6	11	66
7	24	168
	$\Sigma f = 56$	$\Sigma(x \cdot f) = 328$

$$\bar{x} = \frac{328}{56} \doteq 5.9$$

This method can be expressed as a formula:

$$\bar{x} = \frac{\Sigma(x \cdot f)}{\Sigma f}$$

If the classes in the frequency distribution are intervals rather than individual values it is not possible to compute the mean exactly. This is because we cannot determine the exact value of each piece of data. It is, however, possible to make a very good approximation of the mean.

The sum of the numbers in each interval can be found approximately by multiplying the class frequency by the class midpoint. Thus the mean may be approximated by using the same formula as before:

$$\bar{x} \doteq \frac{\Sigma(x \cdot f)}{\Sigma f}$$

But now the x on the right hand side represents the midpoint of the class. The next example illustrates the use of this formula for the height data from the frequency distribution on page 4.

HEIGHT (in cm)	FREQUENCY f	CLASS MARK x	$x \cdot f$
144.5--150.5	15	147.5	2212.5
150.5--156.5	28	153.5	4298.0
156.5--162.5	73	159.5	11643.5
162.5--168.5	44	165.5	7282.0
168.5--174.5	15	171.5	2572.5
174.5--180.5	5	177.5	887.5
	$\Sigma f = 180$		$\Sigma(x \cdot f) = 28896.0$

$$\bar{x} = \frac{28896}{180} \doteq 160.5.$$

How does this answer compare with the value of $\bar{x}$ obtained from the raw data? Can you account for the difference?

3.1.3 Properties of the Mean

The advantages of the mean as a measure of location include:

a. It is the most commonly used measure of location and thus is familiar to many people.
b. It is relatively easy to compute.
c. It lends itself to algebraic manipulation.
d. Each number in the data set has as effect on the mean.
e. The mean is the most stable measure of location under repeated sampling.

The last statement above requires some explanation. As we become more knowledgeable about statistics we find that the data which we have in hand, called a *sample*, is often just a fraction of some larger set of data called a *population*. It is of central importance to use the data in the sample to draw inferences about the population. The study of how this is done is called *inferential statistics*. One of the reasons that the mean is often used in drawing inferences is that the variability of the mean among several samples is less than the variability of other measures of location. This is what we mean when we say "the mean is the most *stable* measure of location under repeated sampling."

The chief disadvantage of the mean as a measure of location is that it is unduly affected by extreme values. For example, the mean of 6, 7, 500 and 3 is 129, which does not seem representative of the original numbers.

Exercise 5. Compute the mean of the data given in Exercise 1 on page 5.

Exercise 6. Compute the mean of the data given in Exercise 1 on page 5 from the ungrouped frequency distribution obtained in part a of that exercise.

Exercise 7. Approximate the mean of the data given in Exercise 1 on page 5 from the grouped frequency distribution obtained in part b of that exercise.

Exercise 8. Compare the results of Exercises 5, 6, and 7.

Exercise 9. Compute the mean of the data given in Exercise 2 on page 6.

Exercise 10. Approximate the mean of the data in Exercise 2 on page 6 from the frequency distribution obtained in that exercise.

3.2 The Median

For a given set of data, a number which is greater than half of the data and less than the other half would be a useful measure of location. In practice there may be no such number. For example, if the numbers in the data set are 3, 4, and 5 then the number in the middle is 4. But only one-third of the data are smaller than 4. In order to insure that the measure we are defining will always exist we must make a slightly more elaborate definition.

The median of a set of data is a number which: a) is not greater than more than half of the data, and b) is not less than more than half of the data. If the variable x is used to represent the individual numbers in the data set then $\tilde{x}$ will be used to represent the median.

3.2.1 Computing the Median from Raw Data

To calculate the median it is first necessary to rank the data from smallest to largest. The median is then the "number in the middle."

If n, the number of numbers, is odd then this number in the middle is easy to find. For example, to find the median of 11, 17, 12, 23 and 13 we rank the

data (11, 12, 13, 17, 23) and observe that the number in the middle is 13. This is the median.

If n is even then there is a small problem. If, for example, the ranked data are 7, 9, 10, 15, 18 and 20 then any number between 10 and 15 satisfies the definition of the median. To be technically correct we should speak of *a* median rather than *the* median. But this ambiguity is avoided if we define the median in this case to be the mean of the two numbers in the middle of the ranked data. By this agreement the median of 7, 9, 10, 15, 18 and 20 is

$$\tilde{x} = \frac{10 + 15}{2} = 12.5$$

In both examples above, no matter whether n is even or odd, the median is the number in the $\frac{1}{2}(n+1)$ position in the ranked data. When n was 5, $\frac{1}{2}(n+1)$ was 3 and the median was the third number in the ranked data. When n was 6, $\frac{1}{2}(n+1)$ was $3\frac{1}{2}$ and the median was halfway between the third and fourth numbers in the ranked data. Thus the procedure for finding the median from raw data may be summarized as follows:

a. Rank the data.
b. Find the number in the $\frac{1}{2}(n+1)$ position in the ranked data.

3.2.2 Computing the Median from a Frequency Distribution

If the data are available to us in a frequency distribution then the data have, in effect, been ranked. If each class in the distribution contains a single value we need only determine the position of the median and find the number in that position.

For example, in the distribution of height data on page 2, n = 180. Thus the position of the median is $\frac{1}{2}(181) = 90.5$, or halfway between the 90th and 91st numbers. Adding the frequencies from the first class onward we find that 77 numbers are in the classes up to

and including 159 cm. and 96 of the numbers are in the classes up to and including 160 cm. Thus both the 90th and 91st numbers are equal to 160 cm. and the median

$$\tilde{x} = \frac{160 + 160}{2} = 160.$$

If the classes in the frequency distribution are intervals then, as with the mean, we cannot calculate $\tilde{x}$ exactly, but only approximate it. The procedure used to approximate the median is as follows:

a. Find the position of the median: $\frac{1}{2}(n+1)$.
b. Find the class which contains the median.
c. Use the formula

$$\tilde{x} = L + \left(\frac{\frac{1}{2}(n+1) - S}{f}\right) w$$

where: L = lower boundary of the class containing the median.
S = sum of frequencies for classes lower than the class containing the median.
f = frequency of the class containing the median.
w = width of the class containing the median.

Applying this rule to the grouped data on heights on page 4 we find:

a. Position of the median $\frac{1}{2}(181) = 90.5$.
b. The median is in the third class (156.5-162.5).
c. L = 156.5, S = 15 + 28 = 43, f = 73, w = 6,

$$\tilde{x} = 156.5 + \left(\frac{90.5 - 43}{73}\right) 6 = 156.5 + 3.9 = 160.4.$$

This answer compares favorably with the exact result, 160, obtained above.

3.2.3 Properties of the Median

The median has the following advantages:

a. It is an easily understood measure of location.
b. It is not affected by extreme values and thus is sometimes more typical of the numbers in the data set than is the mean.

c. Unlike the mean, the median may be used to summarize non-numerical data if there is an order among the categories. Such data are called *ordinal data*. Suppose, for example, that 250 students in a college course receive the following grades: 58 A's, 72 B's, 80 C's, 25 D's, and 15 F's. Then it is meaningful to say that the median course grade is B.

d. In many frequency distributions the smallest or largest class is open-ended. For example, in reporting the number of children in a family the top class might be "10 or more." Since there is no mid-point of this top class it is difficult to approximate the mean of such data. But since the top class is not usually involved in the process of finding the median, it may be found as before.

The chief disadvantage of the median is that it does not lend itself to algebraic manipulation as readily as does the mean. We might also regard the necessity to rank the data as a disadvantage. For large sets of data the ranking procedure is time consuming, even if done on a computer.

Exercise 11. Compute the median of the data on World Series games given in the frequency distribution on page 4.

Exercise 12. Compute the median of the data in Exercise 1 on page 5. Compare $\tilde{x}$ with $\bar{x}$ for this data set.

Exercise 13. Compute the median of the data in Exercise 1 on page 5 from the frequency distribution constructed in part b of that exercise. Compare this with the result obtained in Exercise 12.

Exercise 14. Compute the median of the data in Exercise 2 on page 6. Compare $\tilde{x}$ with $\bar{x}$ for this data set.

3.3 The Mode

The *mode* of a set of numbers is simply the number which appears more frequently than any other. For example, in the data set presented on page 1 (and again on page 2) the mode is 160.

If all of the numbers in the data set are distinct then there is no mode. Even when there is a mode it may be of no particular importance. If a data set consists of 100 values, with two of these being equal and the remainder distinct, it is unlikely to be of any use to note that the value which occurs twice is the mode.

On the other hand, if the mode represents some relatively large fraction of the data, it is useful to report it. In the data on World Series games on page 4 we see that nearly half of the World Series have taken seven games to complete. This is an interesting feature of the data. Thus it makes some sense to mention this if the four class frequencies had been 11, 10, 11 and 12. The importance of the mode as a measure of location is directly related to the relative frequency of this value: The larger the fraction of the data represented by the mode, the more important the mode becomes.

Sometimes a data set will have two values which occur much more frequently than the others. For example, the salaries of employees of a business might fall mainly into two categories, low salaries for laborers and higher salaries for management personnel. Such a data set is said to have two modes, even if the frequency for one mode is somewhat larger than for the other. Such data may also be described as *bimodal*. It is appropriate to report *both* modes for bimodal data.

If the data are in a grouped frequency distribution we may choose the class with the largest frequency

and call this the *modal class*. Alternatively, the midpoint of the modal class may be reported as the mode.

In short, if one or two values, or intervals, represent a relatively large fraction of the data then this is interesting and should be mentioned when describing the data. Otherwise we should not use the mode as a measure of location.

4. CHOOSING A MEASURE OF LOCATION

Now that we have three measures of location at our disposal, which one should we use? The answer to this question depends both on the data set itself and on the use we intend to make of the measure of location once it has been found. If our purpose is simply to describe the data effectively we should use whatever measure or measures are suggested by the data.

The shape of the histogram of a data set is useful in deciding what measure to use. Four possibilities are illustrated in Figure 2.

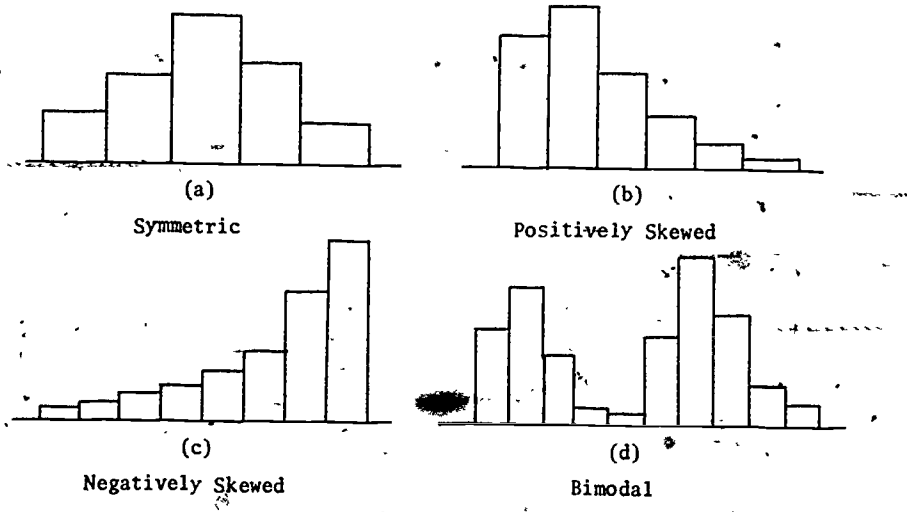


Figure 2.

If the histogram of the data is approximately *symmetric*, as in Figure 2a, then the mean and the median will be approximately equal. If the histogram is approximately symmetric and has a single modal class then the mean, median and mode are all approximately equal. If the data are concentrated toward the lower end of the range with a few larger values, as in Figure 2b, then we say the data are *positively skewed*. The reverse case, illustrated in Figure 2c, is referred to as *negatively skewed* data. The more the data are skewed, the greater will be the difference between the mean and the median.

The histrogram on page 6, which represents the height data given on page 1, is approximately symmetric. For this data set the mean was 160.6, the median was 160 and the mode was 160. The data set summarized in the frequency distribution below is negatively skewed.

Class	Frequency
0.5--100.5	3
100.5--200.5	2
200.5--300.5	7
300.5--400.5	24
400.5--500.5	52

For this data set $\bar{x} \doteq 387$, $\tilde{x} \doteq 417$ and the midpoint of the modal class is 450.5.

The outstanding characteristic of the data represented by the histogram in Figure 2d is that it is biomodal. This fact should be included in any description of the data.

If we intend to follow the calculation of the measure of location with further statistical computations then this fact must be considered when choosing the measure of location. The great majority of statistical tests and procedures are designed to use the mean rather than some other measure of location. Hence

there is a strong inclination to choose the mean in those cases where further statistical investigation is anticipated.

With these facts in mind we list below some suggestions.

1. In general, *use the mean*. It is the most commonly used measure. It is especially appropriate if you expect to do further statistical computations.

2. If the data are highly skewed, use the median. The median is, in general, less affected by a small number of very extreme values than is the mean.

3. If the data are in a frequency distribution which uses an open-ended interval, use the median.

4. If the data have a pronounced mode, mention this fact. If the data have two pronounced modes, mention this also.

5. There is no law which forbids you to report more than one measure of location.

Exercise 15. The frequency distribution below, taken from the 1978 edition of the Statistical Abstract of the United States, gives adjusted gross incomes as reported on individual income tax returns in 1976. Which measure of location is most appropriate for these data, and why?

ADJUSTED GROSS INCOME (IN DOLLARS)	NUMBER OF TAXPAYERS (IN THOUSANDS)
0 to 3,000	15,015
3,000 to 5,000	8,837
5,000 to 10,000	19,891
10,000 to 15,000	14,182
15,000 to 20,000	11,182
20,000 to 25,000	6,662
25,000 to 30,000	3,611
30,000 to 50,000	3,632
50,000 to 100,000	945
100,000 to 500,000	221
500,000 to 1,000,000	4
over 1,000,000	1

Exercise 16. Suppose that two hundred film reviewers were asked to choose, from among the five films listed below, their favorite. Suppose further that the responses were as indicated. What measure of location is most appropriate for these data and Why?

PICTURE	NUMBER
High Noon	2
The Godfather	55
Gone With the Wind	90
The Sound of Music	8
Casablanca	40

Exercise 17. On an opinionnaire 450 people were asked to state whether they "strongly agree," "agree," are "neutral," "disagree" or "strongly disagree" with the following statement: "Gas rationing is one good way to deal with the energy shortage." The results of this (hypothetical) poll are presented below. Which measure of location is appropriate for these data and why?

RESPONSE	NUMBER
Strongly Agree	54
Agree	97
Neutral	150
Disagree	103
Strongly Disagree	46

Exercise 18. The grades of thirty high school students on a French examination are recorded below. Which measure of location is appropriate for these data and why?

80	84	79	81	75	68
76	72	90	96	85	86
88	85	70	92	87	90
80	80	72	73	84	91
64	76	71	76	81	68

Exercise 19. What measure of location would be appropriate for the data given in Exercise 1 on page 5, and why?

Exercise 20. What measure of location would be appropriate for the data given in Exercise 2 on page 6, and why?

---

5. PERCENTILES, DECILES AND QUARTILES

The measures discussed in this section are measures of location or position, but are not properly described as measures of central tendency. These are the percentile scores, decile scores and quartile scores. Percentiles will be described in detail. Deciles and quartiles may be thought of a special cases of percentiles.

5.1. Percentiles

Percentiles are defined and computed in a matter analogous to the median. As with the median, care must be taken to insure that percentiles exist and are unique. To begin with an example, the *eightieth percentile*, denoted by P_{80}, may be thought of as a number which is larger than 80% of the data and smaller than 20% of the data. Similarly, the thirty-fifth percentile, P_{35}, may be thought of as a number which is larger than 35% of the data and smaller than 65% of the data. The formal definition is given below.

If r is any number from 1 to 99 then the *rth percentile* for a set of data is a number, P_r, such that at most r% of the data are less than P_r and at most $(100-r)$% of the data are greater than P_r.

5.2 Computing Percentiles

The method for finding a percentile score is very similar to that for finding the median. In fact you may have already noticed that the fiftieth percentile and the median are identical. To find the rth percentile:

a. Rank the data.

b. Find the number in the $\frac{r}{100}(n+1)$ position in the ranked data.

Suppose for example that we wish to find the 84th percentile score for the height data given on page 1. The data have been ranked in the frequency distribution on page 2.

The position of P_{84} is

$$\frac{84}{100}(n+1) = \frac{84}{100}(181) = 152.04.$$

Thus P_{84} is between the 152nd and 153rd numbers in the ranked data. To avoid ambiguity we will take P_{84} to be four one-hundreds of the way between these two numbers. That is

P_{84} = 152nd number + 0.04 (153rd number - 152nd number).

Counting through the frequency distribution from the smallest class we find that the 152nd number is 167 and the 153rd number is 168. Thus

$P_{84} = 167 + 0.04(168 - 167) = 167 + 0.04 = 167.04$.

Exercise 21. Find P_{24} and P_{75} for the height data on page 2.

If the data are given in a frequency distribution with class intervals then the method for finding P_r is similar to the method for finding the median given on page 12. The position of P_r is, as before,

$$\frac{r}{100}(n+1).$$

First we find the class containing this number, and then we define P_r by

$$P_r = L + \left(\frac{\frac{r}{100}(n+1) - S}{f}\right) w$$

where: L = lower limit of the class containing P_r

S = sum of frequencies for classes lower than the class containing P_r

f = frequency of the class containg P_r

w = width of the class containing P_r .

Exercise 22. Compute P_{30} and P_{89} for the height data in the frequency distribution on page 4.

5.3 Deciles and Quartiles

The median divides the data into halves. The percentiles divide the data into hundredths. Similarly, the deciles divide the data into tenths and the quartiles divide the data into quarters. The *sixth decile*, denoted D_6, is that number such that *six-tenths* of the data are less than D_6. The *third quartile*, Q_3, is that number such that *three-quarters* of the data are less than Q_3. Etc.

It is not necessary to present methods for finding quartile and decile scores as these may be found by computing the corresponding percentile scores.

$D_1 = P_{10}$ $\quad$ $D_3 = P_{30}$

$Q_1 = P_{25}$

$D_2 = P_{20}$ $\quad$ $D_4 = P_{40}$

$D_5 = Q_2 = P_{50} = \tilde{x}$

$D_6 = P_{60}$ $\quad$ $D_8 = P_{80}$

$Q_3 = P_{75}$

$D_7 = P_{70}$ $\quad$ $D_9 = P_{90}$

6. MODEL EXAM

1. Compute the mean and the median of the data below:

8.1	9.0	7.5	6.9	9.0
11.3	10.9	8.4	8.3	9.6
7.9	12.5	11.0	10.6	10.5

2. Construct a frequency distribution for the following set of data using 130.5 as the lower boundary of the first class and having all classes of width 15.

189	233	180	181	200
216	215	190	141	165
193	201	177	217	175
168	138	149	199	223
143	148	203	185	183
192	163	168	166	177
140	193	230	181	173
201	136	158	174	195

3. Compute the mean and the median of the data in problem number two from the frequency distribution.

4. Compute Q_3, D_4 and P_{21} from the raw data in problem number two.

5. What are 'positively-skewed' data?

6. When is the *mode* an important measure which should be reported?

7. ANSWERS TO EXERCISES

1.a.

pH	f
7.95	1
8.00	4
8.05	8
8.10	7
8.15	9
8.20	5
8.25	4
8.30	2

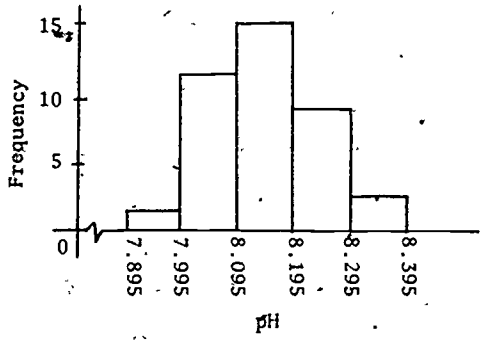
b.

CLASS BOUNDARIES	f
7.895-7.995	1
7.995-8.095	12
8.095-8.195	16
8.195-8.295	9
8.295-8.395	2

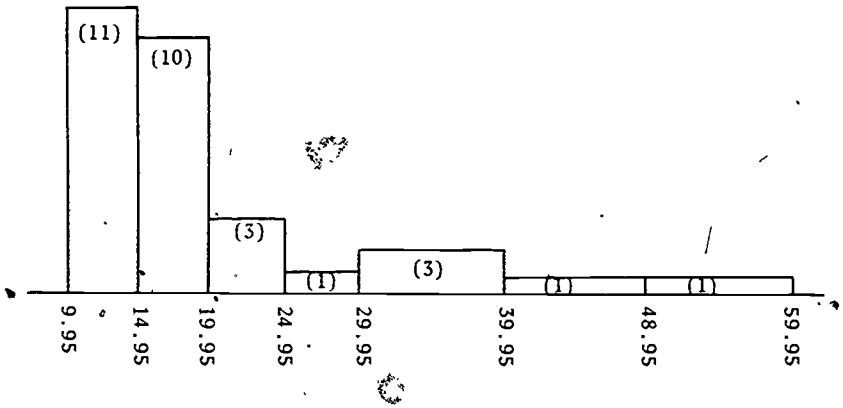
2. The frequency distribution you obtain depends upon your choice of classes. One possible result is shown below.

TIME (in sec.)	f
9.95-14.95	11
14.95-19.95	10
19.95-24.95	3
24.95-29.95	1
29.95-39.95	3
39.95-49.95	1
49.95-59.95	1

3.



4. Your result here depends on your choice of class intervals back in Exercise 2. If you, as I did, chose intervals of varying widths, remember that in a histogram it is the *area* of the rectangle and not its height, which is proportional to the frequency. Note in addition that a vertical axis for frequency is not possible when the classes are of varying widths. The numbers inside parenthesis on this histogram indicate the frequencies of the classes.



5. $n = 40$, $\Sigma x = 325.00$, $\bar{x} = \frac{325}{40} = 8.125$.

6.

x	f	
7.95	1	7.95
8.00	4	32.00
8.05	8	64.40
8.10	7	56.70
8.15	9	73.35
8.20	5	41.00
8.25	4	33.00
8.30	2	16.60
	$\Sigma f = 40$	$\Sigma(x \cdot f) = 325.00$

$\Sigma f = 40$

$\Sigma(x \cdot f) = 325.00$

$\bar{x} = \frac{325}{40} = 8.125$

7.

CLASS	f	x	x · f
7.895-7.995	1	7.945	7.945
7.995-8.095	12	8.045	96.540
8.095-8.195	16	8.145	130.320
8.195-8.295	9	8.245	74.205
8.295-8.395	2	8.345	16.690
	$\Sigma f = 40$		$\Sigma(x \cdot f) = 325.700$

$\bar{x} = \frac{325.7}{40} = 8.1425$

8. The mean obtained in Exercise 6 agrees exactly with the mean obtained in Exercise 5, as it should. The mean of these data is 8.125. The mean obtained in Exercise 7 is only an approxi-

mation to the true mean. This loss of exactness is caused by the loss of information which occurs when the data are grouped into class intervals. Notice that the error of approximation is not large.

9. $n = 30$, $\Sigma x = 618.2$, $\bar{x} = 20.61$.

10.

CLASS	f	x	x · f
9.95-14.95	11	12.45	136.95
14.95-19.95	10	17.45	174.50
19.95-24.95	3	22.45	67.35
24.95-29.95	1	27.45	27.45
29.95-39.95	3	34.95	104.85
39.95-49.95	1	44.95	44.95
49.95-59.95	1	54.95	54.95
	$\Sigma f = 30$		$\Sigma(x \cdot f) = 611.00$

$\bar{x} = \frac{611}{30} = 20.67$.

The answer to this exercise depends upon your choice of class intervals in Exercise 2.

11. $n = \Sigma f = 56$. Position of $\tilde{x} = \frac{1}{2}(n+1) = \frac{1}{2}(57) = 28.5$. The 28th and 29th numbers are both 6. Hence $\tilde{x} = 6$.

12. The data have already been ranked in Exercise 1, part a. $n = 40$. The position of $\tilde{x} = \frac{1}{2}(40+1) = 20.5$. The 20th number is 8.10 and the 21st is 8.15. Thus $\tilde{x} = (8.10 + 8.15)/2 = 8.125$.

We note that the mean and the median are equal. Although exact equality is something of a coincidence, the mean and the median of a data set will be approximately equal whenever the histogram of the data is symmetric. This point will be discussed further in Section 4.

13. The position of the median is 20.5, as in Exercise 12. The median is in the third class. $L = 8.095$, $S = 1 + 12 = 13$, $f = 16$, $w = 0.10$.

$$\tilde{x} = L + \left(\frac{\frac{1}{2}(n+1) - S}{f}\right) w = 8.095 + \left(\frac{20.5 - 13}{16}\right) 0.10 =$$

$$8.095 + 0.047 = 8.142.$$

The approximate value of the median obtained here is reasonably close to the true value obtained in Exercise 12.

14. First we rank the data: 10.7, 10.8, 11.6, 11.8, 13.1, 13.3, 13.9, 14.0, 14.1, 14.4, 14.8, 15.5, 15.7, 15.9, 16.0, 16.1, 16.9, 17.5, 17.7, 18.3, 19.8, 20.3, 21.3, 23.2, 29.8, 34.6, 38.3, 39.7, 42.9, 56.2. The position of $\tilde{x} = \frac{1}{2}(30+1) = 15.5$. The 15th number is 16.0 and the 16th number is 16.1. Thus $\tilde{x} = (16.0 + 16.1)/2 = 16.05$. The mean for these data was 20.61, which is markedly larger than the median.

15. There are two reasons to choose the *median* as the measure of location for these data. One is that the data are positively skewed, as is usually the case with income data. The other is that the last class is open-ended, which prevents the calculation of the mean unless we are willing to guess at an average value (midpoint) for this class.

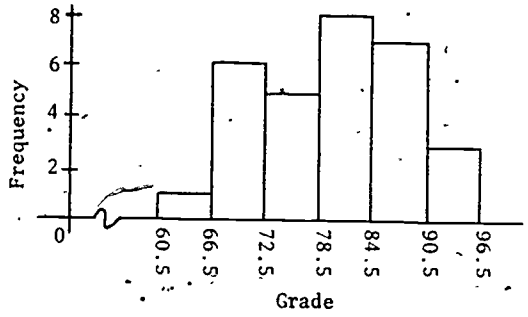
The data also seem to be bimodal, but not to a remarkable degree.

16. In this example the categories are not numerical. In fact they are not even ordered. Thus neither the mean nor the median can be used. This leaves the *mode*. Fortunately there is a pronounced mode: *Gone With The Wind* received the vote of almost half of the people polled.

17. As in Exercise 16, the categories are not numerical. Thus the mean is not a candidate for the measure of location. The categories are, however, ordered. With such ordinal data the median may be used. The position of the median is $\frac{1}{2}(450+1) = 225.5$. The median response is "neutral." This is also the modal response. It seems that this accurately reflects the fact that, according to these responses, opinion on this question is rather evenly divided.

18. A frequency distribution and histrogram for this data set are shown below.

CLASS	f
60.5-66.5	1
66.5-72.5	6
72.5-78.5	5
78.5-84.5	8
84.5-90.5	7
90.5-96.5	3



The histogram above indicates that there is nothing about this data set to indicate we should use a measure other than the mean. Thus we choose the *mean*.

19. As in Exercise 18, we choose the *mean* because there seems to be no strong reason to do otherwise.

20. Choose the *median* because the data are positively skewed.

21. Position of $P_{24} = \frac{24}{100}(180+1) = 43.44$. The 43rd number is 156 and the 44th number is 157. Therefore $P_{24} = 156 + 0.44(157-156) = 156.44$.

Position of $P_{75} = \frac{75}{100}(180+1) = 135.75$. The 135th and 136th numbers are both 165. Thus $P_{75} = 165$.

22. Position of $P_{30} = \frac{30}{100}(180+1) = 54.3$. Thus P_{30} is in third class. $L = 156.5$, $S = 15 + 28 = 43$, $f = 73$ and $w = 6$.

$$P_{30} = 156.5 + \left(\frac{54.3 - 43}{73}\right) 6 = 156.5 + 0.9 = 157.4.$$

Position of $P_{89} = \frac{89}{100}(180+1) = 161.09$. Thus P_{89} is in the fifth class. $L = 168.5$, $S = 160$, $f = 15$ and $w = 6$.

$$P_{89} = 168.5 + \left(\frac{161.09 - 160}{15}\right) 6 = 168.5 + 0.4 = 168.9.$$

8. ANSWERS TO MODEL EXAM

1. a) $n = 15$, $\Sigma x = 141.5$, $\bar{x} = \frac{141.5}{15} \doteq 9.4$.

b) Ranked data: 6.9, 7.5, 7.9, 8.1, 8.3, 8.4, 9.0, 9.0, 9.6, 10.5, 10.6, 10.9, 11.0, 11.3, 12.5.

Position of $\tilde{x} = \frac{1}{2}(15+1) = 8$. $\tilde{x} = 9.0$.

2.

CLASS	f
130.5-145.5	5
145.5-160.5	3
160.5-175.5	8
175.5-190.5	9
190.5-205.5	9
205.5-220.5	3
220.5-235.5	3

3.

CLASS	f	x	x · f
130.5-145.5	5	138	690
145.5-160.5	3	153	459
160.5-175.5	8	168	1344
175.5-190.5	9	183	1647
190.5-205.5	9	198	1782
205.5-220.5	3	213	639
220.5-235.5	3	228	684
	$\Sigma f = 40$		$\Sigma(x \cdot f) = 7245$

a) $\bar{x} = \frac{7245}{40} = 181.125 \doteq 181$.

b) Position of $\tilde{x} = \frac{1}{2}(40+1) = 20.5$

$$\tilde{x} = 175.5 + \left(\frac{20.5 - 16}{9}\right) 15 = 183.$$

4. Ranked data:

136	163	177	190	201
138	165	177	192	203
140	166	180	193	215
141	168	181	193	216
143	168	181	195	217
148	173	183	199	223
149	174	185	200	230
158	175	189	201	233

a) Position of $Q_3 = \frac{3}{4}(40+1) = 30.75$

$Q_3 = 199 + 0.75(200-199) = 199.75$.

4. b) Position of $D_4 = \frac{4}{10}(40+1) = 16.4$

$D_4 = 175 + 0.4(177-175) = 175.8.$

c) Position of $P_{21} = \frac{21}{100}(40+1) = 8.61$

$P_{21} = 158 + 0.61(163-158) = 161.05.$

5. See pages 17-18.

6. See page 16.

STUDENT FORM 2

Unit Questionnaire

Return to:
EDC/UMAP
55 Chapel St.
Newton, MA 02160

Name_____ Unit No._____ Date_____

Institution_____ Course No._____

Check the choice for each question that comes closest to your personal opinion.

1. How useful was the amount of detail in the unit?

____Not enough detail to understand the unit
____Unit would have been clearer with more detail
____Appropriate amount of detail
____Unit was occasionally too detailed, but this was not distracting
____Too much detail; I was often distracted

2. How helpful were the problem answers?

____Sample solutions were too brief; I could not do the intermediate steps
____Sufficient information was given to solve the problems
____Sample solutions were too detailed; I didn't need them

3. Except for fulfilling the prerequisites, how much did you use other sources (for example, instructor, friends, or other books) in order to understand the unit?

____A Lot ____Somewhat ____A Little ____Not at all

4. How long was this unit in comparison to the amount of time you generally spend on a lesson (lecture and homework assignment) in a typical math or science course?

Much Longer	Somewhat Longer	About the Same	Somewhat Shorter	Much Shorter
____	____	____	____	____

5. Were any of the following parts of the unit confusing or distracting? (Check as many as apply.)

____Prerequisites
____Statement of skills and concepts (objectives)
____Paragraph headings
____Examples
____Special Assistance Supplement (if present)
____Other, please explain_____

6. Were any of the following parts of the unit particularly helpful? (Check as many as apply.)

____Prerequisites
____Statement of skills and concepts (objectives)
____Examples
____Problems
____Paragraph headings
____Table of Contents
____Special Assistance Supplement (if present)
____Other, please explain_____

Please describe anything in the unit that you did not particularly like.

Please describe anything that you found particularly helpful. (Please use the back of this sheet if you need more space.)

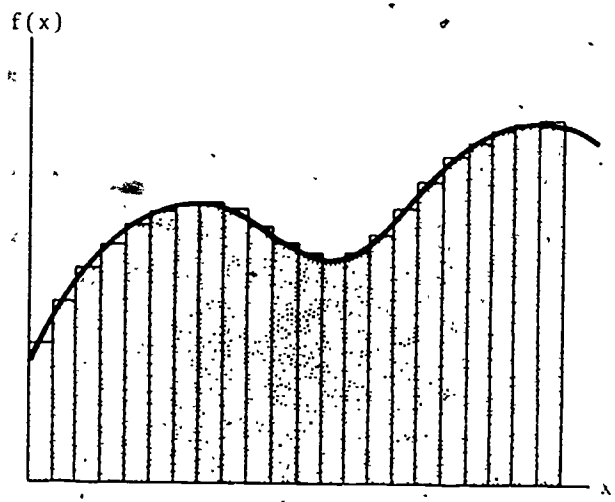
umap

UNIT 443

MODULES AND MONOGRAPHS IN UNDERGRADUATE MATHEMATICS AND ITS APPLICATIONS PROJECT

APPROXIMATIONS IN PROBABILITY CALCULATIONS

by Donald Guthrie and Jolayne Service



APPLICATIONS OF STATISTICS

edc/umap 55chapel st./newton.mass 02160

APPROXIMATIONS IN PROBABILITY CALCULATIONS

by

Donald Guthrie
Department of Psychiatry
and Biobehavioral Sciences
University of California
Los Angeles, CA 90024

and

Jolayne Service
School of Social Sciences
University of California
Irvine, CA 92717

TABLE OF CONTENTS

Intermodular Description Sheet: *UMAP Unit 443*

Title: APPROXIMATIONS IN PROBABILITY CALCULATIONS

Authors: Donald Guthrie and Jolayne Service
Department of Psychiatry and Biobehavioral Sciences
University of California
Los Angeles, CA 90024

Department of Social Sciences
University of California
Irvine, CA 92717

Review Stage/Date: III 7/30/80

Classification: STATISTICS

Suggested Resources: Hand calculator or computer, tables of the standard normal cumulative distribution function and the standard normal density function.

Prerequisite Skills: Elementary acquaintance with concepts of population and sample, random variables, discrete and continuous probability distributions, probability density functions, cumulative distribution functions (in particular, with binomial and normal distributions and tables of the standard normal cumulative distribution function), statistical independence, and central limit theorems. Knowledge of college algebra, including the exponential function and summation notation.

Output Skills: The student will be able to:

1. discuss how approximation is pervasive in statistics,
2. compare "structural" approximations and "mathematical" approximations to probability models,
3. describe and recognize a hypergeometric probability distribution and an experiment in which it holds,
4. recognize when hypergeometric probabilities can be approximated adequately by binomial probabilities (or normal or Poisson probabilities),
5. recognize when binomial probabilities can be approximated adequately by normal or Poisson probabilities,
6. recognize when the normal approximation to binomial probabilities requires the continuity correction to be adequate,
7. calculate with calculator or computer hypergeometric or binomial probabilities exactly or approximately.

MODULES AND MONOGRAPHS IN UNDERGRADUATE
MATHEMATICS AND ITS APPLICATIONS PROJECT (UMAP)

The goal of UMAP is to develop, through a community of users and developers, a system of instructional modules in undergraduate mathematics and its applications which may be used to supplement existing courses and from which complete courses may eventually be built.

The Project is guided by a National Steering Committee of mathematicians, scientists, and educators. UMAP is funded by a grant from the National Science Foundation to Education Development Center, Inc., a publicly supported, nonprofit corporation engaged in educational research in the U.S. and abroad.

PROJECT STAFF

Ross L. Finney	Director
Solomon Garfunkel	Consortium Director
Felicia DeMay	Associate Director
Barbara Kelczewski	Coordinator for Materials Production
Paula M. Santillo	Assistant to the Directors
Donna DiDuca	Project Secretary
Janet Webber	Word Processor
Zachary Zevitas	Staff Assistant

NATIONAL STEERING COMMITTEE

W.T. Martin (Chair)	M.I.T.
Steven J. Brams	New York University
Llayron Clarkson	Texas Southern University
Ernest J. Henley	University of Houston
William Hogan	Harvard University
Donald A. Larson	SUNY at Buffalo
William F. Lucas	Cornell University
R. Duncan Luce	Harvard University
George Miller	Nassau Community College
Walter E. Sears	University of Michigan Press
George Springer	Indiana University
Arnold A. Strassenburg	SUNY at Stony Brook
Alfred B. Willcox	Mathematical Association of America

This module was developed under the auspices of the UMAP Statistics Panel whose members are: Tom Knapp (Chair) of Rochester University; Roger Carlson of University of Missouri, Kansas City; Earl Faulkner of Brigham Young University; Peter Purdue of the University of Kentucky; Judith Tanur of SUNY at Stony Brook; Richard Walker of Mansfield State College, and; Douglas A. Zahn of Florida State University.

This material was prepared with the partial support of National Science Foundation Grant No. SED76-19615 A02. Recommendations expressed are those of the author and do not necessarily reflect the views of the NSF or the copyright holder.

1. INTRODUCTION

1.1 Approximation in Statistics

Approximation plays a central role in the application and interpretation of statistical methods. For instance, parametric probability representations of populations--fundamental tools of statistical analysis-- are usually only approximations of the actual natures of the populations. Sampling distributions in use for these probabalistic models are often themselves approximations to those which are derived mathematically.

There are two principal areas in which approximations are vital in formulating statistical problems: in forming a convenient model of a population when the actual structure of the population is either very complex or unknown; and in developing easy, reasonably accurate methods of computing probabilities when exact methods are cumbersome.

We shall consider experiments consisting of n "trials", where each trial results in one of two possible outcomes (arbitrarily labeled "success" and "failure"). We shall look at two probability models for "the number of successes in the n trials" and study ways to calculate, exactly and approximately, the probability of k successes. While these experiments are of a very special nature, the use of approximations, both structural and mathematical, in this context serve to illustrate the more general application of approximations.

Before turning to approximation of probabilities, however, we shall look at some examples of typical numerical approximations and at a complementary way of making computation more manageable.

1.2 Some Examples of Numerical Approximation

Suppose that, for some reason, we wanted to know

about how large $.7^{10}$ is, but we did not have the time or patience (or the computer) to do all the multiplications. Recalling the algebraic rules for exponents, we can write

$$.7^{10} = (.7^2)^5 = .49^5$$

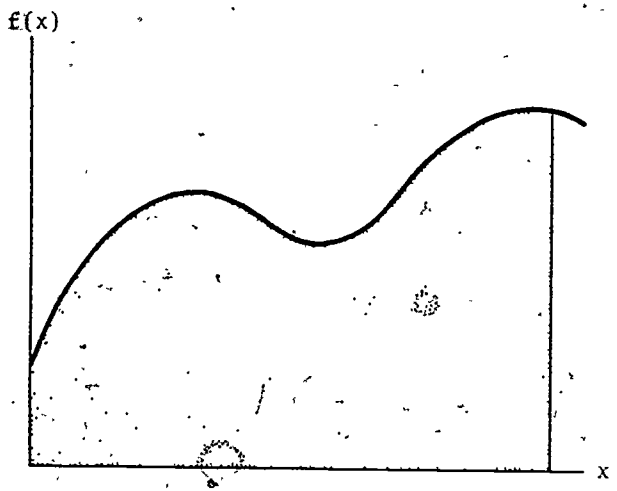
Now, .49 is approximately 1/2. We abbreviate that ".49 ≈ 1/2" (the symbol "≈" means "is approximately equal to"). So

$$.7^{10} \approx (\tfrac{1}{2})^5 = \frac{1}{2^5} = \frac{1}{32} \approx \frac{1}{33} \approx .03.$$

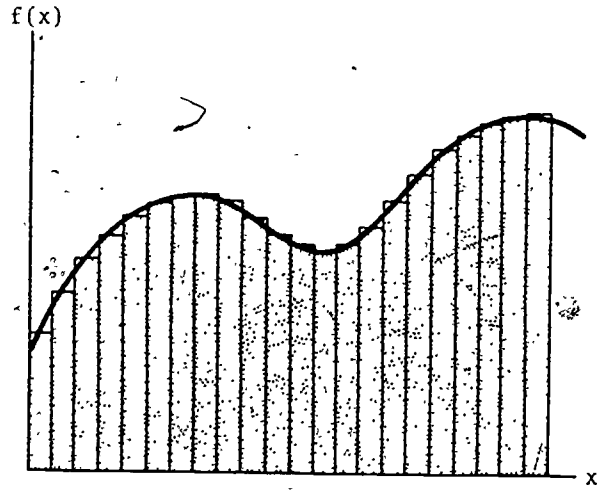
Actually, $.7^{10} = .0282$, to four decimal places, so the approximation is nearly correct. Whether the approximation is close enough depends on the purpose of the calculation. For some applications, especially those which involve further computation using the results of the approximation, a simple approximation may not be close enough to the value being approximated to be dependable.

Numerical approximation may take more complex forms. A frequently encountered mathematical problem is finding the area under a curve, like the shaded area in Figure 1a. We can approximate the area and perhaps simplify the computation by using a series of rectangles whose total area nearly coincides with the area under the curve (see Figure 1b). The height of each retangle at its center is the height of the curve there. Some corners of the rectangles are above the curve (overestimating the area) and some are below the curve (underestimating the area). If the rectangles are narrow enough, the approximation of the area will be quite accurate. (Students of calculus will recognize that the exact area is given by the definite integral of the function defining the curve.)

Some of our probabilistic approximations will use the reverse of this process: we shall use the area under a continuous curve (which happens to be conveniently tabulated) to approximate the area under a series of narrow rectangles.



a. Area to be approximated (shaded).



b. Rectangles whose area approximates the area under the curve.

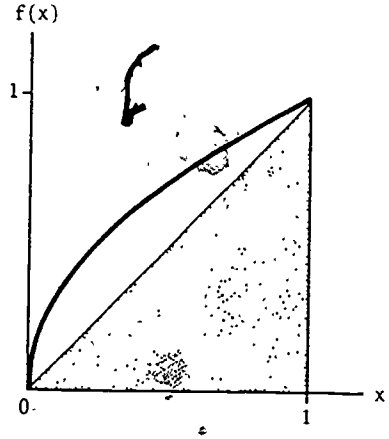
Figure 1. Approximating the area under a curve.

There are some general strategies for designing approximations; they are part of the theory of numerical approximation, which is an important branch of applied mathematics but beyond the scope of this module.

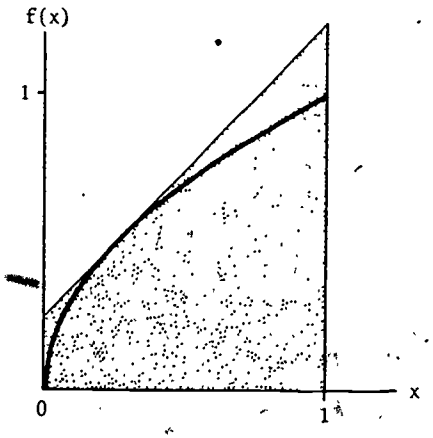
1.3 Exercises

Exercise 1. Approximate the area under the curve defined by $f(x) = \sqrt{x}$ between x=0 and x=1. Try the following methods and compare the approximate areas you compute with the exact area, 2/3.

a) Approximate the area from below, using a straight line:

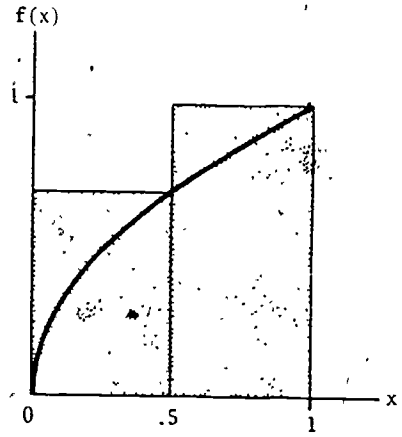


b) Approximate the area from above, using a straight line with the same slope as the line in part (a):

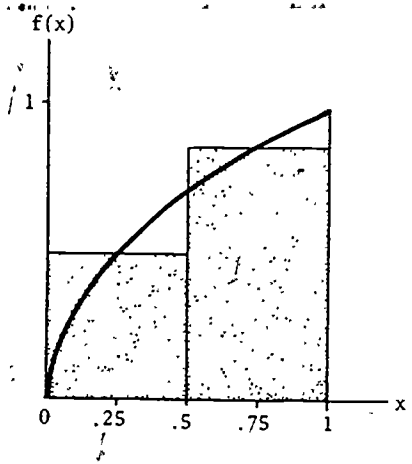


(If you know calculus, you can determine exactly the point at which the line must be tangent to the curve and thus the algebraic representation of the line. If not, you can use graph paper and a ruler to plot f(x), draw the tangent line that has the proper slope, and estimate its height at x=0 and x=1.)

c) Approximate the area using two rectangles, with heights determined by the height of the curve on the right-hand sides of the rectangles:



d) Approximate the area using two rectangles, with heights determined by the height of the curve at the midpoints of the rectangles:



e) Approximate the area using eight rectangles constructed like those in part (c).

f) Approximate the area using eight rectangles constructed like those in part (d).

g) Compare the difference between your answers to (c) and (d) with the difference between your answers to (e) and (f).

1.4 Recursive Formulas

Computing numerical values for a mathematical expression is often easier when the expression is represented as a recursive formula. Simply stated, recursive formulas are "building blocks" which permit the definition (or computation) of the value of a function at some point from the function's value at another point. Usually, some starting value is determined or given, and the function is constructed from this starting value.

For example, consider the function

$$f(k) = k^2$$

for the integers k = 0, 1, 2, ... A recursive representation of the same function could be given by specifying the function's value for 0,

$$f(0) = 0$$

--which is the starting value--and the recursive formula

$$f(k+1) = f(k) + 2k + 1.$$

Table I illustrates the process.

TABLE I

ILLUSTRATION OF RECURSIVE FORMULA $f(k+1) = f(k) + 2k + 1$
(EQUIVALENT TO NON-RECURSIVE FORMULA $f(k) = k^2$.)

k	f(k)	2k + 1
0	0 (starting value)	1
1	1	3
2	4	5
3	9	7
4	16	9
.	.	.
.	.	.
.	.	.

Recursive formulas need not be additive, as our example was. They may involve any kind of mathematical computation. The recursive formulas used in our probability calculations will call for f(k+1) to be determined by multiplying f(k) by several quantities. Multiplicative recursive formulas in particular tend to provide significant reduction in the complexity of computations.

Recursive formulas can also be helpful in suggesting approximations which would hold for large values of one or more of the variables in the expression. Exercise 13 illustrates this use.

1.5 Exercises

Exercise 2. Let $f(0) = 1$ and $f(k+1) = \frac{5-k}{k+1} f(k)$ for $k = 1, 2, 3, 4,$ and 5.

a) Show that $f(k) = \binom{5}{k}$ by computing f(k) recursively, computing $\binom{5}{k}$ directly, and comparing the results.

b) Show algebraically that $f(k) = \binom{5}{k}$. Hint: Prove that

$$\frac{\binom{5}{k+1}}{\binom{5}{k}} = \frac{5-k}{k+1}$$

2. STRUCTURAL APPROXIMATION

2.1 Approximation of Hypergeometric Probabilities by Binomial Probabilities

Suppose that the trials consist of sampling without replacement n items at random from a finite population of N items, K of which are successes. (Sampling without replacement means that an item once chosen for inclusion in the sample cannot be chosen again.) Then the exact probability model for the number of successes is the hypergeometric probability distribution; the probability that k successes are selected is

(1) $$h(k;N,n,K) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} .$$

(We are considering here only values of k that are less than K and also less than n.).

For example, if there are three pink grapefruits and four yellow grapefruits in a bag and three grapefruits are drawn at random, then the probability that exactly one grapefruit in the sample is yellow (a success) and the other two are pink (failures) is given by

$$h(1;7,3,4) = \frac{\binom{4}{1}\binom{3}{2}}{\binom{7}{3}} = \frac{12}{35} = .343 .$$

For this example, N = 7 (the total number of grapefruits in the bag), K = 4 (the number of yellow grapefruits in the bag), n = 3 (the number of grapefruits in the sample), and k = 1 (the number of yellow grapefruits that must appear in the sample to realize the event we described). The mathematical derivation of h(k;N,n,K) is based on counting the total number of possible collections of n items from a population of N items -- which is the denominator, $\binom{N}{n}$ -- and the number of those collections which contain exactly k successes (and n-k failures).

The latter number is the numerator, $\binom{K}{k}\binom{N-K}{n-k}$: there are $\binom{K}{k}$ ways of collecting k successes from among the K successes in the population, and for each of those ways there are $\binom{N-K}{n-k}$ ways of putting together the n-k failures from the N-K failures in the population.

In principle, we could evaluate the hypergeometric probabilities for values of N, K, n and k which should arise. However, for even moderately large values of these four parameters, computation of the binomial coefficients is time-consuming and tedious, and it is useful to have an approximation which involves less tedious calculation.

One of the most convenient methods for simplifying the evaluation of hypergeometric probabilities involves approximating with the binomial probability distribution. This distribution represents the probability of a given number of successes when the results of the trials are statistically independent. If one is sampling with replacement, the probability p of success on any given trial is not affected by the outcomes of previous trials. (In sampling with replacement, an item is "returned" to the population after having been chosen for the sample; so the item could be chosen again.) The trials are independent, and the binomial distribution is applicable. In the hypergeometric situation, if the population size N is small or if the number of trials is an appreciable fraction of N, then the probabilities governing the later trials will be noticeably dependent on the outcomes of the earlier trials. Even when N is large and a very small portion of the population is drawn, the exact probability that k successes will be chosen must be calculated from the hypergeometric probability function, but the effect of dependence is slight when N and K are large. If p is taken to be the proportion of successes in the population (i.e., $p = \frac{K}{N}$), the approximation of the hypergeometric probabilities by binomial probabilities

$$(2) \qquad h(k;N,n,K) \approx b(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

is quite accurate, for $N \geq 20$, or so.

Although N is not large enough for the approximation to be valid, we can demonstrate its application to our previous example. We would approximate h(1;7,3,4) = .343 by

$$b(1;3,4/7) = \binom{3}{1} (4/7)^1 (3/7)^2 = \frac{108}{343} = .315 .$$

TABLE II

ILLUSTRATION OF THE BINOMIAL APPROXIMATION TO HYPERGEOMETRIC PROBABILITIES

N = 7, n = 3, K = 4

(N and K not large enough for approximation to be very accurate)

Number of Successes k	Hypergeometric Probability h(k;7,3,4)	Binomial Approximation b(k;3,4/7)
0	.029	.079
1	.343	.315
2	.514	.420
3	.114	.186
Total	1.000	1.000

Table II shows the exact and approximate probabilities for each of the possible numbers of successes in this example.

This simplification of the calculation of hypergeometric probabilities is based on consideration of the structures of the sampling problems in the two situations. When the population is large -- say, 20 or more times the size of the sample -- sampling with replacement, as in the binomial situation, differs little from sampling without replacement, as in the hypergeometric situation. You are unlikely to select randomly the same item twice from a very large population, even if you are replacing items after sampling them. We can think of such an approximation as a

structural approximation; the structures of the two problems are similar, so the probability distributions are similar.

2.2 Exercises

In performing the following exercises, try to visualize why each of the approximations should be as accurate (or inaccurate) as it is. Use a computer or a calculator to do the calculations. Tabulating the hypergeometric and binomial probabilities is easier when you use the recursive formulas

(3) $$h(k+1,N,n,K) = \frac{(K-k)\ (n-k)}{(k+1)\ (N-K-n+k+1)}\ h(k;N,n,K)$$

and

(4) $$b(k+1;n,p) = \frac{(n-k)p}{(k+1)\ (1-p)}\ b(k;n,p)$$

after calculating $h(0;N,n,k)$ and $b(0;n,p)$ directly.

Exercise 3. Tabulate the hypergeometric probability function and its binomial approximation for:

a) N = 10, n = 5, K = 5
b) N = 10, n = 5, K = 1
c) N = 100, n = 5, K = 50
d) N = 100, n = 5, K = 10

Exercise 4. Repeat parts (c) and (d) of Exercise 3 for n = 20 instead of n = 5. Has the quality of the approximation changed?

Exercise 5. Rose Maybud is choosing at random six members of the United States House of Representatives and determining whether or not each of them supports a particular bill. Explain why this situation is hypergeometric, and identify N, K, n, and k. Which of their values can you determine from our statement of Rose's activity? Would the binomial approximation of the hypergeometric probabilities be adequate? Why?

3. MATHEMATICAL APPROXIMATION

3.1 Approximation of Binomial Probabilities Using the Normal Distribution

When the number of trials n is large, even binomial probabilities are cumbersome to compute, and it helps to have a simple method of approximating them. For large values of n and values of p which are not too close to zero or one, the cumulative binomial distribution distribution function

(5) $$B(k;n,p) = \sum_{i=0}^{k} b(i;n,p)$$

may be approximated by the cumulative normal distribution function thus:

(6) $$B(k;n,p) \approx \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right).$$

The function $\Phi(y)$ is the cumulative distribution function of the standard normal distribution, which has mean zero and variance one. To apply this approximation, you calculate the quantity $y = (k-np)/\sqrt{np(1-p)}$ and refer to a table of the standard normal cumulative distribution function to determine approximately the probability of k or fewer successes in the n trials.

For example, suppose that we are interested in finding the probability of 20 or fewer successes in 56 independent trials, where each trial has probability .45 of resulting in a success. In order to compute this quantity exactly, we would have to add up the binomial probabilities for 21 values of k (0, 1, 2, ..., 20). For each k, we would have to compute the binomial coefficient $\binom{56}{k}$, raise .45 to the power k, and raise .55 to the power 56-k (or at least compute that quantity for k = 0, and then use the recursive formula (4) repeatedly). We might find an answer in a published table of binomial distributions, but such tables do

not cover all possible values of n and p. A computer might be used to perform the calculations, but for values of n much larger than 56, even computer calculation would be rather time-consuming and subject to round-off error. Hence we find the normal approximation attractive.

To apply it, we compute

$$y = \frac{20-56(.45)}{\sqrt{56(.45)(.55)}} = -1.39$$

and refer to a table of the standard normal distribution to find that

$$B(20;56,.45) \approx .081 .$$

By referring to a table of binomial distributions or by computing, we can find the exact value of $B(20;56,.45) = .103$. (For a better approximation see page 15.)

Just as the cumulative binomial distribution may be approximated by the cumulative normal distribution, so may the individual binomial probabilities be approximated by the density function of the normal distribution,

$$(7) \qquad b(k;n,p) \approx \frac{1}{\sqrt{np(1-p)}} \phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right).$$

ϕ is the density function of the standard normal distribution,

$$(8) \qquad \phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} .$$

We approach the derivation of the normal approximation of binomial probabilities somewhat differently from the way we discussed the previous approximation. In that discussion, we noted the structural nature of the binomial approximation of hypergeometric probabilities. The normal approximation, however, is derived from a more intrinsically mathematical formulation, and we consider the nature of the approximation to be more mathematical. That is to say, we chose to employ this particular approximation because of a mathematical derivation rather than an elementary

structural similarity between binomial sampling schemes and those which commonly give rise to normally distributed random variables. A normal random variable is, after all, continuous, while a binomial or hypergeometric random variable is discrete, and it would appear that they are not structurally similar. A less immediately apparent similarity between binomial and normal random variables is revealed, though, by mathematical manipulation. But rather than being a property of these two specific distributions, it applies more generally to the normal distribution. Recall that a Central Limit Theorem states that if $Y_1, Y_2, \ldots, Y_n$ are independent random variables, each with mean μ and finite variance σ^2, then for large n

$$(9) \qquad P(\bar{Y} \leq y) \approx \phi\left(\frac{y-\mu}{\sigma/\sqrt{n}}\right)$$

or equivalently,

$$P\left(\frac{\sqrt{n}\,(\bar{Y}-\mu)}{\sigma} \leq y\right) \approx \phi(y)$$

for any y.

To apply the Central Limit Theorem to the binomial problem, we let Y_i take on the value 1 if the i^{th} trial results in a success or 0 if it results in a failure. Then $\bar{Y}$ is the total number of successes divided by n. The mean of each Y_i is

$$(11) \qquad \mu = \Sigma y p(y) = 0 \cdot (1-p) + 1 \cdot p = p ,$$

and the variance of each Y_i is

$$(12) \qquad \sigma^2 = \Sigma(y-\mu)^2 p(y) = (0-p)^2(1-p) + (1-p)^2 p = p(1-p) .$$

The Central Limit Theorem states that $\bar{Y}$ is approximately normally distributed, so $n\bar{Y}$, the total number of successes in the n trials, is also approximately normally distributed. You should verify that the theorem as stated here leads to the normal approximation given above for binomial probabilities.

The difference in the application of the two types of

approximation--structural and mathematical--is therefore more conceptual than practical.

3.2 Accuracy of the Normal Approximation

The normal approximation to the binomial distribution is quite accurate for situations in which there are both large values of n and values of p not too close to zero or one. Most statisticians regard the approximation as satisfactory whenever np(1-p) is greater than 5. When this condition is violated, one of two alternative approximations may be applicable.

3.3 The Continuity Correction to the Normal Approximation

The first alternative approximation is a refinement of the normal approximation. It involves the use of a "continuity correction". Instead of finding $\Phi(y)$ for $y = \frac{k-np}{\sqrt{np(1-p)}}$, we evaluate it for a slightly different y:

$$B(k;n,p) \approx \Phi\left(\frac{k-np+.5}{\sqrt{np(1-p)}}\right) \tag{13}$$

In effect, this modification assigns to k half the probability between k and k+1 in the normal approximation. (See Figure 2.) Although it generally improves the accuracy of the normal approximation, this refinement is less important for larger n, since the effect on y of the added 1/2 diminishes as n increases. (Compare Exercise 1, part (g).) The continuity correction extends the validity of the normal approximation to considerably smaller n.

To illustrate the application of the continuity correction, we take another look at the example of Section 3.1. The value of y would now be

$$y = \frac{(20-56(.45)+.5)}{\sqrt{56(.45)(.55)}} = -1.262$$

and

$$B(20;56,.45) \approx \Phi(-1.262) = .103 .$$

Notice that this value is the same as the exact value to three decimal places--much closer than the approximation (.081) which was obtained without using the continuity correction.

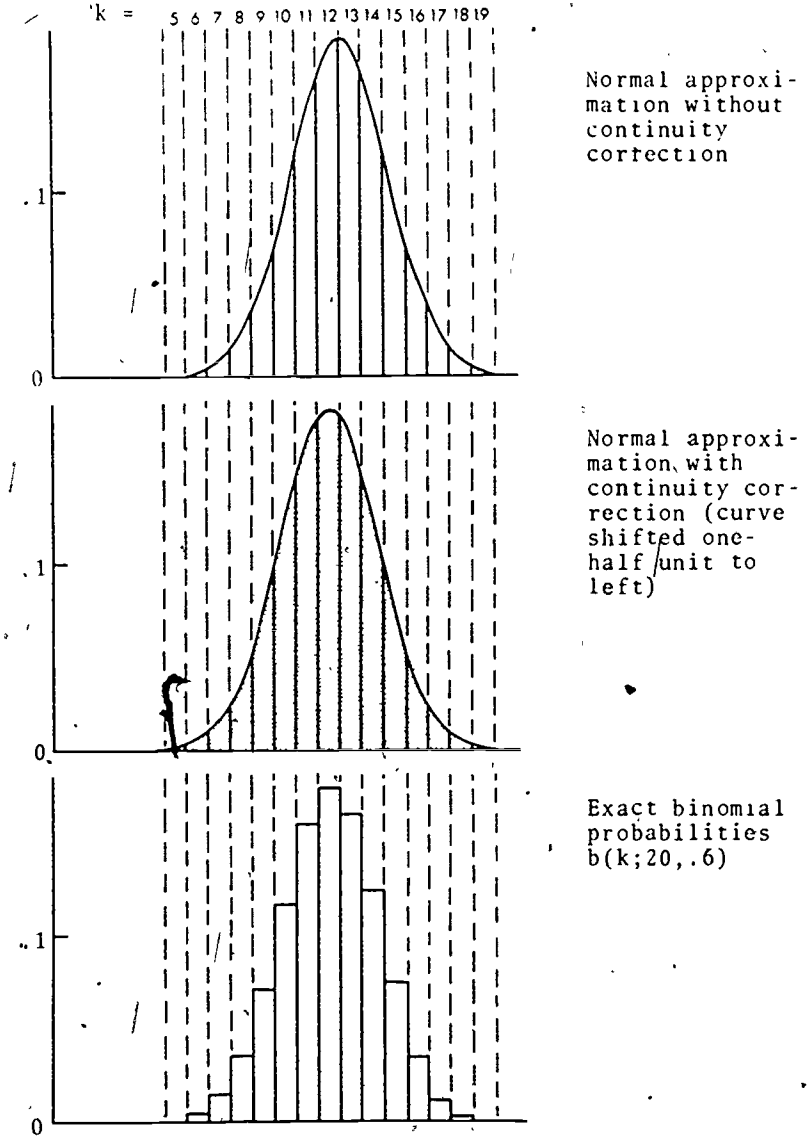


Figure 2. Normal approximations to binomial probabilities for n = 20, p = .6. (Area between lines under curve is probability assigned to k successes.)

3.4 Approximation of Binomial Probabilities by Poisson Probabilities

The second alternative approximation may be applicable when values of p are very small (near zero) or large (near one). We need consider only small values of p; if p is large, we can interchange our definitions of "success" and "failure" and apply the discussion below. (We can make the exchange because "success" and "failure" are arbitrary designations, and it will suffice because a very large probability of "success" implies a very small probability of "failure".)

When n is fairly large, p is small, and np is moderate (perhaps somewhere between 0.5 and 5), the probability of k successes in n trials may be approximated by the Poisson probability distribution:

$$b(k;n,p) \approx p(k;np) = \frac{(np)^k e^{-np}}{k!} \tag{14}$$

The values of p(k;np) are easily computed with a calculator or by a computer.

In illustrating the Poisson approximation, we shall suppose that we want to obtain an approximation of the probability of no successes or one success in one hundred independent trials, each trial with probability of success .02. To apply the Poisson approximation, we find np = 100(.02) = 2 and compute the approximations of the probabilities of zero successes and one success, obtaining

$$\begin{aligned} B(1;100,.02) &= b(0;100,.02) + b(0;100,.02) \\ &\approx p(0;2) + p(1;2) \\ &= \frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} \\ &= .406 \end{aligned}$$

The exact probability, computed from the binomial distribution, is .403; the uncorrected normal approximation is .238, and the corrected normal approximation is .361. In this example, the Poisson approximation is considerably more accurate than either of the normal approximations.

The basis for the continuity correction is essentially mathematical--it exploits the particular way in which binomial probabilities begin to resemble normal probabilities as n becomes large. Although the Poisson approximation may be derived mathematically, we can see it as manifested more intuitively in structure. If we imagine that we are holding constant the number of successes likely to be observed but allowing the number of trials to increase, then the experiment begins to resemble a process in which successes occur "at random" across time. Such a process gives rise directly to a Poisson distribution. In this sense, the Poisson approximation is structural, although its derivation is frequently represented mathematically. The analogy between the Poisson approximation and the Poisson process of stochastic-process theory is discussed in most elementary probability texts.

3.5 Exercises

To do the following exercises, use the recursive formula (4) for computing binomial probabilities and the corresponding formula

$$p(k+1;np) = \frac{np}{k+1} p(k;np) \tag{15}$$

for computing Poisson probabilities.

Exercise 6. Tabulate the cumulative binomial distribution function and its normal and Poisson approximations for n = 5, 20 and 50 for each value of p = .5, .25, and .1 . For which values of n and p does each approximation appear to be valid? Which method of approximation gives better results in the "tails" of the distribution when p is small? Compare the results of using differences between successive values of k in the normal approximation to the cumulative binomial distribution with the results of using the direct approximation of b(k;n,p) described by equation (7).

Exercise 7. Recompute the normal approximations of Exercise 6 using the continuity correction, and describe its effect on the accuracy of the approximations.

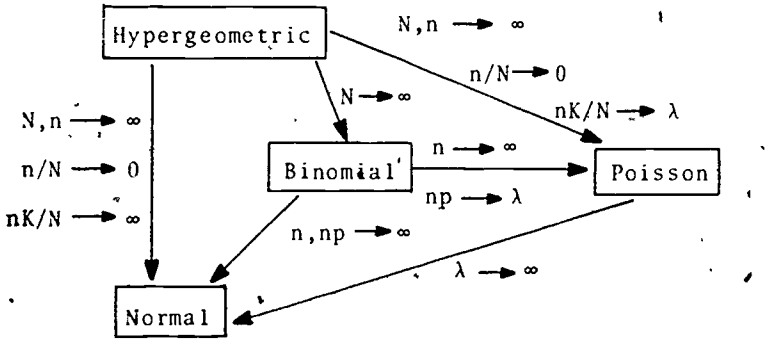
Exercise 8. A jury panel of 100 members was selected from a community in which 25% of the jury-eligible residents own no land. 90 of the panel members were land owners. How likely is it that non-land-owners are that scarce on a panel when selection is truly random?

Exercise 9. Suppose that in the community of Exercise 8, 4% of the jury-eligible residents have completed fewer than 8 years of school. What is the probability that every member of a randomly selected jury panel has completed 8 or more years of school?

4/ CONCLUSION

4.1 Summary

The following diagram summarizes the approximations we have discussed.



That is, the hypergeometric probabilities are approximated by the binomial probabilities for large populations. The binomial probabilities in turn have normal and Poisson approximations; so, therefore, do the hypergeometric probabilities. The diagram shows that Poisson probabilities have a normal approximation for large values of the parameter λ, but we have not discussed that approximation here.

In all of the populations we discussed, the numerical values are either zeros or ones, representing dichotomous outcomes--success or failure, yellow or pink, etc. There are approximation techniques for other kinds of populations. Many such techniques are in common use in statistics, especially techniques based in some way on Central Limit Theorems. Approximate statistical methods, based on approximate probability calculations, are widely used by statisticians. Discussion of the theoretical bases for approximate statistical methods is beyond the scope of this module; however, the techniques have the same two bases--structural and mathematical approximations.

From these and similar approximations, you should be gaining the feeling that it is possible for several probability models whose similarity is not immediately apparent to reflect a given sampling problem. As you progress in your study of inferential statistical methods, it will become more and more necessary for you to rely on the ideas of approximation in choosing models for populations and in deriving approximate sampling distributions for the statistics you will be using in reaching conclusions about the populations. The approximations here of hypergeometric and binomial distributions are useful as presented, for determining the probabilities of given numbers of successes, but examining them should in addition give you some familiarity with the advantages and limitations of approximation in general.

4.2 Exercises

Exercise 10. How might one obtain a normal approximation to hypergeometric probabilities? For what values of N, n, and K would it be valid?

Exercise 11. A committee of 25 people is to be drawn at random from a group consisting of 120 men and 80 women. Obtain an approximation of the probability that more than half of the committee members will be men.

Exercise 12. Wilfred Shadbolt is inspecting brackets. He tests

30 of them, choosing the 30 randomly (without replacement) from a lot of 5000. If the 5000 include 150 defective brackets, what is the probability that at least one defective bracket will be among the 30 tested?

Exercise 13. Show that

a) as N becomes very large (while K/N = p remains constant), the coefficient of h(k;N,n,k) in formula (3) approaches the coefficient of b(k;n,p) in formula (4).

b) as n becomes very large and p becomes very small (while np remains constant), the coefficient of b(k,n,p) in formula (4) approaches the coefficient of p(k,np) in formula (15).

(Rigorous demonstration of these propositions, each of which corresponds to a segment of the diagram of Section 4.1, requires some calculus.)

77

5. ANSWERS TO EXERCISES

Exercise 1.(a)

Area of triangle = 1/2 × 1 × 1 = 1/2.

Exercise 1.(b)

Slope of tangent line = 1. To find tangent point, set $\frac{df}{dx} = \frac{1}{2\sqrt{x}} = 1$, and solve to obtain x = 1/4. Line intersects vertical axis at f(1/4) -1/4 = 1/4, height of line at x = 1 is f(1/4) + 3/4 = 5/4. Area of trapezoid is 1 × (1/4 + 5/4)/2 = 3/4.

Exercise 1.(c)

f(1/2) = .7071; f(1) = 1. Area of first rectangle = .3536; area of second rectangle = .5. Approximate area = .8536.

Exercise 1.(d)

f(1/4) = .5; f(3/4) = .8660. Area of first rectangle = .25; area of second rectangle = .4330. Approximate area = .6830.

Exercise 1.(e)

x	f(x)	Area of rectangle
.125	.3536	.0442
.250	.5000	.0625
.375	.6124	.0765
.500	.7071	.0884
.625	.7906	.0988
.750	.8660	.1083
.875	.9354	.1169
1.000	1.0000	.1250
		.7206

Approximate Area = .7206

Exercise 1.(f)

x	f(x)	Area of rectangle
.0625	.2500	.0313
.1875	.4330	.0541
.3125	.5590	.0699
.4375	.6614	.0827
.5625	.7500	.0938
.6875	.8292	.1036
.8125	.9014	.1127
.9375	.9682	.1210
		.6691

Approximate Area = .6691, which is very close to 2/3.

78

Exercise 1.(g)

The answers to (c) and (d) are farther apart than the answers to (e) and (f). Taking the height of a rectangle to be f(x) at the center of the rectangle rather than at the edge is more critical to the success of the approximation when fewer, broader rectangles are used.

Exercise 2.(a)

k	f(k)		$\frac{5-k}{k+1}$	$\binom{5}{k} = \frac{5!}{k!\,(5-k)!}$
0	1	×	5	$\frac{5\times4\times3\times2\times1}{1\times5\times4\times3\times2\times1} = 1$
1	5	×	2	$\frac{5\times4\times3\times2\times1}{1\times4\times3\times2\times1} = 5$
2	10	×	1	$\frac{5\times4\times3\times2\times1}{2\times1\times3\times2\times1} = 10$
3	10	×	2/4	$\frac{5\times4\times3\times2\times1}{3\times2\times1\times2\times1} = 10$
4	5	×	1/5	$\frac{5\times4\times3\times2\times1}{4\times3\times2\times1\times1} = 5$
5	1		0	$\frac{5\times4\times3\times2\times1}{5\times4\times3\times2\times1\times1} = 1$

Exercise 2.(b)

Suppose that $f(k) = \binom{5}{k}$. Then $f(0) = \binom{5}{0} = \frac{5!}{0!\,5!} = 1$, and $f(k+1) = \binom{5}{k+1}$. Therefore

$$\frac{f(k+1)}{f(k)} = \frac{\binom{5}{k+1}}{\binom{5}{k}} = \frac{\frac{5!}{(k+1)!\,(4-k)!}}{\frac{5!}{k!\,(5-k)!}}$$

$$= \frac{k!\,(5-k)!}{(k+1)!\,(4-k)!}$$

$$= \frac{k!\,(5-k)\cdot(4-k)!}{(k+1)\cdot k!\,(4-k)!}$$

$$= \frac{5-k}{k+1}.$$

So $f(k+1) = \frac{5-k}{k+1}\,f(k)$, which is the recursive formula sought.

Exercise 3.(a)

k	Exact Hypergeometric	Binomial Approximation
0	0.0040	0.0313
1	0.0992	0.1563
2	0.3968	0.3125
3	0.3968	0.3125
4	0.0992	0.1563
5	0.0040	0.0313

Exercise 3.(b)

0	0.5000	0.5905
1	0.5000	0.3280

Exercise 3.(c)

0	0.0281	0.0313
1	0.1529	0.1563
2	0.3189	0.3125
3	0.3189	0.3125
4	0.1529	0.1563
5	0.0281	0.0313

Exercise 3.(d)

0	0.5838	0.5905
1	0.3394	0.3280
2	0.0702	0.0729
3	0.0064	0.0081
4	0.0003	0.0004

Exercise 4.(c)

k	Exact Hypergeometric	Binomial Approximation
3	0.0004	0.0011
4	0.0021	0.0046
5	0.0089	0.0148
6	0.0278	0.0370
7	0.0661	0.0739
8	0.1216	0.1201
9	0.1746	0.1602
10	0.1969	0.1762
11	0.1746	0.1602
12	0.1216	0.1201
13	0.0661	0.0739
14	0.0278	0.0370
15	0.0089	0.0148
16	0.0021	0.0046
17	0.0004	0.0011

Exercise 4.(d)

k	Exact Hypergeometric	Binomial Approximation
0	0.0951	0.1216
1	0.2679	0.2702
2	0.3182	0.2852
3	0.2092	0.1901
4	0.0841	0.0898
5	0.0215	0.0319
6	0.0035	0.0089
7	0.0004	0.0020

Exercise 5.

Rose is sampling randomly without replacement from a finite population of dichotomous outcomes. N is the total number of members of the U.S. House of Representatives, which is 435. K is the number of Representatives supporting the bill in which Rose is interested. n is the number of Representatives in Rose's sample, which is 6. k will be the number of Representatives in Rose's sample who support the bill. The binomial approximation would be adequate, because 6 is a small fraction of 435.

Exercises 6 & 7. n = 20 p = .25 [np(1-p) = 3.75]

CUMULATIVE PROBABILITIES

k	y	Exact Binomial	Uncorrected Normal Approx.	Corrected y	Corrected Normal Approx
0	-2.5820	0.0032	0.0049	-2.3238	0.0101
1	-2.0656	0.0243	0.0194	-1.8074	0.0354
2	-1.5492	0.0913	0.0607	-1.2910	0.0984
3	-1.0328	0.2252	0.1508	-0.7746	0.2193
4	-0.5164	0.4148	0.3028	-0.2582	0.3981
5	0.0000	0.5172	0.5000	0.2582	0.6019
6	0.5164	0.7858	0.6972	0.7746	0.7807
7	1.0328	0.8982	0.8492	1.2910	0.9016
8	1.5492	0.9591	0.9393	1.8074	0.9646
9	2.0656	0.9861	0.9806	2.3238	0.9899
10	2.5820	0.9961	0.9951	2.8402	0.9977
11	3.0984	0.9991	0.9990	3.3566	0.9996
12	3.6148	0.9998	0.9998	3.8730	0.9999
13	4.1312	1.0000	1.0000	4.3894	1.0000

INDIVIDUAL PROBABILITIES

k	Exact Binomial	Normal Approx	Poisson Approx
0	0.0032	0.0073	0.0067
1	0.0211	0.0244	0.0337
2	0.0669	0.0620	0.0842
3	0.1339	0.1208	0.1404
4	0.1897	0.1803	0.1755
5	0.2023	0.2060	0.1755
6	0.1686	0.1803	0.1462
7	0.1124	0.1208	0.1044
8	0.0609	0.0620	0.0653
9	0.0271	0.0244	0.0363
10	0.0099	0.0073	0.0181
11	0.0030	0.0017	0.0082
12	0.0008	0.0003	0.0034
13	0.0002	0.0000	0.0013

n = 50 p = 0.1 [np(1-p) = 4.5]

CUMULATIVE PROBABILITIES

k	y	Exact Binomial	Uncorrected Normal Approx.	Corrected y	Corrected Normal Approx.
0	-2.3570	0.0052	0.0092	-2.1213	0.0169
1	-1.8856	0.0338	0.0297	-1.6499	0.0495
2	-1.4142	0.1117	0.0786	-1.1785	0.1193
3	-0.9428	0.2503	0.1729	-0.7071	0.2398
4	-0.4714	0.4312	0.3187	-0.2357	0.4068
5	0.0000	0.6161	0.5000	0.2357	0.5932
6	0.4714	0.7702	0.6813	0.7071	0.7602
7	0.9428	0.8779	0.8271	1.1785	0.8807
8	1.4142	0.9421	0.9214	1.6499	0.9505
9	1.8856	0.9755	0.9703	2.1213	0.9831
10	2.3570	0.9906	0.9908	2.5927	0.9952
11	2.8284	0.9968	0.9977	3.0641	0.9989
12	3.2998	0.9990	0.9995	3.5355	0.9998
13	3.7712	0.9997	0.9999	4.0069	1.0000
14	4.2426	0.9999	1.0000	4.4783	1.0000

INDIVIDUAL PROBABILITIES

Exact Binomial	Normal Approx.	Poisson Approx.
0.0052	0.0117	0.0067
0.0286	0.0318	0.0337
0.0779	0.0692	0.0842
0.1386	0.1206	0.1404
0.1809	0.1683	0.1755
0.1849	0.1880	0.1755
0.1541	0.1683	0.1462
0.1076	0.1206	0.1044
0.0643	0.0692	0.0653
0.0333	0.0318	0.0363
0.0152	0.0117	0.0181
0.0061	0.0034	0.0082
0.0022	0.0008	0.0034
0.0007	0.0002	0.0013
0.0002	0.0000	0.0005

83

Exercise 8.

The hypergeometric probabilities can be approximated by

$$B(10;100,.25) \approx \Phi\left(\frac{10-100(.25)+.5}{\sqrt{100(.25)(.75)}}\right) = \Phi(-3.35) = .0004$$

Exercise 9.

The hypergeometric probability can be approximated by

$$b(0;100,.04) \approx p(0;4) = \frac{4^0 e^{-4}}{0!} = .018$$

Exercise 10.

Approximate the hypergeometric probabilities with binomial probabilities, and approximate the binomial probabilities with one of the normal approximations. N should be very large, K should be an appreciable fraction of N, and n should be large (but still a small fraction of N).

Exercise 11.

$$\begin{aligned} P(\text{number of men} \geq 13) &= P(\text{number of women} < 12) \\ &= H(12;200,25,80) \\ &\approx B(12;25,.4) \\ &\approx \Phi\left(\frac{12-25(.4)+.5}{\sqrt{25(.4)(.6)}}\right) \\ &= \Phi(-1.02) \\ &= .154 \end{aligned}$$

Exercise 12.

$$\begin{aligned} P(\text{at least one bracket defective}) &= 1 - P(\text{no brackets defective}) \\ &= 1 - h(0;5000,30,150) \\ &\approx 1 - b(0;30,.03) \\ &\approx 1 - p(0;.9) \\ &= 1 - \frac{.9^0 e^{-.9}}{0!} \\ &= 1 - .407 = .593 \end{aligned}$$

27

81

28

Exercise 13.(a)

$$\frac{(K-k)(n-k)}{(k+1)(N-K-n+k+1)} = \frac{\frac{(K-k)}{N}(n-k)}{(k+1)(\frac{N-K-n+k+1}{N})} = \frac{(\frac{K}{N} - \frac{k}{N})(n-k)}{(k+1)(1 - \frac{K}{N} - \frac{n+k+1}{N})}$$

As N becomes very large, $\frac{k}{N}$ and $\frac{n+k+1}{N}$ become so small as to be negligible, so the expression above is approximately

$$\frac{\frac{K}{N}(n-k)}{(k+1)(1-\frac{K}{N})}$$

Because $\frac{K}{N} = p$, we can write that as $\frac{p(n-k)}{(k+1)(1-p)}$, which is the coefficient of b(k;n,p) in formula (4).

Exercise 13.(b)

$$\frac{(n-k)p}{(k+1)(1-p)} = \frac{np - kp}{(k+1) - (k+1)p}$$

As p becomes very small (but np remains constant), kp and (k+1)p become so small as to be negligible, so the expression above approaches $\frac{np}{k+1}$, which is the coefficient of p(k;np) in formula (15).

6. MODEL UNIT EXAM

1. In what sense is the binomial approximation to hypergeometric probabilities structural? In what sense is the normal approximation to binomial probabilities structural?

2. You are working for an automobile dealer. Invent a hypergeometric random variable related to your work, and describe what N, K, n, and k are. Can you approximate its distribution adequately with a binomial distribution? How would you change your answer to the first question to make the random variable genuinely binomial? What would p be?

3. Thomas Tolloller plays a gambling game at which he has probability p = .492 of winning $1 and probability p = .508 of losing $1. What is the probability that, after 100 plays, he has won more than he has lost? What is the probability that, after 100 plays, he has won <u>exactly</u> as many times as he has lost?

4. Thomas Tolloller plays another game at which he is told he has a 1/38 chance of winning on each play. After 100 plays, he has won only once. How likely is winning no more than once in 100 plays if the game is as described?

7. ANSWERS TO MODEL UNIT EXAM

1. The sampling schemes in binomial and hypergeometric situations are similar. The binomial and normal distributions are both sampling distributions of sums, and they can be shown mathematically to be similar for large sample sizes.

2. For example, Y could be the number of people in a random sample of 15 of this year's customers who bought Model PQR. (the random sample is chosen without replacement). N would be the total number of this year's customers; K would be the number of this year's customers who bought Model PQR; n would be 15, the number of customers in the sample; and k would be the number of customers in the sample who bought Model PQR. If the dealership is active this year (selling more than 75 cars, say), then the binomial approximation should be adequate. To make Y genuinely binomial, the sample should be chosen with replacement (i.e., a customer could appear in the sample more than once.) $P = \frac{K}{N}$.

3. $$B(49;100,.508) \approx \phi\left(\frac{49-100(.508)+.5}{\sqrt{100(.508)(.492)}}\right) = \phi(1.260) = .397$$

and

$$b(50;100,.492) \approx \frac{1}{\sqrt{100(.492)(.508)}}\ \phi\left(\frac{50-100(.492)}{\sqrt{100(.492)(.508)}}\right)$$

$$= \frac{\phi(.160)}{4.999} = .079.$$

4. $$B(1;100,1/38) = b(0;100,1/38) + b(1;100,1/38)$$

$$= p(0;100/38) + p(1;100/38) = .261.$$

STUDENT FORM 1

Request for Help

Return to:
EDC/UMAP
55 Chapel St.
Newton, MA 02160

Student: If you have trouble with a specific part of this unit, please fill out this form and take it to your instructor for assistance. The information you give will help the author to revise the unit.

Your Name _____ Unit No. _____

Page _____
- ○ Upper
- ○ Middle
- ○ Lower

OR

Section _____
Paragraph _____

OR

Model Exam
Problem No. _____
Text
Problem No. _____

Description of Difficulty: (Please be specific)

Instructor: Please indicate your resolution of the difficulty in this box.

○ Corrected errors in materials. List corrections here:

○ Gave student better explanation, example, or procedure than in unit. Give brief outline of your addition here:

○ Assisted student in acquiring general learning and problem-solving skills (not using examples from this unit.)

Instructor's Signature _____

Please use reverse if necessary.

STUDENT FORM 2

Unit Questionnaire

Return to:
EDC/UMAP
55 Chapel St.
Newton, MA 02160

Name_____ Unit No._____ Date_____

Institution_____ Course No._____

Check the choice for each question that comes closest to your personal opinion.

1. How useful was the amount of detail in the unit?

____Not enough detail to understand the unit
____Unit would have been clearer with more detail
____Appropriate amount of detail
____Unit was occasionally too detailed, but this was not distracting
____Too much detail; I was often distracted

2. How helpful were the problem answers?

____Sample solutions were too brief; I could not do the intermediate steps
____Sufficient information was given to solve the problems
____Sample solutions were too detailed; I didn't need them

3. Except for fulfilling the prerequisites, how much did you use other sources (for example, instructor, friends, or other books) in order to understand the unit?

____A Lot ____Somewhat ____A Little ____Not at all

4. How long was this unit in comparison to the amount of time you generally spend on a lesson (lecture and homework assignment) in a typical math or science course?

Much Longer	Somewhat Longer	About the Same	Somewhat Shorter	Much Shorter
____	____	____	____	____

5. Were any of the following parts of the unit confusing or distracting? (Check as many as apply.)

____Prerequisites
____Statement of skills and concepts (objectives)
____Paragraph headings
____Examples
____Special Assistance Supplement (if present)
____Other, please explain_____

6. Were any of the following parts of the unit particularly helpful? (Check as many as apply.)

____Prerequisites
____Statement of skills and concepts (objectives)
____Examples
____Problems
____Paragraph headings
____Table of Contents
____Special Assistance Supplement (if present)
____Other, please explain_____

Please describe anything in the unit that you did not particularly like.

Please describe anything that you found particularly helpful. (Please use the back of this sheet if you need more space.)