DOCUMENT RESUME

ED 304 367 SO 019 695

TITLE	Introduction to Statistics. Learning Packages in the Policy Sciences Series, PS-26. Revised Edition.
INSTITUTION	Policy Studies Associates, Croton-on-Hudson, NY.
REPORT NO	ISBN-0-936826-24-X
PUB DATE	87
NOTE	65p.
PUB TYPE	Statistical Data (110) -- Guides - Classroom Use - Materials (For Learner) (051)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Curriculum Guides; Higher Education; Instructional Materials; *Public Policy; Resource Materials; *Statistics
IDENTIFIERS	*Policy Research

ABSTRACT

The primary objective of this booklet is to introduce students to basic statistical skills that are useful in the analysis of public policy data. A few, selected statistical methods are presented, and theory is not emphasized. Chapter 1 provides instruction for using tables, bar graphs, bar graphs with grouped data, trend lines, pie diagrams, percentages to describe differences, and the mean, median, and mode. Range, variance, standard deviation, z-scores, and coefficient of variance are described in chapter 2. Chapters 3 and 4 examine relationships between variables, while chapters 5 and 6 describe measures of association and include the: (1) del statistic; (2) t-test; (3) scatterplot; (4) linear regression analysis; (5) regression line; and (6) Pearson's R superscript 2. Selected formulas and exercises are provided and tables and figures are included. (JHP)

LEARNING PACKAGES IN THE POLICY SCIENCES

PS 26

INTRODUCTION TO STATISTICS

Revised Edition

POLICY STUDIES ASSOCIATES

INTRODUCTION

TO

STATISTICS

Third Revised Edition

of

Descriptive Statistics for Public Policy Analysis (PS-20)

LEARNING PACKAGES IN THE POLICY SCIENCES

PS-26

Policy
Studies
Associates

P. O. Box 337
Croton-on-Hudson, NY 10520

ISBN 0-936826-24-X

OBJECTIVES AND MATERIALS

Primary Objective:

To introduce you to basic statistical skills used in the analysis of public policy issues.

Upon Completion of This Package, You Will Be Able To:

- Arrange small data sets into tabular form.
- Organize large data sets into frequency distributions.
- Construct and interpret graphic displays of data.
- Calculate and distinguish among several basic statistics.
- Perform data analysis and interpretation using these statistics.

Additional Material Required:

None, although access to a pocket calculator would be desirable.

Time Span:

Five weeks.

TABLE OF CONTENTS

FORMULAS

PREFACE

Several people are to be thanked for contributing to the present version of this learning package. Most important are Amy Kafton and Gary Hammerstrom, the authors of the previous edition of this package, Descriptive Statistics for Public Policy Analysis. The use of that edition in classes and as a guide to self-study by individual students provided a wealth of information about what students did and did not need to know about basic applied statistics. The hundreds of students who used the earlier package and provided comments ranging from improved formulas to spotting typographical errors were also an immense help.

These contributions show how these packages can serve as a step in the direction of cooperation among teachers and students in the preparation of more effective learning materials for applied policy analysis.

INTRODUCTION

In order to understand and participate in public policy issues, you have to be familiar with the procedures of statistical analysis and be able to discriminate between and interpret the most common statistics. The purpose of this learning package is to equip you with the basic statistical skills needed for public policy analysis. It should be emphasized that this is an introduction to statistics and not the equivalent of a complete course. Only a few methods of analysis are presented and much of the theory behind them has been omitted. The number of formulas has been kept to a minimum and in no case are they proved or derived.

The major educational assumption behind this package is that students learn best by doing. The material has been broken into six chapters with an exercise following each. Depending on the purposes of the course, the exercises may be accomplished in some combination of the following:

- Students may complete the exercise with data gathered through a survey of class members;
- Students may complete the exercise with data that each has gathered as part of a research project;
- Students may complete the exercises with data provided by the instructor pertinent to a specific public policy or social science discipline: or
- Students may complete the exercises with completely hypothetical data.

Further information on statistics may be found in such texts as:

Blalock, H.M., Jr., Social Statistics, New York: McGraw-Hill, 1972.

Runyon, R.P., and A. Haber, Fundamentals of Behavioral Statistics, Reading, MA: Addison-Wesley, 1976.

CHAPTER I: DESCRIPTION

Why do people collect and analyze numerical data when they study public policy? For two primary reasons: (1) To describe social conditions that have importance to policy, and (2) to describe relationships that help us understand policy problems.

It is possible to find many examples of the first use of numerical data. No one today is content with a statement like: "Lots of people are unemployed." In this, and in countless other areas of daily life, we seek the greater precision of numbers in describing the world around us (e.g., "Unemployment is 7.6% this year, compared to 8.1% last year").

Many people who are philosophically opposed to the use of numbers (or are too lazy to learn how to use numbers) will be quick to point out that it is easy to lie with statistics. This is undoubtedly the truth: an unethical person can use numbers as well as words. The use of quantitative symbols is no assurance of either the competence or honesty of the speaker. The use of numbers does, however, assure that the speaker must be precise and clear about what is being communicated. At the very least, the use of numbers makes it possible to know exactly what the speaker is trying to say. In many cases, we can go back to some other sources--or even do our own research--to find out if the speaker is telling the truth.

A. Using Tables

Data must be presented in a form that is both interesting and clear. Stringing numbers together in a text does not accomplish either of these goals. For example, consider the following hypothetical example:

> "A recent study by the organization called Committee on the Aging: Research and Education (CARE) indicates that in 1920 the percentage of the United States population over 60 years of age was 6% of the population; in 1940 the percentage was 8%; in 1960 it was 13%; in 1980 it was 15%; and in 2000 it is forecast that it will be 20%."

Such information is difficult to understand or analyze when presented in this way. If arranged in tabular form (Table 1.1), it is much easier to interpret.

Table 1.1

PERCENTAGE OF THE U.S. POPULATION OVER 60 YEARS OF AGE

Year	Percentage
1920	6%
1940	8%
1960	13%
1980	15%
2000 (forecast)	20%

SOURCE: Committee on the Aging, Nursing Home Care in the U.S., Washington, D.C.: Superintendent of Documents, 1974.

From this table it is easy to see clear trends in the variable (a characteristic which may take on different values) over the time series (the pattern of a series of values arranged in a time sequence).

Tabular methods are a widely used and accepted means of organizing small sets of data for rapid visualization and understanding. A table requires:

- A title which clearly explains its content or purpose.
- Data elements carefully listed under headings which clearly specify units of measure.
- Documentation of the data source.

Tables can be used to present information that describes social conditions. They are most useful when numerical information is presented either for years, as in the table above, or for different locations (e.g., different cities, states, or countries). The use of tables in written or oral presentations must also include the meaning of the information in the table and what it says about social conditions as they relate to the public policy issue.

B. Using Bar Graphs

An even more striking way of presenting data is to use a bar graph, which is a series of parallel bars (or similar markings) placed either vertically or horizontally to indicate frequencies. Figure 1.1 is an example of the same information about the population of the United States as above, presented here in bar graph form.

Figure 1.1

BAR GRAPH SHOWING THE PERCENTAGE OF THE U.S. POPULATION
OVER 60 YEARS OF AGE

```
PERCENTAGE

20|                                                x
18|                                                x
16|                                    x           x
14|                        x           x           x
12|                        x           x           x
10|                        x           x           x
 8|             x          x           x           x
 6|   x         x          x           x           x
 4|   x         x          x           x           x
 2|   x         x          x           x           x
 0|   x         x          x           x           x
  ________________________________________________________________
     1920      1940       1960        1980        2000 (forecast)

                              YEAR
```

SOURCE: Committee on the Aging, Nursing Home Care in the U.S., Washington, D.C.: Superintendent of Documents, 1974.

In the construction of a bar graph, the length of the bars and the space between them should be consistent and allow for clear visual inspection. In the example above, each "X" represents approximately 2 percentage points. Each bar does not necessarily represent a number precisely, but it does give a general picture of the pattern. Bar graphs are useful for helping the reader see the differences and similarities between observational units such as years (as in the example above), sex (comparing males to females), cities, states, countries, types of businesses, groups of people, geographical locations (e.g., north versus south, or urban versus rural). Anytime you want to compare two or more units with respect to some variables, you may want to use a bar graph.

C. Using Bar Graphs with Grouped Data

When you are presenting information on frequencies for each year (as in previous tables) or the frequency of well-established categories, such as party affiliation or place of residence, deciding which category each bar represents is not a problem. But, when you have frequencies in continuous, numerical categories, the job is more difficult. Consider the following hypothetical data about the age distribution in the United States population of those who are age 60 or older.

Table 1.2

DISTRIBUTION OF AGES OF THOSE WHO ARE AT LEAST 60 YEARS OF AGE

Age	Percentage
60	5.94
61	6.94
62	4.94
63	4.94
64	6.88
65	4.36
66	4.00
67	4.72
68	5.36
69	3.36
70	4.00
71	4.72
72	5.36
73	3.36
74	4.36
75	2.22
76	2.22
77	2.00
78	3.11
79	3.11
80	1.11
81	1.11
82	2.00
83	2.11
84	2.11
85+	5.6

SOURCE: Committee on the Aging, *Nursing Home Care in the U.S.*, Washington, D.C.: Superintendent of Documents, 1974.

It clearly does not make sense to present a bar for each age category, since each category represents such a small percentage. What should be done in this case is to group the information into a smaller number of categories (between 3 and 7 is usually a good number) and present the bar graphs showing the distribution by category.

In this case, we will divide the information into four categories: 61-64; 65-74; 75-84; and 85+. The categorization could be based on having an equal range for each category, or, as in this case, on the basis of some dividing points that are important for policy reasons. (The age groups here divide people into groups who are eligible for different forms of benefits.) After the classification is complete, the bar graph can be presented in the same way as above, in Figure 1.2. (In this bar, each "X" represents about 2 percentage points.)

Figure 1.2

BAR GRAPH OF DISTRIBUTION OF THE AGES
OF THOSE WHO ARE AT LEAST 60 YEARS OF AGE

```
PERCENTAGE

45|            x
40|            x
35|            x
30|     x      x
25|     x      x
20|     x      x      x
15|     x      x      x
10|     x      x      x
 5|     x      x      x      x
 0|_____x______x______x______x___

      60-64  65-74  75-84   85+

             AGE CATEGORY
```

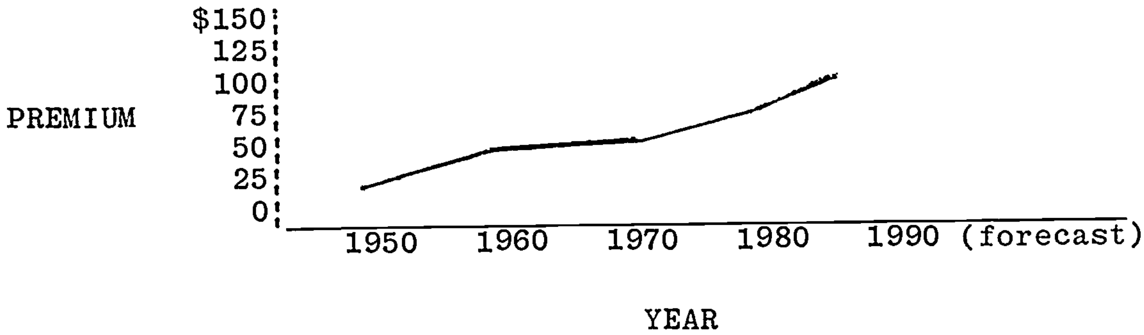
SOURCE: Committee on the Aging, Nursing Home Care in the U.S., Washington, D.C.: Superintendent of Documents, 1974.

D. Using Trend Lines

A trend line is derived by plotting time--in years, months, or days--on the horizontal axis, and something which is changing over time on the vertical axis. This type of graph shows growth or decline of some condition over time. The trend can also be projected into the future. This type of display is useful in monitoring and forecasting social conditons. Figure 1.3 is an example of a trend line and forecast using hypothetical data.

Figure 1.3

HEALTH INSURANCE PREMIUMS IN NEW YORK STATE:
TREND OF HEALTH PREMIUM COSTS



SOURCE: "Trends in Insurance Premiums," Albany: New York State Insurance Council, 1982.

E. Using Pie Diagrams

A pie diagram can be used to show how the component parts of a sum are divided. The apportionment of government spending or the ethnic composition of a political party are both examples of subjects that can be illustrated with this technique. A pie diagram is not difficult to construct if you remember that the total of 100% is described by a circle of 360 degrees. Thus, each percentage point is equal to an arc of 3.6 degrees. To illustrate, we will construct a pie diagram to show the distribution of people over 60 years of age by age group. This is the same information used in Figure 1.3, presented in a different way.

Table 1.3

AGE DISTRIBUTION OF POPULATION OVER 60

Age Group	Percent
60-64 years	29.7
65-74 years	43.6
75-84 years	21.1
85 years & over	5.6
	100. %

If each percent figure in Table 1.3 is multiplied by 3.6, we will know the size of the arc that must be drawn for each segment of the pie.

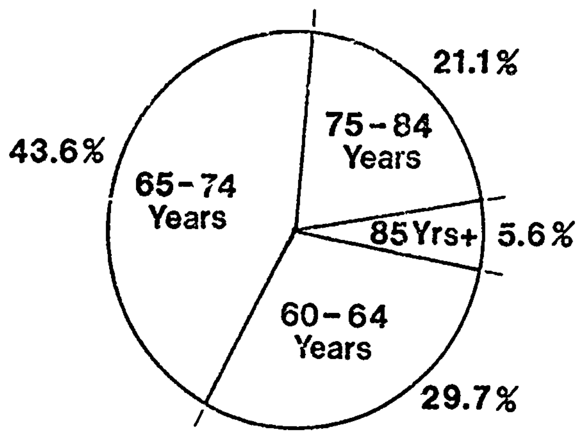
29.7	x	3.6	=	106.92
43.6	x	3.6	=	156.96
21.1	x	3.6	=	75.96
5.6	x	3.6	=	20.16

A protractor should be used to measure the necessary angles on the circle. It is a good idea to plan ahead so that you have your narrowest angles at the sides where they will be the easiest to label. The labels can either go inside or outside the circle, and should be on the outside if the angles are especially narrow.

As illustrated in Figure 1.4, the pie diagram can be a very effective technique for the visual display of data. However, some caution should be observed. Pie diagrams containing more than eight segments, or containing several segments with very small arc (less than five degrees), are difficult to label and often appear so cluttered as to be difficult to interpret. A pie diagram (or any graph) which is difficult to interpret or confusing is of no assistance in analyzing data or in presenting the results of analysis.

Figure 1.4

DISTRIBUTION OF U.S. POPULATION AGE 60+



SOURCE: Committee on the Aging, Nursing Home Care in the U.S., Washington, D.C.: Superintendent of Documents, 1974.

F. Using Percentages to Describe Differences

Percentages can be a powerful tool in assessing differences between two sets of numbers. You may want to determine the differences between estimated and actual budget figures, or between one year's crime rate and another. Calculating the percentage difference can give you a precise indicator of the difference between two points of observation. For example, the original estimate of the Federal budget deficit for 1983 was $113.65 billion, but the actual deficit was $195.4 billion. The percentage of error, or difference between the estimate and the actual, was 73%. The number of felonies in New York City was 637,451 in 1981, while the number dropped to 538,051 in 1984. The difference between the two years was a 16% drop.

To find the percentage difference, make the following calculation:

$$\frac{\text{New - Old}}{\text{Old}} \times 100 = \text{\% Difference}$$

For the example of crime in New York:

$$\frac{538{,}051 - 637{,}451}{637{,}451} = -.16 \text{ or } -16\%$$

G. Using the Mean, Median, and Mode

The arithmetic mean--or the "average," as it is commonly called--can be useful in describing social conditions and public policies. It can be used to summarize conditions over a number of time periods or across a number of points of observation. For example, the mean budget deficit for a five-year period from 1980 to 1984 was $119.74 billion as Table 1.4 indicates.

Table 1.4

U.S. FEDERAL GOVERNMENT BUDGET DEFICIT
BETWEEN 1980 and 1984

Year		Deficit in Billions
1980		59.5
1981		57.9
1982		110.6
1983		194.5
1984		175.3
	SUM	598.7

$$\text{MEAN} = \frac{598.7}{5} = 119.74$$

In addition to providing a summary of a series of observations, the mean can be useful in making judgments about improvements or lack of improvement in social conditions. For example, a budget deficit for one year that was below the mean of 119.74 would be viewed as an improvement in narrowing the deficit. One important point about using the mean is the range of numbers that contribute to the mean. In the example above, the range is very large with 57.9 in 1981 and 194.5 in 1983. Taking into account the range, you can see that the mean is a very rough way of generalizing about the five-year period. Several other statistics can be used to augment the interpretation of the mean, including the standard deviation. We will consider this measure later in the package.

Merely reporting the mean of a single set of observations is not very useful. It is much better to compare the means of two different sets of observations. For example, the $119.74 billion average deficit for the years 1980 to 1984 is hard to evaluate without some comparison. If we point out that the mean deficit between 1975 and 1979 was $46.6 billion, we immediately see how much the deficit on average has increased. The average between 1980 and 1984 is more than twice as high as the average between 1975 and 1979.

Another measure used to summarize is the median. The <u>median</u> is the middle number of a set of numbers. For example, we might have the following test scores, arranged in ascending order:

68 74 85 90 95

In the case of an odd number of values the middle, (85 in the above example), is the median. In the case of an even number, the procedure changes slightly. Suppose we had six numbers:

68 74 85 88 90 95

Again, we arrange the set in ascending order. Now, the median is the mean of the middle two numbers:

$$85 + 88 / 2 = 86.5$$

The median is especially useful as a summary when some of the numbers have extremely high or low values. Consider the numbers above. The median is 86.5 and the mean 83.3, not much different. But suppose the largest number were 195, rather than 95. This would make the mean equal to 100, but the median would still be 86.5, a much better representation of all the numbers. The median is commonly used as a summarizing measure for such unevenly distributed numbers as income.

A third summary measure of the data set, useful in certain cases, is the mode. The mode is the element which occurs most frequently. Take the following data set:

45 48 50 42 50 68 87 94

The number 50 occurs twice, while no other number occurs more than once. Therefore, 50 is the mode of this data set. But what would the mode have been if there had been two 94's as well as two 50's? In this case there are two modes, 50 and 94. This data set, in other words, is bi-modal.

For example, suppose a randomly selected group of high school students were surveyed. Among the information gathered is the students' year in school, with the following results:

9 12 9 11 10 9 9

The mode of this set is 9. This is useful to know because it tells you that there are more respondents to the survey in 9th grade than in any other grade.

Exercise I

A) Construct a table of some data relevant to a chosen public policy issue.

B) Construct a bar graph of data relevant to a public policy issue.

C) Prepare a trend line graph measuring some condition relevant to public policy.

D) Construct a pie chart that shows the relative size of the components of some social condition or policy.

E) Provide an example of percentage difference with two numbers you have located.

F) Calculate the mean, median, and mode from at least five numbers for a social condition or public policy.

G) For each of the above exercises, provide a one-paragraph interpretation of the figure of calculations.

NOTE: In this package, "interpretation" involves the following steps:

o Briefly summarize the main point of the figure or statistic.

- Discuss some of the principal reasons for this finding.
- Draw one inference of a forecast, evaluation, or prescription for public policy from the findings.
- Offer any reasons why the results (whether they are favorable or not) may not be valid. Indicate what future research is suggested by the results.

CHAPTER II: MEASURES OF DISPERSION, NORMAL DISTRIBUTION, AND STANDARD DEVIATION

We have learned how to summarize a set of numbers with the mean, median, and the mode. All three summarize an entire distribution of scores by describing the most typical or central value of the distribution. These statistics are powerful because they can reduce huge arrays of data to a single easily understood number. Also, remember that this kind of data reduction or summarizing function is the central purpose of descriptive statistics. But by themselves the measures of central tendency are incomplete summarizers of data. To fully describe a distribution of scores, measures of central tendency must be paired with measures of dispersion. Whereas the mean, median, and mode are designed to locate the central score, measures of dispersion provide an indication of the amount of heterogeneity or variety within the distribution of scores. The fluctuation of scores around the measure of central tendency is known as the variability or dispersion.

A. The Range(R)

The range(R) is the simplest measure of dispersion. It is calculated by subtracting the smallest observation in the data set from the largest and adding one. The range tells us how far the data are dispersed, but can be deceiving because it is based on only two scores in the distribution, the highest and the lowest. Since almost any sizable distribution will contain some atypically high and low scores, the range might be quite misleading. Also, R yields no information about the nature of the scores between the two extremes.

B. Variance and Standard Deviation

The most commonly used measure of distribution is the variance and the standard deviation.

The variance is defined as the average of the squared deviation from the mean. It is calculated as follows:

1) Find the difference between each individual's score and the mean.

2) Square this difference, then find the sum of these squares.

3) Divide their sum by the number of individuals.

Note that the variance is not in the same units as the original observations since they have been squared. The standard deviation is the square root of the variance. It is a number interpreted in the same units as the variable being measured.

FORMULAS

Variance = $\frac{\sum(X-\overline{X})^2}{N}$ Standard Deviation = $\sqrt{\frac{\sum(X-\overline{X})^2}{N}}$

X = Case N = Number of Cases

$\overline{X}$ = Mean of Cases

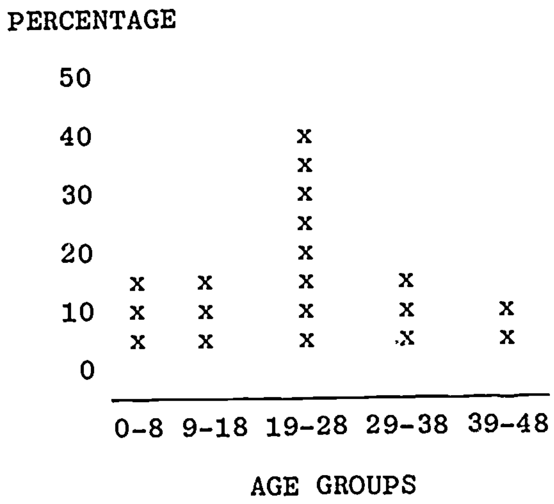
Like the mean, the standard deviation has mathematical properties which make it useful for higher statistical operations and it is therefore often the most desirable measure of dispersion. The size of the standard deviation represents the extent of the variability of individual observations around the mean.

Suppose that you are a social worker, trained to perform rural community development work and employed by a benevolent foundation through a series of renewable grants. You will be expected to mobilize the people of a community to tackle projects which will generally improve their living conditions such as building schools, digging wells for irrigation, and improving roads. You will work in a poor, rural area of the United States and want to plan your projects in detail before you leave home. You have narrowed your choices down to two possible locations: either a Native American reservation village in South Dakota or a small, backwoods town in Appalachia. The foundation research office has told you that the average age of the people in both villages is 23 years. Which location and what kind of project would you choose? In terms of central tendency, they seem identical. But are there important differences between the two?

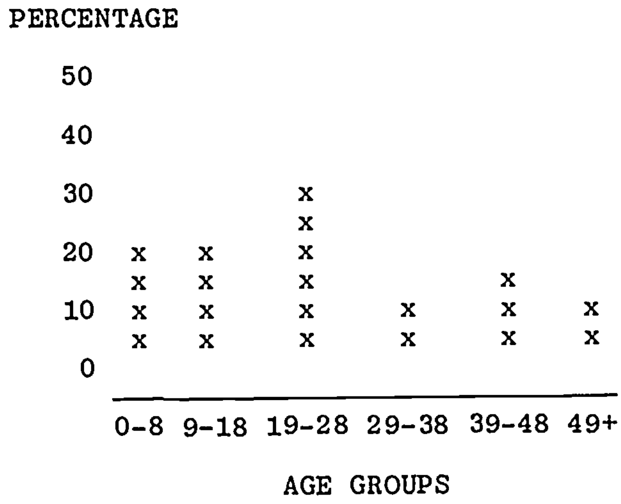
To check further, you ask the research office to provide you with the actual ages of all the people in the two villages. Using these data, you can construct a measure of dispersion of ages in the two villages, called "standard deviation" (S.D.). The S.D. is a kind of average distance of all individuals from the mean of the group.

To help visualize the two distributions, we can display the distribution of their ages on bar graphs:

Age Distribution of Town in Appalachia



Age Distribution of Town in South Dakota Indian Village



SOURCE: Funding Agency (hypothetical information)

The graphs show that the population of the South Dakota village is more widely dispersed. There are fewer people between 19 and 38 years of age, but more children and more older people. Thus, heavy construction projects take much longer in the South Dakota village since that location has about half as many people in the most active worker age bracket.

The larger number of young children in the South Dakota village implies that a program focusing on nutrition, personal hygiene, and health care would be appropriate here. The bar graph is a way of illustrating the different distributions of the two village populations. In addition, statistics provide some ways to summarize these distributions.

For the Appalachia village, the S.D. was 6.7 years. For the South Dakota village, the S.D. was 12.8 years. This is a clear statistical confirmation about the wider variability of ages in the latter village.

The mean and standard deviation can also be used to interpret individual scores. In most numerical distributions (which are characterized by the terms "normal distribution" or "bell-shaped curve"), about 67% of all individual scores will fall within the range of +/- 1 standard deviation on either side of the mean. About 95% of all scores will fall within +/- 2 standard deviations on either side of the mean. For example, suppose the average test score in a class was 75, and the standard deviation was 5.0. This means that about 67% of the individual scores are in the range 70-80 (+/- 1 S.D.), and about 90% fall within the range 65-85 (+/- 2 S.D.). This knowledge helps us interpret an individual score of 80 as being fairly close to the mean, while a score of 88 is substantially (more than 2 S.D.s) greater than the mean.

C. Z-Scores

Thus, the standard deviation and the normal curve are useful for the interpretation of an individual observation within a group of observations. They are also useful for the comparison of one person's responses on different occasions. Suppose an individual scored 95 on a history exam and 90 on a political science exam. The class mean on both tests was 75, and the standard deviations were 20 and 10 respectively.

History Test		Political Science Test	
Score	= 95	Score	= 90
Mean	= 75	Mean	= 75
S.D.	= 20	S.D.	= 10

In order to make a meaningful comparison of the scores obtained from the separate distributions, it is necessary to express each score in terms of standard deviation units from the mean. We can do this by transforming raw scores to z scores by dividing the difference between a given score and its mean by the standard deviation of that distribution.

FORMULA

$$\frac{\text{Score} - \text{Mean}}{\text{S.D.}} = z$$

For the history exam: $z = \frac{95 - 75}{20} = +1.0$

For the political science exam: $z = \frac{90 - 75}{10} = +1.5$

Although the student scored lower in raw units on the second test than on the first, he had a higher z score on the second exam due to the smaller standard deviation in the distribution of the results of the second test. Thus, it could be said that he did better on the second test.

D. Coefficient of Variation

The coefficient of variation provides us with a measure of variability relative to the mean and is especially useful for comparing the variability of distributions having different means:

FORMULA

$$\text{Coefficient of Variation (V)} = \frac{\text{S.D.}}{\text{Mean}}$$

Suppose we are given the following information about the standard deviation of per capita income in two U.S. counties:

County "A," Georgia:	S.D. = $10
County "B," New Jersey:	S.D. = $15

It would appear from this information that income is more widely dispersed in the New Jersey county than in the Georgia county. However, if we were then given the following information about mean income in these two counties it would change matters considerably.

County "A," Georgia:	Mean = $1,000
County "B," New Jersey:	Mean = $5,000

It seems probable to say that a S.D. of $10 where average income is only $1,000 shows much wider dispersion than a S.D. of $15 where average income is $5,000. It is obvious that in this case the standard deviation makes more sense in relationship to the mean of the data, and that we should calculate the coefficient of variation:

County "A," Georgia:	V = S.D./Mean = 10/1000 = .01
County "B," New Jersey:	V = S.D./Mean = 15/5000 = .003

The results show that income in the Goergia county is over three times more widely dispersed than in the New Jersey county.

Exercise II: Measures of Dispersion

Find two comparable sets of indicators of at least 10 cases each (e.g., two sets of test scores, the income of 10 cities in two different states, or some comparable figures that have policy relevance).

A) Make a bar graph of the two distributions.

B) Compute the mean, median, mode, range, standard deviation, and the coefficient of variation for the two sets of indicators.

C) Identify two units, one in each set of indicators, that you think are interesting. Compute the z-scores for these two.

D) In no more than a page, interpret what these measures indicate about the two sets of data, and about the two cases for which you computed z-scores. (Refer to the guidelines in Exercise I.)

E) Find the appropriate measure of tendency for each of the four variables displayed in the table below. Report the appropriate statistic for each of the four variables.

DATA FROM THE COUNSELING CENTER SURVEY

Student	Sex	Marital Status	Satisfaction With Services	Age
A	Male	Single	4	18
B	Male	Married	2	19
C	Female	Single	4	18
D	Female	Single	3	19
E	Male	Married	1	20
F	Male	Single	3	20
G	Female	Married	4	18
H	Female	Single	3	21
I	Male	Single	3	19
J	Female	Divorced	3	23
K	Female	Single	3	24
L	Male	Married	3	18
M	Female	Single	1	22
N	Female	Married	3	26
O	Male	Single	3	18
P	Male	Married	4	19
Q	Female	Married	2	19
R	Male	Divorced	1	19
S	Female	Divorced	3	21
T	Male	Single	2	20

Key: (4) Very Satisfied (3) Satisfied (2) Dissatisfied
(1) Very Dissatisfied

CHAPTER III: USING STATISTICS TO DECIDE IF THERE IS A RELATIONSHIP

That numbers are a more precise way of describing things should be clear enough. The second major use of numbers, determining relationships, is a little more complicated. Consider the following non-numerical description: "Lots of people don't like the current president's policies." We should all be sufficiently sensitive to the need of precision to be uncomfortable with such a vague assertion. Let's suppose we can gather information in order to be able to improve the statement as follows: "A survey of a random sample of adults has found that 40% express opposition to the president's policies." Is this really a lot? Or a little? We can argue for a long time about that, but at least we know precisely what was found out by one inquiry into that question.

So far, so good. Suppose, however, that we want to know more. We want to know whether the sort of people who are more likely to be opposed to the president's policies depends on certain voter characteristics. We underline the word "depends" because the policy analyst would say presidential support is a dependent variable, one that is affected by the characteristics of the people who have views on presidential policies.

By the way, a variable is defined as any characteristic of people or groups which varies. (Unfortunately, not all social science definitions are so clear or reasonable. Make sure you enjoy straightforward ones like this as much as possible.) A variable can be something that there is more or less of, such as level of support for presidential policies (in percentage of respondents saying they approve of the president); age (in years); financial status (in money earned); crime rate (in frequency of reported crimes); or size (in population). Alternatively, a variable may be some characteristic which has only a few possible distinct categories, such as gender (male/female), political party affiliation (Democrat/Republican/other), or type of political system (presidential/parliamentary/authoritarian).

A variable, then, is a characteristic of people or units which can take on different values; a dependent variable is a variable whose values are believed to depend on one or more other variables.

Let us go back to the variable called presidential support. What are some of the things on which it might depend? In 1985 the president, Ronald Reagan, was a handsome former movie actor with an appealing personality. It might, therefore, have been the case that females were more supportive of the president than males.

This last statement (aside from being insulting to women) is a <u>hypothesis</u>, defined as an assertion of relationship between two variables. Note that it doesn't matter whether the assertion is insulting or not; it doesn't even matter whether it is accurate or not. Its accuracy is something to be determined by actual research. (As a matter of fact, most actual research has indicated that females were significantly less supportive of Reagan.) What the statement must have in order to be a hypothesis is:

- A clearly stated dependent variable.

- A clearly stated variable which is said to affect the dependent variable. (This variable which affects the dependent variable is called by social scientists an <u>independent variable</u>. Remember this name. In social science, independent variables go together like love and marriage--sometimes even closer than love and marriage.

- A clearly stated relationship between the independent variable and the dependent variable.

Note that the hypothesis does not have to specify exactly how the two variables will be collected and measured. Support for the president, like almost every variable, can be measured in many different ways. All that is required is that the two variables are capable of being measured in some reasonable fashion.

But if any one of the preceding three characteristics is missing, the statement is not a hypothesis, and is not susceptible to social science analysis. For example, consider the following revised versions of the earlier statement about gender and support for the president's policies: read each one, and then try to figure out why it is not a hypothesis. (The answers will be given below; but don't look down at the answers until you try to figure it out for yourself.)

Non-hypothesis 1: Women have different views of politics than men do.

Non-hypothesis 2: Some people are more supportive of the president's policies than other people.

Non-hypothesis 3: There is a relationship between a person's gender and the person's level of support for the president's policies.

Non-hypothesis 4: More people should support the president's policies.

Non-hypothesis 5: Most persons who back the president will tend to give support for his policies.

Reasons Why Each Is Not a Hypothesis:

Non-hypothesis 1: The independent variable (male/female) is clear enough, but no dependent variable is clearly stated. The first commandment of social science is "if you ain't got a dependent variable, you ain't got nothin'!"

Non-hypothesis 2: Here the dependent variable is clear (level of support for the president) but the independent variable is impossibly vague and unspecific.

Non-hypothesis 3: This is a tricky error, and one which often trips up rookie social scientists. The dependent variable is clearly stated (level of support for the president). The independent variable is clearly stated (male/female). So what's the problem? The problem is the vagueness of the asserted relationship between the two variables. Just saying that there is a relationship is not enough. To be a hypothesis, a statement must assert the nature of the relationship as clearly and as precisely as possible.

Non-hypothesis 4: This is a value statement expressing a personal belief. It cannot be tested by social science research--ever.

Non-hypothesis 5: This is a tautology: the alleged independent and dependent variables are really different ways of saying the same thing--support for the president.

Most of the time any dependent variable may be hypothesized to be affected by several different independent variables. For example, here are several reasonable hypotheses about support for the president's policies:

- Wealthy people are more likely to be supporters of the president's policies.

- Younger people are more likely to oppose the president's policies.

- People living in the Northeastern part of the U.S. are more likely to oppose the president's policies than those living elsewhere.

- The higher the level of a person's education, the more likely a person is to support the policies of the president.

Which hypothesis should we work with? The answer depends on the interest and purpose of the researcher. If you are involved in women's politics, you might be interested in finding out how women feel about the president's policies. If you are a fund-raiser for the president, you might be interested in finding out about how people with lots of money feel about the president. All statements that satisfy the three criteria of a hypothesis are equally worthy of research. The basis for choosing is what you are interested in finding out.

Using Percentages of Subcategories

Percentages are often used to indicate the levels of some important condition. Percentage enables us to make comparisons when two or more absolute quantities are unequal. For example, unemployment levels are reported in percentages of those who are looking for work but who cannot find work. By translating our data into percentages we can compare unemployment, over time and in different localities.

Percentages can also be used to show the relationship between two variables, something that is often an important part of analyzing policies.

Suppose we take the one variable already analyzed in a previous chapter--the age categories of those 60 and over--and suppose we also know something else about the same population, namely the number of people above and below the poverty line. We will also simplify the age distribution into two categories--the percentage of people 60-74 years of age (73%), and the percentage of people 75 years and over (27%). This information can be presented as follows.

AGE DISTRIBUTION OF PEOPLE 60 YEARS OF AGE AND OLDER

Age	Percentage
60-74 years of age	73%
75 years & over	27%

For this same group of people 60 years of age and over, we also have information on their annual income. People whose incomes are so low as to make it difficult to meet basic living expenses are said to be "below the poverty line." Those with higher incomes are said to be "above the poverty line." The distribution of people in our hypothetical group is as follows:

PEOPLE 60 YEARS AND OLDER WHO ARE ABOVE AND BELOW THE POVERTY LINE

Below the Poverty Line	55%
Above the Poverty Line	45%

These two bits of information are useful enough. But we also may want to know whether there is any relationship between the two age categories and being under the poverty line. Specifically, is there evidence, as some people suggest, that people in the older age group are more likely to be below the poverty line?

The first thing we must point out is an error that people sometimes commit: that is to look at two tables such as those above and make a conclusion about the relationship between the two variables--in this case, age and being below the poverty line. For example, since a majority of the group are in the younger age category, and because a majority of the group are also below the poverty line, some people may erroneously conclude that the younger group are more likely to be below the poverty line. In fact, there is nothing that can be learned about the relationship between any two variables from merely looking at how each variable is distributed.

In order to find out whether either younger or older people are more likely to fall below the poverty line, we have to use percentages in a cross-tabulation of the two variables in what is called a "contingency table." A cross-tabulation in a contingency table shows the number of people in each category of age and poverty level.

Suppose in a sample of 200 people, we have the following actual frequencies: People who are 60-74 and below the poverty line; those who are 60-74 and above the poverty line; those who are 75 and older and below the poverty line; and, finally, people who are 75 and older and above the poverty line. This information can be efficiently portrayed in a contingency table such as below. Note that the categories of the dependent variable, the variable we want to describe--in this case, poverty level--make up the rows of the table.

The categories of the independent variable make up the columns of the table. In this case, the independent variable category--age--helps to explain who is in poverty. This is a rule you should always follow in constructing a contingency table.

Table 3.1

RELATIONSHIP BETWEEN AGE AND BEING BELOW THE POVERTY LEVEL

SOCIAL STANDING	AGE	
	75+	60-74
Below Poverty Level	50	60
Above Poverty Level	4	86
	54	146

SOURCE: Committee on the Aging, Special Report, Washington, D.C., 1985.

This table shows the same information as before--the percentage of the two different age groups, and the percentages of people above and below the poverty line. But it also shows something else of importance--the number of people in each of the two age categories cross-classified by the number of people above and below the poverty line. We have to do one more thing to make the relationship clearer. This is to compute the "column percentages" for the table. Column percentages enable us to compare for each age group the relative proportion of those below and above the poverty level. To compute the column percentages, divide 146 (representing all those age 60-74) by 60 (the number of those age 60-74 who are below the poverty level). This is 60/146 = .41, reported as a percentage as 41%. The percentage in column 1, row 2--those 60-74 who are above the poverty level--is 86/146 = .59, or 59%. The same percentage is used for the column percentage of the second column. The table below shows the two sets of column percentages.

Table 3.2

RELATIONSHIP BETWEEN AGE AND BEING BELOW THE POVERTY LINE
(with Column Percentages)

SOCIAL STANDING	AGE 75+	AGE 60-74
Below Poverty Level	50 (93%)	60 (41%)
Above Poverty Level	4 (7%)	86 (59%)
	54 (100%)	146 (100%)

SOURCE: Committee on the Aging, Special Report, Washington, D.C., 1985.

Notice that the two sets of column percentages total to 100 (as percentages always must) when you look down each column. But when you look across the rows, the percentages do not add to 100. This means that you can use column percentages to compare down each column. By looking at column percentages you can easily see that even though the raw number of those aged 60-74 below the poverty line is greater than those 75 and older who are below the poverty line, the percentage of the older group below the poverty line is much greater--93%--compared to 41% for the younger group. The contingency table, with column percentages, has clearly shown us what we wondered about--the people 75 or older are much more likely to find themselves below the poverty level than the group of people who are 60-74.

Here are some things to check for in making a cross-table:

- The overall title of the table should describe either the two variables being studied or the relationship which was found between the two. Thus, an alternative title of Table 3.2 could have been: "Older People Are More Likely to Fall Below the Poverty Level."

- Each variable should be clearly labeled in a place where the name obviously belongs to the rows and the columns.
- Present percentages along with the raw figures in each cell. The percentages should be according to the independent (column) variables. Do not present more than one percentage in each cell.
- The source of the variables should be clearly stated below the table.
- Any omissions, such as not answering or don't knows, should be clearly explained.
- Any combining of categories which were originally separate should be explained.
- All appropriate statistics should be presented immediately beneath the table.

Exercise III

Find at least 10 individuals (people, cities, nations, or some other unit) for which you have two indicators (such as population and size, or education and income, or unemployment and crime rate). Decide which of the two indicators is the more reasonable independent variable. Using that for the columns and the other variable for the rows, construct contingency table. Compute column percentages. In no more than one paragraph, summarize and interpret the relationship shown in the table. (Refer to the guidelines in Exercise I.)

CHAPTER IV: DECIDING WHETHER A RELATIONSHIP IS WORTH PURSUING OR WHETHER IT'S BETTER JUST TO FORGET ABOUT IT

No, the title doesn't indicate that this has suddenly become a guide to your personal life. It is still about statistics. In fact, it is about one of the most important things statistics can do for you--giving you some guidelines for whether a relationship between two variables is strong enough for you to conclude that it really helps explain why, for example, there is support for a president, why crime rates are at the level they are, or whatever it is that you are studying. Determining whether there is a relationship between two variables is a tricky proposition, as we shall see.

Before considering the techniques used in testing hypotheses, let's stop for a minute to consider a simple question. How do we decide what hypotheses are worth our valuable time and energy as the subject of our research? Nothing in statistics, or in all of social science for that matter, gives the answer. The answer is: whatever turns you on because of your personal interest about human behavior, because of an important interest in something you read, because of your job, or anything else in your life.

One good source of hypotheses is statements made about public life and policy. Many statements are not made exactly like hypotheses, but they clearly imply a hypothesized relationship. Examples: "Capital punishment is a deterrent to murder." "Special pre-school programs don't help students after the second or third grade." "Don't pay ransoms to terrorists; it just encourages them in the future." Try restating each of these as formal hypotheses. If you are sensitive to what is said by the people around you, or to what is said in the news and elsewhere, you will begin to pick up lots of statements that are more or less disguised hypotheses.

One further point about social science terminology. You will frequently hear people say: "My theory is that..." What usually follows is a hypothesis. A theory is a general set of beliefs and assumptions, from which specific hypotheses are derived. For example, to return to our hypothesis about females being more supportive than males of a handsome, charming president, a hypothesis derived from a theory about physical attractiveness leading females to respond to political figures in terms of their physical appearance and behavior. Each hypothesis derived from that theory is a partial test of the validity of the theory--but it isn't the theory itself, which is almost never tested in the social sciences.

If you hear someone say something like: "My theory is that southerners commit more murders than northerners," you should reply: "That's not a theory, that's a hypothesis." This will (1) impress any social scientists who happen to be listening, (2) demonstrate that your tuition money has been well spent, and (3) irritate everyone who hears you.

Now let's consider a hypothesis that we are pretty sure is false: Women are more likely to be left-handed than men. We would expect to find the pattern summarized in the table below.

Table 4.1

EXPECTED RELATIONSHIP IN GATHERING DATA TO TEST THE HYPOTHESIS: "WOMEN ARE MORE LIKELY TO BE LEFT-HANDED THAN MEN."

HANDEDNESS	GENDER: Female	GENDER: Male
Left	Higher proportion of females than in the male-left cell.	Lower proportion of males than in the female-left cell.
Right	Lower proportion of females than in the male-right cell.	Higher proportion of males than in the female-right cell.

The wording of how many people should be expected in each cell is pretty cumbersome. In fact, it's time to start practicing what we preach and to begin to fill in the tables with numbers rather than words. The first thing we want to know is what will be the marginal totals for the cell? The marginal totals are <u>the total number of cases in each category of the variables we use</u>--typically placed out at the margin of the cross-table. There are approximately equal numbers of men and women in contemporary America. (Trivia experts know that statement is not exactly correct, but it's close enough for a learning package.) We also know that about 20 percent of the population is left-handed. Consequently, in a random sample of 100 people, we would expect to find the marginal totals in Table 4.2.

Table 4.2

EXPECTED MARGINAL TOTALS FROM TABLE 4.1

	GENDER		
HANDEDNESS	Female	Male	
Left			20
Right			80
	50	50	100

We get those expected marginal totals from what we know about the distribution of males/females and people's handedness. This general knowledge does not tell us what the totals in each cell will be: that is what we have to find out.

Before continuing, one point in Table 4.2 has to be mentioned because it will be used later. This is the number of degrees of freedom (df) for the table. In statistics, degrees of freedom is not the high philosophical question it is in some quarters. It is merely a technical attribute of a cross-table. The table, as with all 2 x 2 tables, has a df of 1. What that means is that you can arbitrarily put any number (not larger than the marginal total) in any one of its four cells, and still be consistent with the marginal totals. But once you have put that one number in, you cannot arbitrarily add any other numbers. For example, if the lower right-hand cell (male, right-handed) has 40 cases, we know that the number of male, left-handed people in this sample is 10 (since the total number of males is 50). Furthermore, the number of right-handed people is 80, and the number of left-handed females is 10. In short, when we have a two-row and two-column table and the marginal frequencies for the two variables, we can freely place only one number in a cell; after we do that, the frequencies in all the other cells are determined. As the number of rows and/or columns of a table increases, the df also increases. The formula is: the number of rows, minus 1, multiplied by the number of columns, minus 1 = (r-1) x (c-1) as it is usually stated in mathematical terms. For a 2 x 4 table, the df is 3. What is the df for a table of three rows and three columns?

With that rather tedious, but important technical point behind us, let's see how statistics can help us determine the significance of a relationship between two variables.

We'll go back to our study of sex and handedness, testing the hypothesis that women are more likely to be left-handed than men. Suppose for the moment that the hypothesis is not correct--that there is no systematic relationship between handedness and sex. In that case, what frequencies would we find in each cell of our 2 x 2 table? The answer is that each cell will contain the proportion of people of each category of gender (male/female) multiplied by the proportion of each category of handedness (left/right). It's really not that bad. The calculation goes like this: 50 percent of the 100 people in the sample are male; 80 percent of the 100 people are right-handed; therefore, the number of people in the male-right-handed cell we would expect is (50 x 80) /100, or 40. And so it goes for each cell, multiplying the corresponding row and column totals, and dividing by the total number in the sample. That gives us the expected frequencies as shown in Table 4.3.

Table 4.3

EXPECTED FREQUENCIES IF THERE IS NO SYSTEMATIC RELATIONSHIP BETWEEN GENDER AND HANDEDNESS (Calculation Shown in Each Cell)

	GENDER		
HANDEDNESS	Female	Male	
Left	$\frac{50 \times 20}{100} = 10$	$\frac{50 \times 20}{100} = 10$	20
Right	$\frac{50 \times 80}{100} = 40$	$\frac{50 \times 80}{100} = 40$	80
	50	50	100

Clearly, this table shows no systematic relationship. The proportion of left- and right-handed people is exactly the same for females and males. But what if instead of being divided 10-40, the females were divided 11-39 and the males divided 9-41? Now the proportions are not exactly the same; there is, in fact, a slightly larger proportion of female left-handers (in accord with our hypothesis). But is such a small difference big enough? What about 12-38 and 8-42? At some point a difference would be big enough to satisfy even the most skeptical. But proceeding on an intuitive basis alone will produce many differences of opinion, argument, and, who knows, even fisticuffs. Statisticians feel that most people already have enough things to fight about, so they have come up with a procedure that all agree upon as the way to determine how big a difference is big enough to be called "significant" rather than just an artifact of chance.

This procedure is based on measuring the amount by which an actual set of cross-tabulated observations deviates from the cross-tabulations that would be obtained if there were no systematic relationship between the two variables. The bigger the deviation from chance, the more likely a significant relationship exists.

The way to calculate this measure is called <u>chi^2</u>. ("Chi" is prounounced ki, to rhyme with guy.)

<u>FORMULA</u>

$$\text{chi}^2 = \sum (f_o - f_e)^2 / f_e$$

f_o is the observed frequency

f_e is the expected frequency

The first step in calculating chi^2 is to compute the expected frequences in each cell of a table for which you have calculated the actual frequencies. In Table 4.4 below, we have pretended that we collected a distribution slightly different from the expected frequencies for the cross-tabulation of gender with handedness from a random sample of the population.

Table 4.4

ACTUAL FREQUENCIES OF
A CROSS-TABULATION OF GENDER AND HANDEDNESS

HANDEDNESS	GENDER: Female	GENDER: Male	
Left	11	9	20
Right	39	41	80
	50	50	100

The calculation of chi^2 is to do the following mathematical operation for each cell of the table: Find the difference between the actual and the expected value, square the difference, and divide that number by the expected value. The result is a measure of the relative difference between expected and actual values in each cell of the table. Add together all the numbers for each cell; this total is the value of chi^2 for that table. It should be clear that the larger the value of chi^2, the bigger the discrepancy between the actual values and the expected values, and, therefore, the more likely that some systematic realtionship is present between the two variables. To interpret when chi^2 is big enough, we consult a table of significant values of chi^2. To use such a table, part of which is reproduced on the following page, look down the first column until you find the numbers of degrees of freedom your table has. Then read across until you find the two numbers that encompass the actual chi^2 you have calculated. At the top of the columns are the approximate probabilities that the relationship you have is due to chance rather than systematic relationship between the two variables.

Table 4.5

CHI SQUARE PROBABILITIES

Degrees of Freedom	Probability .1	.05	.01	.005
1	1.64	2.71	5.41	6.63
2	3.22	4.60	7.82	9.21
3	4.64	6.25	9.84	11.34
4	5.99	7.78	11.67	13.28
5	7.29	9.24	13.39	15.09
6	8.56	10.64	15.03	16.81
7	9.80	12.02	16.62	18.47
8	11.03	13.36	18.17	20.09
9	12.24	14.68	19.68	21.67
10	13.44	15.99	21.16	23.21
11	14.63	17.27	22.62	24.72
12	15.81	18.55	24.05	26.22
13	16.98	19.81	25.47	27.69
14	18.15	21.06	26.87	29.14
15	19.31	22.31	28.26	30.58
20	25.04	28.41	35.02	37.57
25	30.67	34.38	41.57	44.31
30	36.25	40.26	47.96	50.89

SOURCE: R.R. Fisher, Statistical Methods for Research Workers, 14th ed., New York: Hafner Press, 1970.

The larger the chi^2 values, the lower is the probability that the relationship is due to chance. For example, in a table in which df = 3, a chi^2 of 5.22 falls between the first and second columns. (It is greater than 4.64 and less than 6.25.) By looking at the top of the columns, we can say that the probability that a value of chi^2 of 5.22 was obtained by chance is less than .1 (10%, as it is frequently written) and greater than .05.

Let's go through the calculations for the sample in the preceding Table 4.4 (page 32).

Cell	Expected Frequency (f_e)	Observed Frequency (f_o)	$(F_e - F_o)^2 / F_e$	=
Female Left-handed	10	11	1/10	= .1
Female Right-handed	40	39	1/40	= .03
Male Left-handed	10	11	1/10	= .1
Male Right-handed	40	41	1/40	= .03

The resulting value of chi^2 is 26. As you can see from the table, this is quite tiny. The probability of this result being due to chance is much larger than the .1 which is the largest column. The reason that nothing greater than .1 is given is that anything with more than a 10% probability of being due to chance is pretty much rejected as statistically insignificant. Conversely, anything which has a .05 or less probability of being due to chance is normally accepted as being worth talking about. The .05 probability of chance is widely accepted by statisticians as the threshold between research results which are a hit or a miss.

Suppose that the gathering of information produced the results shown in Table 4.6, with a much larger proportion of men (but still a minority of men), turning out to be left-handed.

Table 4.6

IMAGINARY DISTRIBUTION OF RESULTS IN TESTING THE GENDER AND HANDEDNESS HYPOTHESIS

HANDEDNESS	GENDER: Female	GENDER: Male	
Left	15	5	20
Right	35	45	80
	50	50	100

$chi^2 = 6.26$ $df = 1$ $p < .01$

Since the marginals are still the same, the expected frequencies are the same. In this case, the deviations of actual values are much greater. In fact, $chi^2 = 6.26$, which with a df = 1 table is somewhere between .01 and .005 (note the two zeros) of being due to chance--a result well within conventional limits of being a statistically significant finding. Note that we report the probability of chance (p) as less than the higher figure rather than an exact number.

Exercise IV

Identify a data set of at least 20 cases that measure some policy condition and that has two categoric variables. Prepare a contingency table that meets all the criteria discussed in Chapter III. Also prepare a table showing the expected frequencies. Calculate the chi^2 and report it under the contingency table of actual frequencies, along with the df and the probability of error. Interpret the results in one paragraph. (Refer to the guidelines in Exercise I.)

CHAPTER V: MEASURES OF ASSOCIATION--NOMINAL AND ORDINAL DATA

So far we have been concerned with using chi^2 to see if there is any relationship between two variables that is different from chance and analysis. The final two chapters focus on describing the relationship between two variables; that is, measures of coefficients, indicate the degree of association or correlation between variables which are of nominal, ordinal, interval, or ratio level of measurement. Correlation does not tell us about causality, but it does indicate the covariation, or direction and extent of change that occurs in each variable when the other changes.

A. The Del Statistic

The first descriptive statistic we will consider is for relationships portrayed in a contingency table. When the variables are simply catagoric, without any order, these are called nominal. Many different statistics can be used to analyze cross-table (or contingency table) data. Only one, however, can handle any size table with any assumptions about the data. This statistic is del.

The del statistic can be used in any table, as long as you can hypothesize a relationship between the two variables.* The hypothesis establishes that certain cells are consistent with the expected results of the hypothesis, and certain cells of the table are error cells because any cases within them have traits inconsistent with the hypothesis. Let's look back at Table 4.6, and hypothesize that females are more likely to be left-handed. This makes the female-right and male-left cells error cells.

The principle underlying del is to see how large the error cells are, relative to the expected value in those cells, just as we did for all cells in calculating chi^2. The formula for del is $E_c - E_o / E_c$, where E_c stands for the frequency within all error cells expected by chance and E_o is the actual number of observations in all the error cells. Going back to Table 4.6 in the section on chi^2, we can calculate the del as follows: E_c is 40 + 10 = 50; E_o is 35 + 5 = 40. The formula then becomes 50 - 50/59 = .20, which is a slight support for the hypothesis.

A more formal, generalized procedure for calculating del is on the following page.

---

*There are a number of other frequently used statistics. These include chi^2 (described above) and Gamma. For further information on these statistics, see J. M. Blalock, Jr., Social Statistics, New York: McGraw-Hill, 1972.

Summary of Del Calculations

These calculations can be used for tables of any number of rows and columns. All that is necessary is to identify, with a clear hypothesis, which are the error cells--that is, the cells corresponding to observations inconsistent with your hypothesis.

FORMULA

del = $E_c - E_o / E_c$

E_c is the frequency within all error cells expected by chance

E_o is the actual number of observations in all the error cells

Given Variable 1 (independent variable) divided into categories A and B; and Variable 2 (dependent variable) divided into categories C and D. Hypothesis: People who are in category A of Variable 1 are more likely to be in category D of Variable 2. Example. Males are more likely to be taller (than females), or people who are above average in wealth were more likely to vote Republican (compared to those with below average wealth).

VARIABLE 1

VARIABLE 2	Category A	Category B	
Category C	f_1 E_c for this cell = $\frac{A \times C}{TOTAL}$ (error)	f_2 (hypothesized)	C
			+
Category D	f_3 (hypothesized)	f_4 E_c for this cell = $\frac{B \times D}{TOTAL}$ (error)	D
	A	+ B	= TOTAL

INTERPRETATION OF DEL

1.00-.80 -- Very Strong Support
.80-.60 -- Strong Support
.60-.40 -- Support
.40-.20 -- Some Support
.20-.10 -- Slight Support
.10-.00 -- No Support

$$\text{Del} = \frac{E_c - E_o}{E_c}$$

$$E_o = f_1 + F_4$$

$$E_c = \frac{A \times C}{TOTAL} + \frac{B \times D}{TOTAL}$$

(If the score is negative, this is evidence for a relationship opposite that which was hypothesized.)

Del Statistic with Larger Than 2 x 2 Tables and with Ordinal Data

When we first introduced the del statistic, we promised that it could be used on tables of any size. And so it can. Suppose, for example, you have the following hypothesis you wish to test: Catholics and Jews are more likely than Protestants to vote for the Democratic party.

That hypothesis would lead to the table shell presented below. In this table, there are three error cells. To calculate the E_c you would multiply the appropriate row and column totals for each cell (as you learned before) and divide by the total number of cases.

Table 5.1

TABLE SHELL AND ERROR CELLS FOR THE HYPOTHESIS: "CATHOLICS AND JEWS ARE MORE LIKELY THAN PROTESTANTS TO VOTE FOR THE DEMOCRATIC PARTY."

	RELIGION		
VOTING BEHAVIOR	Catholic	Jewish	Protestant
Democratic			error
Republican	error	error	

When you are dealing with nominal variables that have more than two categories, the extension of the calculation for del is quite straightforward. Sometimes, however, you are interested in something that measures the relationship between two ordinal variables. In that case, the selection of the error cells is a little trickier. Consider, for example, the following hypothesis: The higher people's social standing, the more likely they are inclined to favor increased military spending. Let us further suppose that we have measured social standing with three ordinal categories--high, middle, and low--and that opinion about military spending has been classified as more or less. This would lead to the below table 5.2 in which two of the cells are clearly error cells, two are not error cells, but the remaining two are questionable.

Table 5.2

TABLE SHELL TO ANALYZE THE DATA FOR TESTING THE HYPOTHESIS: "THE HIGHER PEOPLE'S SOCIAL STANDING THE MORE THEY ARE INCLINED TO FAVOR INCREASED MILITARY SPENDING."

	SOCIAL STANDING		
OPINION	High	Middle	Low
Spend More	not an error	?	error
Spend Less	error	?	not an error

The more devious among you may be able to figure out a couple of possible tricks to solve this problem. One would be just to ignore the middle category and analyze a 2 x 2 table. The reason this is not a good solution is that it would involve throwing away some data--all those fine souls who are in the middle rank of social standing. When you go out and gather your own data at great time and cost, you will not be so willing to throw away data just to make it easier to compute the statistics.

Another apparent solution might be to convert opinion about military spending into three categories: spend more/spend less/ spend the same. But this is truly a step away from solving the problem. It would tell you that one more cell was a non-error, but it would also increase the number of question-mark cells. Just look at Table 5.3!

Table 5.3

TABLE SHELL FOR TESTING THE HYPOTHESIS: "THE HIGHER PEOPLES' SOCIAL STANDING THE MORE THEY ARE INCLINED TO FAVOR INCREASED MILITARY SPENDING"

	SOCIAL STANDING		
OPINION	High	Middle	Low
Spend More	not an error	?	error
Spend the Same	?	not an error	?
Spend Less	error	?	not an error

So much for that tempting but foolhardy solution. Of course, it may well be that you will have data in the form shown in either Table 5.2 or Table 5.3 and you will want to be able to decide how to calculate expected error unequivocally, without throwing away those cases that don't nearly fit into the error/non-error classification as we have been using it so far.

The correct way out of this difficulty is to make use of the fact that we have here two ordinal variables. This means that the order of the categories is important. The order of high, middle, and low makes sense, but high, low, middle would be putting the categories out of order. What is important in the relationship between two ordinal variables is not the frequencies in any given cell, but the frequencies in the comparative ranks in the relationship between the two. To explain what we mean a little more clearly, let's consider a table with some numbers.

Table 5.4

ILLUSTRATIVE DISTRIBUTION OF DATA GATHERED TO TEST THE HYPOTHESIS: "THE HIGHER PEOPLE'S SOCIAL STANDING THE MORE THEY ARE INCLINED TO FAVOR INCREASED MILITARY SPENDING."

OPINION	SOCIAL STANDING: High	Middle	Low	
Spend More	20	15	35	70
Spend the Same	5	20	10	35
Spend Less	10	10	45	65
	35	45	90	170

When dealing with ordinal data, be sure, as in Table 5.4, to set up the table so the cell represented by the leftmost column and the upper row is not an error, according to the hypothesis. In the table above, high social status people who want to spend more on defense represent such a non-error cell. What the logic of identifying error for ordinal variables tells us to do is to look for mistakes in paired comparison between sets of observations. According to the hypothesis, it is unnecessary that no people of high social standing favor spending less. For support of the hypothesis, it is necessary that, in comparing high standing with middle, high with low, and middle with low, the pairs of observations will all favor more spending. Those cases that do not favor spending represent the frequency of error. The rule of thumb is to total, for all cells, the product of each cell frequency and the total frequencies of all those cells that are to the left of the cell, but not above it. For example, consider the 15 middle-class respondents who want to spend more; they themselves are not an error. But they are an error (with respect to the hypothesis) in comparison to the 20 upper-class respondents who want to spend more, the 5 upper-class who want to spend the same, and the 10 upper-class who want to spend less. To calculate the error from this comparison, multiply 15 x (20 + 5 + 10), or 525.

The full range of comparisons is shown below:

20 x 0 =	0
5 x 0 =	0
10 x 0 =	0
15 x (20+5+10) =	525
20 x (5+10) =	300
10 x 10 =	100
35 x (10+15+5+20+10+10) =	2800
10 x (5+20+10+10) =	450
45 x (10+10) =	900
TOTAL OBSERVED ERROR =	5425

Calculating the E_o follows the same logic. First calculate expected frequencies in each cell, as has been done before, with the marginal totals. Then use those expected numbers to make the same comparisons done above with the observed error. Then use the same del formula: $E_c - E_o / E^c$.

It is admittedly a little tedious to do it this way. But it is a much more efficient way to use the extra information available from the ordinal data.

Table 5.5 shows the calculation of expected frequencies and the application of the del formula, using the totals from Table 5.4.

Table 5.5

EXPECTED FREQUENCIES FROM TABLE 5.4

OPINION	SOCIAL STANDING High	Middle	Low	
Spend More	14.4	18.5	37.1	70
Spend the Same	7.2	9.3	18.5	35
Spend Less	13.4	17.2	34.4	65
	35	45	90	170

The following fully completed Table 5.6 includes both del and chi^2 calculation.

Table 5.6

TESTING THE HYPOTHESIS: "THE HIGHER PEOPLE'S SOCIAL STANDING THE MORE THEY ARE INCLINED TO FAVOR INCREASED MILITARY SPENDING."

OPINION	SOCIAL STANDING High	Middle	Low	
Spend More	20 (57%)	15 (33%)	35 (39%)	70
Spend the Same	5 (14%)	20 (45%)	10 (11%)	35
Spend Less	10 (29%)	10 (22%)	45 (50%)	65
	35	45	90	170

del = .11 $chi^2 = 26.9$ df = 4 $p < .005$

The chi^2 shows a significant deviation from chance. The del of .11 shows a slight tendency for support of the hypothesis.

B. T-Test

In studying situations on which we have to make decisions, we often want to know whether two groups have a significant difference in some important measure. Examples of this type of concern include such questions as the following:

- Do males and females have significantly different GPAs (Grade Point Averages)?
- In a particular workplace, is there a significant difference between the average salaries earned by men and women?
- Do upper-division and lower-division students study for significantly different amounts of time?

The description of the differences between two groups in answer to the above and countless similar questions is easy enough. All you need to do is to calculate the mean score for each group and to see which is greater. However, just observing a difference is not enough. You must also be able to determine whether any difference is statistically significant. The statistic called the t-test is a simple way to decide this.

The way to figure out whether a certain difference is significant is to follow the steps below:

Steps in Calculating a T-Test

1. Formulate a hypothesis that one of two groups will have a higher average on some variables than the second group. (Examples: Males have greater weight than females; southern states have higher murder rates than northern states.)

2. Find the mean of group 1 and the mean of group 2. (If the difference in means is not the same as hypothesized, you can stop right here.)

3. Compute the variance for group 1 and the variance for group 2. Variance is computed by first subtracting the score of each case from the mean and squaring that difference. Next, sum these squared differences and divide the total by the number in the group. (If you were to take the square root of this variance, you would have the standard deviation.)

4. Subtract 1 from the number of cases in group 1 and divide the resulting number into the variance of group 1. Do the same thing with group 2.

5. Add the two resulting numbers together and find the square root of the total.

6. Divide this square root into the difference between the two means (step 2).

7. The resulting number (ignoring whether it is positive or negative) is the t-score. It must be evaluated according to a t-test table, summarized below. Df in the table stands for degrees of freedom, which in this case is the number of cases in group 1, plus the number of cases in group 2, minus 2.

The .05 level of significance is interpreted the same as with the chi^2 square statistic; it is a level that indicates you will be wrong no more than 1 time in 20 if you conclude that a difference you have observed is a difference that would hold up in the population from which you have sampled. Table 5.7 below displays the minimum required T-value for .05 level of significance.

Table 5.7

MINIMUM REQUIRED T-VALUES FOR .05 LEVEL OF SIGNIFICANCE

Df	Minimum T-Value	Df	Minimum T-Value
1	6.314	16	1.746
2	2.920	17	1.746
3	2.353	18	1.734
4	2.132	19	1.729
5	2.015	20	1.725
6	1.943	21	1.721
7	1.895	22	1.717
8	1.860	23	1.714
9	1.833	24	1.711
10	1.812	25	1.708
11	1.796	26	1.706
12	1.782	27	1.703
13	1.771	28	1.701
14	1.761	29	1.699
15	1.753	30	1.697
		31	1.644

These calculations work so that a difference is more likely to be significant under the following conditions: (1) the larger the difference between the two means, (2) the smaller the variance of the scores in each of the two groups, and (3) the larger the number of cases in the two groups.

FORMULA

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S^2_{x_1}}{N_1 - 1} + \dfrac{S^2_{x_2}}{N_{x_2} - 1}}}$$

$\overline{X}_1$ and $\overline{X}_2$ are the means of group 1 and group 2

$S^2_{x_1}$ and $S^2_{x_2}$ are the variances of group 1 and group 2

N_{x_1} and N_{x_2} are the number of individuals in group 1 and group 2

An Application

Here is an example of this calculation. We will be interested in testing the hypothesis that in the sample of voters on which we have gathered data, those who are affiliated with the Democratic party have lower annual incomes than those affiliated with the Republican party.

1. The hypothesis is as follows: On the average, Democrats have lower annual incomes than Republicans.

2. In our hypothetical example, we find that the average annual income of the Democrats is $15,000; the average annual income of Republicans is $15,040. Clearly, the difference is in the direction we have hypothesized. We shall therefore proceed to see if this difference is statistically significant.

3. We calculate the variances of the two groups and find that the variance in average annual income for Democrats is 16,000 and the variance in average annual income for Republicans is 9,000.

4. ? our two samples, we have 35 Democrats and 42 Republicans. For the Democrats, 35-1 is divided into 16,000, giving a score of 470.59. For the Republicans, 42-1 is divided into 9,000, giving a score of 219.51.

5. The total of these two numbers is 690.10. The square root is 26.27.

6. This square root is divided into 40 (the difference between the mean annual income for Democrats and Republicans, step 2). The answer is 1.52.

7. The resulting t-score, 1.52, is evaluated according to the number of degrees of freedom in the table above. For a total of 35 in one group and 42 in the other group, the degrees of freedom is 76, which is off the chart. Therefore, a t-score of 1.644 would have been required for statistical significance. Since the score of 1.52 is less than that, we conclude that while there is a difference in the direction we have hypothesized, that difference is not statistically significant with an acceptable probability of error (.05 or less).

Note that the difference was close to being significant. If the difference in means had been just a little larger, the t-test would have been significant. For example, if the difference had been $44 (rather than $40), the t-score would have been 1.67, exceeding minimum requirement. Likewise, if the variances had been smaller, the t-test would have been significant. Or, if the size of the samples had been larger, the t-test would have been significant. For example, if the number of cases had been 150

Democrats and 200 Republicans, the t-test would have been 3.25, well above the 1.644 required for statistical significance.

Exercise V

(A) Using the data from Exercise IV or a comparable data set, create a contingency table. State a hypothesis. Compute the del statistic; compute ordinal del if that is appropriate. Interpret the results.

(B) Identify a data set for which the T-test is appropriate, or at least 10 cases (e.g., differences of males and females on a test; the difference of northern and southern states on average income). State a hypothesis. Compute a T-test and interpret the results. (Refer to the guidelines in Exercise I.)

CHAPTER VI: MEASURES OF ASSOCIATION--INTERVAL DATA

This chapter describes the ways of measuring the association between two numerical variables, such as height and weight, age and income, unemployment and inflation. Such variables are called interval or ratio variables.

A. Using Scatterplots

A scatterplot, or scatter diagram, is a graph in which one variable is scaled along the Y (or vertical) axis and the other is scaled along the X (or horizontal) axis. Pairs of values can then be represented as points on the graph. The pattern or "scatter" which the points describe suggests types of association of the variables.

The following table indicates the number of hours or study per week for 10 students, and their respective grade point averages on a scale of 2.00 to 4.00. We are interested in determining the association, if any, between the two variables. In other words, do students who spend more hours studying tend to get higher grades?

Table 6.1

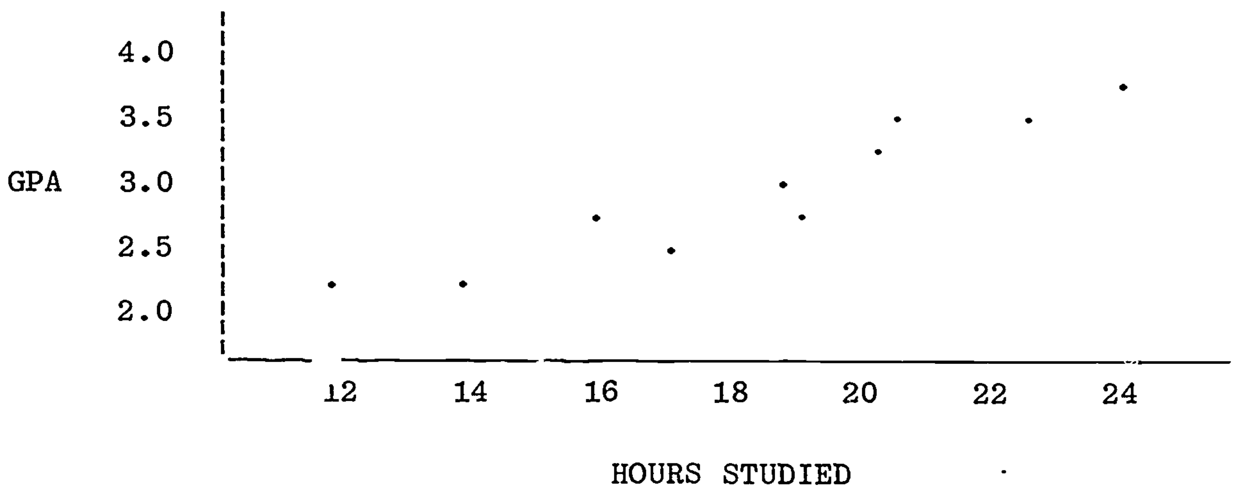
HOURS OF STUDY AND GRADE POINT AVERAGE

Student	Hours Studied (X)	Grade Point Average (Y)
A	19	2.9
B	14	2.3
C	21	3.5
D	12	2.1
E	24	3.8
F	19	2.7
G	20	3.1
H	16	2.7
I	23	3.6
J	17	2.6

Using this data, we can construct the below scatterplot.

Figure 6.1

SCATTERPLOT OF HOURS OF STUDY AND GRADE POINT AVERAGE



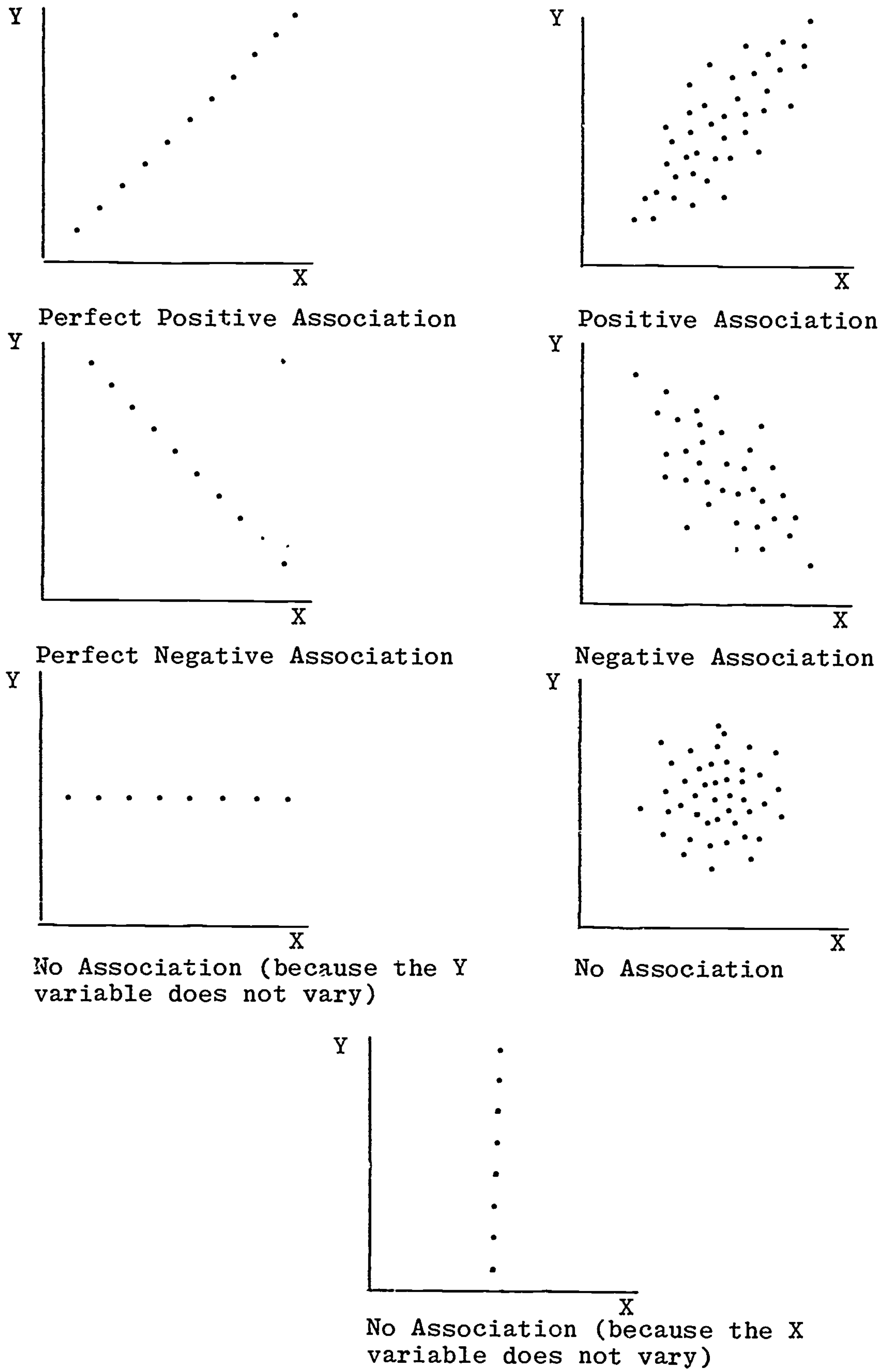
You will note that the points tend to line up from lower left to upper right. This tells us that "as hours studied increases, grade point average increases." This kind of propositional statement defines a positive association.

Thus far, we have graphed only situations in which both variables have been positive. However, it is possible to encounter negative values for variables. For example, population growth rates can be negative, as well as government budgets and corporate earnings if spending exceeds revenues. In some situations, there may be no relationship between variables.

In a social program that does not work, for example, increasing the money spent will not improve social conditions. In general, money spent on programs to rehabilitate criminals does not lead to lower crime rates.

Figure 6.2 presents seven hypothetical patterns of data on scatterplots. One of these seven will emerge for any scatterplot that is developed. To make an interpretation, you need to make a judgment about the pattern of data to determine whether there is a relationship and, if so, in what direction. Formal statistical calculations can be applied to determine a precise mathematical representation of the pattern of data (e.g., linear regression and Pearson's R^2), but making a judgment is often sufficient.

Figure 6.2: SCATTERPLOTS



B. Linear Regression Analysis

In addition to presenting the relationship between the two interval variables graphically in a scatterplot, one more bit of information can be presented. This is called the linear regression equation, which in general estimates the amount by which the dependent variable increases (or decreases) as the independent variable increases. The formulas are as follows:

FORMULAS

Linear Regression equation (straight line) $Y = a+b(X)$

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \qquad a = \bar{Y} - b\bar{X}$$

X = independent variable $\quad Y$ = dependent variable

$\bar{X}$ = the mean of the independent variable

$\bar{Y}$ = the mean of the dependent variable

b is the slope

a is the point of the vertical axis (representing the dependent variable) where the line would cross if the independent variable had a value of 0.0

You can, for example, calculate the regression equation that tells how much an increase in a person's studying increases overall grade point average (GPA).

This equation is: GPA = .286 + .14(Hours).

This means that to make an estimate of a person's GPA, on a basis of the person's number of hours studied, you must take the number .14, multiply it by the hours the person studied and then add .286 to the product. Let's see how well this works with a couple of examples. Consider student A, whose hours studied is 19. She would have an estimated GPA of .14(19) + .286--or 3.00. This is quite close to the actual value of A's GPA of 2.9. Or consider student J, whose hours studied--17--translates into an estimated GPA of 2.7, compared to his actual GPA of 2.6. Almost every estimated score is quite close to the actual score. This relationship can be seen from looking at the scatterplot on page 49, in which nearly every point is close to a straight line.

In the foregoing equation, the number by which hours is multiplied--.14--is called the slope because it represents ho steeply the dependent variable (in this case GPA) increases in association with increases in the independent variable (hours studied). If the slope coefficient were larger, say .30, or even 2.00, the line would be much steeper. The slope can also be negative, indicating that as the independent variable increases across a set of observations, the values of the dependent variable tend to decrease. (See the plot labeled perfect negative association, p.50.) Among nations of the world, researchers usually find a negative relationship between per capita wealth and the amount of internal strife and violence, supportive of the proposition that the greater the wealth of a nation, the less its domestic instability.

The formula for the above slope is designated by the symbol "b" in regression equations.

The constant term, .286 in the foregoing example, is the point of vertical axis (representing the dependent variable) where the line would cross if the independent variable had a value of 0.0. It is usually represented by the letter a in a regression equation. The general form of the linear (straight line) regression equation is written $Y = a+b(X)$. Y stands for the dependent variable; X stands for the independent.

Here is the procedure for computing the b value and the a value from a set of data with an example using the hours studied and GPA for 10 students.

Step 1: Compute the following calculations:

Student	Hours Studied (X)	X^2	GPA (Y)	(XY)
A	19	361	9	55.1
B	14	196	2.3	32.2
C	21	441	3.5	73.5
D	12	144	2.1	25.2
E	24	576	3.8	91.2
F	19	361	2.7	51.3
G	20	400	3.1	62.0
H	16	256	2.7	43.2
I	23	529	3.6	82.8
J	17	289	2.6	44.2
Sums	185	3553	29.3	560.7

N (the number of observations) is 10.

Step 2: Multiply N times the sum of XY: 10 * 560.7 = 5607.

Step 3: Multiply the sum of the independent variable times the sum of the dependent variable: 185 * 29.3 = 5420.5.

Step 4: Subtract the product in step 3 from the product in step 2. This is the numerator, 5607 - 5420.5 = 187.

Step 5: Multiply N times the sum of X^2: 10 * 3553 = 35530.

Step 6: quare the sum of X: 185 * 185 = 34225.

Step 7: Subtract the square in step 6 from the product in step 5. This is the denominator, 35530 - 34225 = 1305.

Step 8: Divide the numerator (step 4) by the denominator (step 7). This is the value of b: 187/1305 = .143295. (Report it to two decimal places = .14.)

To calculate the value of the constant term, a, follow steps 9 and 10:

Step 9: Subtract b x (the sum of X) from the sum of Y. This is the numerator.

29.3 - (.1432 * 185) = 29.3 - 26.438 = 2.862

Step 10: Divide the numerator from step 9 by N. This is the value of a: 2.862/10 = .286.

Interpreting a regression equation under the right conditions can frequently lead to important substantive conclusions about how much of a unit change in the independent variables produces how much of the dependent variable. Examples include how much money spent on educational programs improves average achievement scores of the students in the programs; how much time spent studying leads to higher GPA; how much time spent on campaigning produces more votes in the ensuing election.

But before the equation can have meaning, the following conditions must be present:

- There must be an essentially straight-line relationship between the two variables (as evidenced by your knowledge of the topic and by inspecting the scatterplot.)

- The Pearson's R^2 (described below) must be reasonably strong, indicating that the two points are reasonably close to the hypothetical straight line.

- The number of widely discrepant cases must be small (such as very high on the independent and very low on the dependent variable in an otherwise positive relationship). Such discrepant cases are called outliers. If they are found within a scatterplot, they should always be examined very closely. There may be something special and different about the outlier cases that makes them different from the general pattern. It may be reasonable to exclude them from the general analysis as long as you report what you are doing.

Also remember that the interpretation of the regression equation makes sense only within the range of data that have been gathered. Suppose we estimated what would happen if a student were to study 27 hours; the equation would estimate the GPA as .14(27) + .286 = 4.12. Since this is a higher GPA than humans are allowed to earn, it obviously makes no sense. It is a mistake that arises from extending the equation behond the range of the independent and dependent variables actually observed.

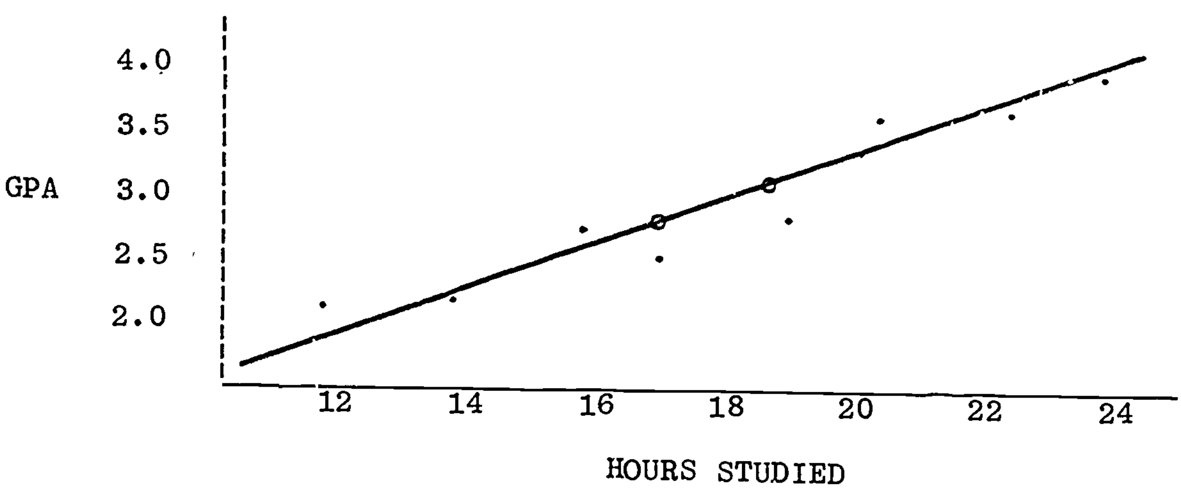
C. Drawing the Regression Line

The equation can be used to draw the regression line on a scatterplot. Select any two values of the independent variable and calculate the "predicted" value of the dependent variable for these two independent values. Enter the "predicted" points on the scatterplot and draw a straight line through those two points from the Y axis through the range of the dependent variable. For the scatterplot on page 49 of hours and GPA, the steps would be as follows:

For student A, the value of 19 hours computes to a predicted score of 3.00; for student J, the value of 17 hours computes to a predicted score of 2.72. Enter these two points on the scatterplot and draw the regression line.

Figure 6.3

SCATTERPLOT WITH REGRESSION LINE DRAWN



D. Pearson's R^2

Pearson's R^2 is a correlation statistic appropriate to interval data. Person's R^2 varies from 0.00 to +1.00. It is an important statistic when doing regression because it generally measures how close the points are to the computed regression line.

When interpreting the meaning of R^2 as it applies to the relationship between two or more variables, it is helpful to remember the following:

1. Since a variable is defined as any characteristic of people or groups which varies (see p. 19), then R^2 is a statistic which helps to explain characteristics that influence variations in the dependent variable.

For example, suppose you wish to understand what the key factors may be in predicting income. You may suspect that income levels differ according to sex, race, and education but you do not know for certain. If you calculate the R^2 for each of these independent variables against the dependent variable, income, you will know whether any, some, or all of these variables explain variations in income.

If the R^2 for sex and income is .67, you can say that 67% of the variation in one's income is explained by whether they are male or female. If the R^2 for education is .23, you can say that only 23% of the variation in income can be explained by education.

If you check the R^2 for all these variables in income, you may find that the R^2 is .73, explaining 73% of the variation in the dependent variable.

Since there is little increase between sex and income and all of the variables and income, you may conclude that more in-depth statistical research should be done to check for explanations as to why more is not explained by these variables. You may find that the results change dramatically when you look at differences between R^2 for women and men, different races, or varying sources of education.

2. When doing analysis for R^2 or regression equations, use the following format to plug the names of your variables into:

R^2: The R^2 for _______________ is __________. This means that _____% of the variation in _______________ (the dependent variable) is explained or predicted by _______________ (the independent variable).

$Y = a + bX$: A one-unit increase in _______________ (X) causes _______________ (Y) to increase/decrease by ________ (b) units. X is the independent variable and Y is the dependent variable.

FORMULA

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

$\hat{Y}$ = The predicted value of each case of the dependent variable

$\bar{Y}$ = The mean of the dependent variable

In order to calculate R^2, follow these three next steps. (Refer to the calculations from the hours and GPA data in the table below.):

Step 1: Calculate the difference between each predicted value and the mean of the dependent variable. Square each difference and then sum the squared differences: = 2.66.

Step 2: Calculate the difference between each actual value and the mean of the dependent variable. Square each difference and then sum the squared differences: = 2.86.

Step 3: Divide the number obtained in step 1 by the number in step 2. The resulting number is R^2: 2.66/2.86 = .93.

The very large R^2 in this example (.93 out of a possible 1.00) statistically confirms the visual impression we have when viewing Figure 6.3, namely that the points all fall very close to the calculated regression line.

Student	Hours Studied (X)	X^2	GPA (Y)	(XY)	Predicted GPA-Mean	Predicted GPA-Mean^2	GPA-Mean	GPA-Mean^2
A	19	361.	2.9	55.1	0.016	0.000256	-0.03	0.0009
B	14	196.	2.3	32.2	-0.684	0.467256	-0.63	0.3969
C	21	441.	3.5	73.5	[illegible]96	0.087616	0.57	0.3249
D	12	144.	2.1	25.2	[illegible]64	0.929296	-0.83	0.6889
E	24	576.	3.9	91.2	[illegible]16	0.512656	0.87	0.7569
F	19	361.	2.7	51.3	0.016	0.000256	-0.23	0.0529
G	20	400.	3.1	62.	0.156	0.024336	0.17	0.0289
H	16	256.	2.7	43.2	-0.404	0.163216	-0.23	0.0529
I	23	529.	3.6	82.8	0.576	0.331776	0.67	0.4489
J	17	289.	2.6	44.2	-0.264	0.069696	-0.33	0.1089

Pearson's R^2 can generally be interpreted the same as the del statistic, as discussed on page 37.

Exercise VI

Select at least 10 cases for which you have two interval variables that measure some policy condition, and about which you have enough information to identify which variable is the dependent and which is the independent. Using these data, construct a scatterplot. Calculate the regression equation and draw the regression line. Compute Pearson's R^2. In one paragraph, interpret the results. (Refer to the guidelines in Exercise I.)

POLICY STUDIES ASSOCIATES

Policy Studies Associates ~~established in 1976 to strengthen learning resources that will help students develop policy anaylsis skills and techniques and apply ~~ese to important public issues. Toward this end, PSA is publishing a series of learning packages in policy studies designed especially for undergraduate use—the *Policy Sciences Series*, which emphasizes techniques in policy analysis. The series is a continuation of learning packages originally sponsored by the Public Affairs Program of the Maxwell Graduate School of Citizenshi~ and Public Affairs at Syracuse University

Policy Studies Associates in a cooperative non-profit undertaking of a small group of faculty members and others concerned with improving the quality of education on public policy issues in schools, colleges, and universities. The Associates at present include:

William Coplin, Maxwell School, Syracuse University
Michael O'Leary, Maxwell School, Syracuse University
Ward Morehouse, Council on International and Public Affairs, Inc. and Columbia University

PSA is an operating program of the Council on International and Public Affairs. Current and forthcoming titles in the *Policy Sciences Series* are listed below:

PS-8 *Forecasting with Dynamic Systems*
Michael K. O'Leary

PS-9 *The Good Federalism Game: Participant's Manual for a Simulation of Intergovernmental Relations*
Roger M. C~ven and George G. Wolohojian

PS-11 *Designs for Evaluating Social Programs*
Lawrence P. Clark

PS-12 *An Introduction to Surveys and Interviews*
Lawrence P. Clark

PS-13 *The Analysis of Policy Arguments*
Ralph S. Hambrick, Jr. and William P. Snyder

PS-19 *Library Research for the Analysis of Public Policy*
Renee S. Captor

PS-21 *An Introduction to Computor Analysis in the Social Sciences and Business Using SAS*
Josephine M. LaPlante

PS-22 *An Introduction to Benefit-Cost Analysis for Evaluating Public Expenditure Alteraatives*
Josephine M. LaPlante and Taylor R. Durham

PS-23 *Political Analysis through the Prince System*
William D. Coplin and Michael K. O'Leary

PS-24 *Introduction to Political Risk Anaylsis*
William D. Coplin and Michael K. O'Leary

PS-25 *The Formulation and Presentation of Alternatives for Public Programs*
Leonard I. Ruchelman

PS-26 *Introduction to Statistics: Revised Edition*

PS-27 *Database Management for Public Policy*