# CHAPTER 8

# TABULAR REPRESENTATION OF UNIVARIATE AND BIVARIATE QUALITATIVE VARIABLES: SINGLE AND CROSS-CLASSIFICATION SCHEMES

Data analysis usually, but not always, implies statistical treatment of the collected observations. Hence, we will commence this chapter with the nature and use of statistics in social research. Statistics is an applied branch of mathematics much like medicine is an application of the principles of physics, chemistry, anatomy, physiology, and biology. This practical function is geared to provide meaningful guidelines for assembling, describing, and inferring salient numerical characteristics of data distributions. The term "statistics" itself has at least three different meanings.[1] Statistics, as conveyed in the mass media, often refers to large collections of data as illustrated by the generic notion of vital statistics. Vital statistics are numerical data indicating such socially important actuarial phenomena as birth, abortion, mortality, marriage, divorce, and other kinds of frequency counts. This use of the term as a plural noun is entirely legitimate but is not consistent with the manner in which research methodologists generally employ it. Another definition of statistics, this time a singular noun, is the body of methods or techniques that researchers use in making descriptive, associative, and inferential statements about data. This second usage, the one employed here, conceptualizes statistics as a method, tool, or ideology that is employed in data analysis. The third manner in which the term is used is when a numerical index from a sample of observations is computed.

It is no exaggeration to say that the world in which we live is increasingly a statistical/numerical one. In fact, every day we are virtually inundated with statistical information. If we are to stay afloat we must become "numerate." Some of this information is easily digested and understood but much of it appears magical, arcane, or esoteric and is frequently misinterpreted or erroneously construed. Since most persons will occupy the consumer/interpreter rather than producer role of statistics, it is important to have a grasp of certain fundamental principles of statistical reasoning and interpretation.

This book is geared to help you in your role of methodological and statistical consumer. The kinds of questions asked about numerical observations will be advanced along with the reasons why these queries are considered important. Furthermore, although computational formulae will be provided, equal treatment will be given to how the resulting statistical indices are interpreted--what they mean as well as their routine calculations.

The data analysis chapters will cover tabular representation of univariate and bivariate qualitative variables, statistical computations for univariate, bivariate, and multi-variate data, hypothesis testing and parameter estimation.

After data have been collected the first step is to assemble and organize them into a meaningful configuration. Usually this entails a data reduction process whereby the collection of statistical observations is whittled down into a form that enables the "consumer" to grasp salient patterns. By summarizing and condensing the information, one may more easily understand and "digest" the observations. One of the most popular and practical ways for reducing raw data into meaningful patterns is to present them in tabular form. The simplest and most basic table is constructed for a univariate distribution. Univariate refers to the response distribution of a single ("uni")variable. Usually it is a dependent variable and is the center of the researcher's attention.

The Univariate Table.

Table 8.1 is the simplest tabular arrangement and is widely-used in mass media reports. Notice that it contains the categories of the variable of interest. In this instance the variable is said to be a natural dichotomy meaning it contains two "natural" sub-classifications. Secondly, it contains a column heading which specifies the number or frequency of cases in the respective categories of the variable. Thirdly, it contains a title that indicates the population or sample from which the frequency counts stemmed, the characteristic whose distribution is being reported, and the source.

TABLE 8.1

BIRTHS IN UNITED STATES

| Sex | Number |
|---|---|
| Male | 1,608,326 |
| Female | 1,528,639 |

Source: Division of Vital Statistics, National Center for Health Statistics, Public Health Service.

TABLE 8.2

LEADING CAUSES OF DEATH
United States: Estimates

| Category | Number | Proportion | Percent |
|---|---|---|---|
| Diseases of heart and blood vessels | 1,062,160 | .538 | 53.8 |
| Cancer | 351,294 | .178 | 17.8 |
| Accidents | 116,297 | .059 | 5.9 |
| Pneumonia/Influenza | 62,599 | .032 | 3.2 |
| Diabetes* | 38,225 | .019 | 1.9 |
| All other causes | 342,428 | .174 | 17.4 |
| Total | 1,973,003 | 1.000 | 100.0 |

*Deaths from suicide and cirrhosis of the liver exceed those for diabetes for persons under age 65.

Source: National Center for Health Statistics, U.S. Public Health Service, HEW and The American Heart Association.

(Table 8.1 here)

At this juncture you probably confront one of the puzzling features of data presentation. Absolute numbers--frequencies--like those in Table 8.1 are not only difficult to interpret in vacuo but are also of limited value in summarizing the essence of the data. In order to meaningfully interpret statistical data, it becomes necessary to make relevant comparisons of the numbers. Choosing an appropriate criterion for comparing numbers is a fundamental research problem and often more complex than it superficially appears.[2]

What does it mean to compare two (or more) numbers?[3] Ordinarily two numbers are compared by: (1) subtracting one from the other and/or (2) dividing one number by the other. When numbers are compared via subtraction the larger of the two can be determined as well as the magnitude of the difference between them. Using this comparison technique for the characteristics displayed above, we conclude that there were 79,687 (1,608,326 - 1,528,639 = 79,687) more male births than female births, The second comparison operation entails dividing one number by the other. In doing so, we arrive at the relative size of the number and are provided with information conveying the quantity of one number relative to (or per) some standard quantity of the other. Ordinarily, division is preferable since absolute frequencies or absolute differences depend on how many observations were initially collected. The problems inherent in the subtraction procedure generally are overcome when we compare numbers via division. Statisticians have devised various techniques, all based on the comparison of numbers by division, which are useful for making sense out of absolute numerical values.

## The Basic Problem in Comparing Numbers

Since the operation of division entails dividing one number, technically the numerator, by another number, technically the denominator, the basic problem resides in deciding which number is dividend and which number is divisor.

Let us call the characteristic of interest the criterion variable and the number whose influence we want to remove the norming variable (or base). The four procedures--1)ratios, 2) proportions, 3)percents, and 4)rates--for comparing numbers via division represent different solutions to the problem of choosing a meaningful base.[4]

Suppose we want to compare the number of males in Table 8.1 with the number of males in another group. If the total sizes of the two groups were significantly different, say 95 vs. 1,608,326, it would be virtually meaningless to focus upon the absolute frequencies in the respective categories of the two distributions. A purposeful comparison of the number of males in two different groups could be achieved by comparing the number of males in each group with the number of females or with the total number of persons involved. The first type of comparison is subsumed under the concept of ratios and the second under the concept of proportions (and their extension to percentages).

## Ratios

If we compare, through division, the number of elements in one category with the number of elements in a second category we derive a statistical index termed a ratio. A ratio provides us with a numerical value expressing the number of elements in one category (i.e., numerator) to the number of elements in another category (i.e., denominator). In short, a ratio expresses the number of elements in the numerator per elements in the denominator. One of the most familiar demographic indexes is the sex ratio which expresses the number of male(s) per female(s). To illustrate, (see Table 8.1) the census reported that there were 1,608,326 male births and 1,528,639 female births. The following formula may be used for computing the sex ratio (at birth) in the U.S.

$$\text{sex ratio} = \frac{\text{number of males}}{\text{number of females}} = \frac{1,608,326}{1,528,639} = 1.05$$

Hence, the sex ratio was 1.05. This is interpreted to mean that for every one female there was 1.05 male(s).

Two questions arise. Firstly, would it be permissible to alter the numerator and denominator, that is, let the number of females be the numerator and the number of males be the denominator? Could the equation be?

$$\text{sex ratio} = \frac{\text{number of females}}{\text{number of male}} = \frac{1,528,639}{1,608,326} = .95$$

The answer is clearly yes, although the former formula is the standard computing procedure found in the population literature and the latter interpretation would be different. In the latter formula, the numerical value would be interpreted as so many females per male. Specifically, there were .95 female(s) per one male. The two formulas will generate mirror image numerical values.

The second qualification revolves around the confusing and near meaninglessness of the computed sex ratio. Since fractions of persons (in this case) are ludicrous, statisticians generally multiply the computed ratio by some <u>numerical base</u> such as 100, 1000, 10,000, or 100,000 and then interpret the figure in relation to the numerical base. The sex ratio usually takes the form:

$$\text{sex ratio} = \frac{\text{number of males}}{\text{number of females}} (100) = \frac{1,608,326}{1,528,639} = 105$$

This figure means there are 105 males for every 100 females. Equivalently, we could say that there are 95 females for every 100 males. Ratios then are sensible statistics used to reduce large amounts of data to single summary statistical indices.

The generic formula for computing ratios is:

$$\text{ratio} = \frac{\text{number of elements in one category}}{\text{number of elements in another category}} \times \text{numerical base}$$

Other commonly utilized ratios include marriage, divorce, mortality, and birth indices. In psychology the well-known intelligence quotient is a ratio obtained by dividing mental age by chronological age. Similarly, the anthropological measure termed the cephalic index is a ratio generated by dividing the width of the cranium by the length of the cranium. One limitation of the ratio occurs when the denominator is very small. For example, the U.S. Military Academy (West Point) recently admitted female cadets. If the total enrollment at the academy consisted of 2000 males and 10 females the sex ratio would be 200. Such numbers loom virtually meaningless for interpretative purposes.

## Proportions and Percents

Whereas the ratio compares the number of elements in one category with the number of elements in another category the proportion compares the number of elements in one category with the total number of elements in all categories. Take special note of the changing denominator. A proportion is a comparison between a part (or parts) and the whole, or a subset and a set. The proportion of male births in the U.S. population would be determined by dividing the number of male births by the total number of births. In formula form:

$$\text{Proportion of males} = \frac{\text{number of males}}{\text{total \# of births}} = \frac{1,608,326}{3,136,965} = .513$$

Since a proportion is, by definition, a part of a whole the proportions added together will produce a sum of 1.00. To illustrate, if we compute the proportion of female births in the U.S. population we have:

$$\text{proportion of females} = \frac{\text{number of females}}{\text{total \# of births}} = \frac{1,528,639}{3,136,965} = .487$$

Now adding the two proportions, .513 + .487 = 1.00. The same logic holds true for characteristics with more than two categories. For example, in Table 8.2 the proportions are .538 + .178 + .059 + .032 + .019 + .174 = 1.00.

(Tables 8.2 and 8.3 here)

Like ratios, proportions are intuitively more meaningful when multiplied by a numerical base. Conventionally, proportions are multiplied by 100 so that we can speak of so many males per 100 people (with a ratio we spoke of so many males per 100 females). When a proportion is multiplied by 100 it becomes the familiar and commonly used statistic called a percent. Percents when summed yield 100. For example, the percent of male and female births is 51.3 and 48.7 respectively. Adding the two together-- 51.3 + 48.7--equals 100 percent. The same holds true when dealing with polychotomies. Using Table 8.2 as an illustration, 53.8 + 17.8 + 5.9 + 3.2 + 1.9 + 17.4 = 100%.

Proportions and percents solve the small base problem often confronted when working with ratios and, because of that, are generally desirable. In most cases the denominator will be larger than the numerator since it contains the total number of elements in all categories and can never be smaller than the numerator.

The advantage of proportions and percents is that they enable one to meaningfully compare elements in two or more distributions and are called relative frequencies , relative percentages or simply, relative numbers.[5]

Simple computational formulae for proportions and percents are, respectively:

$$\text{Proportion } (p) = \frac{\text{number of elements in a category}}{\text{total number of elements}}$$

$$\text{Percent } (P) = \frac{\text{number of elements in a category}}{\text{total number of elements}} \times 100$$

A final caveat regarding proportions and percents is in order. Since the denominator is inclusive (i.e., includes all elements) dramatic impressions can

TABLE 8.3

BIRTHS IN UNITED STATES

| Sex | Number | Proportion | Percent |
|---|---|---|---|
| Male | 1,608,326 | .513 | 51.3 |
| Female | 1,528,639 | .487 | 48.7 |
| Total | 3,136,965 | 1.000 | 100.0 |

TABLE 8.4

ATTITUDE TOWARD PARDONING BY POLITICAL PARTY PREFERENCE

| | | Party Preference (X) | | | | |
|---|---|---|---|---|---|---|
| | | Republican | | Democrat | | Totals |
| | | N | % | N | % | |
| Attitude (Y) | Wrong | 156 | 36.6 | 602 | 72.8 | 758 |
| | Right | 270 | 63.4 | 225 | 27.2 | 495 |
| | Totals | 426 | 100.0 | 827 | 100.0 | 1253 |

be conveyed as well as misleading ones when the base is small. If we randomly selected ten individuals and asked them for their presidential preference and nine endorsed the Democratic candidate we could report that 90% of the sample favored so and so. The pitfall revolves around the idiosyncrasies of sampling which frequently occur with small samples. (As an aside, statistical theorems like the law of large numbers and the central limit theorem, apply only when the number of samples and/or the size of the samples are large).[6] In short, one must be wary of statistics, whether they be proportions, percents, ratios, or others, based upon small samples. It is recommended practice to report the total sample size and some statisticians have gone so far as to suggest that percents and proportions only be computed when the total size is at least thirty.[7]

Rates

The consumer of statistical information often confronts the reporting of events, happenings, occurrences and facsimiles. For example, typical events that appear in the mass media include marriages, births, divorces, crimes, etc. Before verbalizing the problem one may compute and report the divorce rate to be 436.35, 4.51, and 11.28 per thousand (actual number = 970,000). The bugaboo in these apparently different statistical indices is analogous to the basic problem alluded to earlier, namely, selecting an appropriate denominator upon which to base the statistic. The base for the respective statistics presented above are, respectively, the number of marriages in a given year (2,223,000), the total population (215,000,000), and the number of females 14 years of age and older (86,000,000). The dilemma resides in the selection of an appropriate norming variable. The number of divorces that can occur in a given calendar year is a function of the potential number that could occur or, saying it differently, the number of elements that are exposed to the risk of such an event. Clearly, then, non-married persons

cannot be potential divorce casualities; nor can those below the legal age of marriage.

It is worthy of reiterating at this juncture that one of the easiest ways to "lie" with statistics emanates from the lack of insight regarding the potential number of "risky" cases. It should be apparent from the illustrations above that erroneous, false, and downright incorrect impressions can be advanced from what might appear to be objective indicators. Therefore, in your reading you should adopt a critical posture whereby you attempt to carefully scrutinize the implications of a reported statistic. As far as rates are concerned what is most critical is to compare the number of events in which you are interested (e.g., divorce, marriage, births) with the number of <u>potential</u> events that could occur. This leads us to forward the following generic formula for computing rates:[8]

$$\text{rates} = \frac{\text{number of cases for criterion variable}}{\text{number of cases for norming variable}} \times \text{numerical base}$$

The <u>criterion variable</u> is simply the occurrences in which you are interested in (e.g., divorces, deaths, births) whereas the <u>norming variable</u>, the more problematic of the two, is the number of occurrences at risk. The numerical base is a number multiplied by the ratio or rate of occurrences that facilitates the interpretation of the statistic. Since events occur during some time period, we usually make explicit the time frame being considered, and say that so many events per 1000 (assuming that is the numerical base) actually took place.

While we have acknowledged the major problem in computing rates, we have failed to mention that frequently the norming variable--the number of potential events that could occur--is not easy to choose for practical as well as theoretical reasons. For example, although there is no panacea to

this quandry rates may be divided into two type; 1) crude rates, and 2) refined rates. Crude rates, as one might surmise, are those that are "quick and dirty", meaning little or no concern is taken with the pool of potential cases that could occur. To illustrate, the crude birth rate is computed as follows:

$$\text{birth rate} = \frac{\text{number of births in a given year}}{\text{total population at midyear}}$$

Take special notice of the denominator. It includes the entire population, many of which (e.g., males, non-sexually mature females, females beyond menopause) are impossible of experiencing the event. The same thing occurs with the crude divorce rate which, once again, includes persons not at risk. One must be cautious of interpreting such figures because of their hazardous and misleading nature. In their favor, however, is that they permit at least a modicum of insight into some social phenomenon. See Vignette 9.1 .

Refined rates are generally better and more meaningful since deliberate steps are taken to determine the number of cases at risk. Even with refined rates a creative imagination is important. Among refined divorce rates we find two scions: 1) norming on the number of marriages in the year in which the number of divorces occurred, and 2) married females 14 years of age or older. Even though refined rates are more revealing, one must adopt a critical posture in interpreting them. Moreover, their calculation entails more work.

The computation of rates serves at least two important research functions. Firstly, it enables one to infer occurrences in the same population over a period of time. To illustrate, we may wish to know trends in marriages and divorces over the past quarter of a century. Rates allow us to make induc-

Vignette 9.1

# How to Measure Divorce Rates

By Morris Armstrong
eHow Contributor

There are a number of ways to calculate the divorce rate. The methodology for each may be perfectly correct, and yet the answer may still be misleading. When people ask what the divorce rate is, are they inquiring about all divorces, first time divorces, divorces where the marriage was long lived or divorces where the marriage was short lived? The categories seem almost endless.

## Instructions

1. Divide the total number of divorces that occur in a given time period by the number of marriages that have occurred in the same period. This is a simple calculation that is easy to perform, but it has little statistical merit. This method takes two unrelated events and attempts to make them appear relevant.

   In 2004, there were 1,126,358 divorces and 2,311,998 marriages. Using this method results in a divorce rate for that year of 48.72 percent.

2. Divide the number of divorces in a time period by the entire population and multiply the result by 1000. This method is called the "Crude Divorce Rate." In 2004 the population of the United States was 296 million people. There were 1,126,358 divorces that year, which would make the divorce rate 3.8 per thousand using this method. One major problem with this number is that the total population, which is used as the divisor, will include people who are not yet married, and others who will never get married.

3. Divide the number of divorces in a given time period by the total number of marriages at the end of the time period and multiply by 1000. This result is the "Refined Divorce Rate." It attempts to filter out a the large segment of the population that may create a distorted picture when the Crude Divorce Rate is used. The Refined Divorce Rate for 2004 was 18.34 per thousand.

tions about changing social trends. Secondly, it enables us to compare different populations in terms of those experiencing a certain event. For example, holding constant the year under scrutiny, we may determine the age categories in which most divorces occur relative to other age categories.[9]

Percentage Change

It is possible to compute the change over time in the occurrence of some event by dividing the difference between the two frequency counts by the frequency that existed at the former time period. For example, according to the Bureau of Census, the population of metropolitan Columbus, Ohio, was 1,017,847 in 1975 and 916,228 in 1970. What is the percentage change in the population of the standard metropolitan statistical area (SMSA) of Columbus? To compute the percentage change the population at the earlier time period (i.e., 1970) is subtracted from the more recent time period (i.e., 1975) and the difference is divided by the population at the earlier time. Hence,

$$\frac{1,017,847 - 916,228}{916,228} = \frac{101,619}{916,228} = .11$$

If the decimal is multiplied by 100 it becomes possible to express the change in the population in percentage terms. Substantively, Columbus grew 11% between the 1970 and 1975 census.

Generically the percentage statistic can be computed using the formula:

$$\text{percentage change} = \frac{n_2 - n_1}{n_1} \times 100$$

where: $n_2$ = number of events at time two

$n_1$ = number of events at time one

100 = numerical base

Many statistical measures are important or become particularly so when they are compared with other data. For example, suppose a corporation is

trying to decide between building a new plant in New York City or Columbus, Ohio. They want to be assured that the area where the site will be located is growing. All other things equal, the percentage change computation provides some insight. New York's population in 1975 was 9,973,577 and 11,571,899 in 1970. Using the above formula, we compute the percentage change in New York to be -13.8. This means there was a 14% decline in the New York SMSA between 1970 and 1975.

The Bivariate Table

Just as it is possible to display the distribution of a single variable in tabular form, it is quite common to find the distribution of two variables presented in a table. The bivariate table is one of the most frequently found ones in both popular forums as well as in more scholarly journals. The term "bivariate" simply means two ("bi") variables ("variates") are under examination. More specifically, a bivariate table displays the joint occurrences of the categories of two distinct variables. The objective of this section is twofold: 1) to elaborate the structure of a bivariate table and 2) to provide some guidelines for reading such a table.

Table 8.4 is the most rudimentary bivariate table. It is called a 2 x 2 table or a four-fold table because each variable has two subdivisions (each variable has been dichotomized) and the body of the table contains four frequency counts in the "cells" of the table. More specifically, political preference has been dichotomized into Republican and Democratic categories; attitude toward pardon has been divided into two parts, namely, "wrong" and "right." In general a table like this one is referred to as an r x c table where r = the number of rows and c = the number of columns. A table with three rows and six columns would be called a 3 x 6 table while a table with five rows and four columns would be a 5 x 4 table.

(Table 8.4 here)

The structure of the bivariate table is so important that expounding upon its key ingredients is worthwhile. The important components of the table, also called a contingency table because one variable is presumed to be contingent upon (i.e., contingent in this context means dependent upon) the other variable, are the: 1) heading, 2) stub, 3) independent (X) variable, 4) dependent (Y) variable, 5) cells (or cell frequencies), 6) marginals (or marginal totals), and 7) grand total.

Heading. The heading of the table is the variable listed at the top, along with its categories. In Table 8.4 the variable is party preference and the categories of the variable are Republican and Democrat. It is common practice, although not universally practiced, to place the presumed independent variable (and its sub-divisions) along the horizontal dimension of the tabular display. Since political preference is likely to "cause" one's attitudes toward the pardoning this convention has been followed.

Stub. The stub of the table is the variable located at the side, along with its categories. In Table 8.4 the variable is attitude (toward pardoning) and the categories of the variable are "wrong" and "right". Once again, it is conventional to locate the dependent variable along the vertical dimension of the tabular display. It is sensible to assume that attitudes towards pardoning are dependent upon one's political preference rather than to argue the converse.

Cells. The body of the table, that is, the intersection of a particular row category with a particular column category, contains cells and the entries are called cell frequencies. Notice what these cells represent. Take the cell frequency at the intersection of row one and column one ($n_{11}$). The number there is 156. This is interpreted as the number in the total sample who were both Republican and held a negative attitude toward the par-

doning. In a similar vein the 225 cell frequency (at the intersection of row two and column two--$n_{22}$) represents the number of cases who were both Democratic and held a positive attitude toward the pardoning.

The same logic applies to the other two frequencies ($n_{12}$ = 602 and $n_{21}$ = 270) in the body of the table. Because of their interpretation, the cell frequencies are more descriptively termed joint or conditional frequencies because they show the number of times that values of one variable occurredgiven that a particular value of the other variable occured.

Marginals. The marginals or marginal totals appear at the bottom of the columns and the side of the rows. Hence, in Table 8.4 there are two column marginal totals (426=Republicans and 827=Democrats) and the two row marginal totals (758="wrong" and 495= "right"). Later on you will see that marginal totals are used in the computation of various statistics.

Grand Total. The grand total is simply the total number of cases with which the researcher is dealing. In Table 8.4 it is 1253 and can be obtained by summing the column or row marginal totals (e.g., 426+827, 758+495=1253) or the four individual cell frequencies (e.g., 156+602+270+225=1253). The grand total, like the marginals, is also used in computing certain statistical indices.

Constructing a Bivariate Table

Before offering aids for making sense of Table 8.4 let us briefly indicate how a bivariate table is constructed. There are two ways in which a bivariate table can arise.[10] Firstly, we can imagine that two sets of measurements (e.g., political party preference and attitude toward the pardoning have been collected from each and every member of the sample. Through tabulating the responses into two categories simultaneously we arrive at a bivariate data arrangement. Following through on this we might visualize having collected the following information:

| Subject Number | Political Party Preference | Attitude Toward the Pardoning |
|---|---|---|
| 1 | Republican | Right |
| | Democrat | Wrong |
| 2 | Republican | Wrong |
| 3 | | |
| ... | | |
| 1253 | Democrat | Wrong |

Having collected data in this manner, we could then simultaneously classify each case into the appropriate cell of a 2 x 2 contingency table. For illustrative purposes, we could classify the first three subjects indicated above as follows:

| | Republican | Democrat |
|---|---|---|
| Wrong | 1 | 1 |
| Right | 1 | |

We would continue this operation until the entire sample of 1253 was classified.

A second way one can visualize the construction of a bivariate table is to imagine that we have the distribution of responses (e.g., attitudes toward the pardoning) in two (or more) populations (e.g., one sample of Republicans and another sample of Democrats). Following through on this second approach, we may imagine the following information:

| Republican Number | Attitude Toward the Pardoning |
|---|---|
| 1 | Wrong |
| 2 | Right |
| 3 | Wrong |
| ... | . |
| 426 | . |

Notice that the sample consists entirely of Republicans, of which there were 426, and we tally the attitudes of this social category toward the action, The prodedure would look something like this (for the first three cases):

Attitudes of Republicans Toward the Pardoning

| | |
|---|---|
| Wrong | 2 |
| Right | 1 |

Similarly, we can imagine a sample comprised entirely of Democrats along with their attitudes toward the issue and tally the responses as follows:

| Democrat Number | Attitude Toward the Pardoning |
|---|---|
| 1 | Wrong |
| 2 | Right |
| 3 | Wrong |
| .... | . |
| | . |
| 827 | . |

Locating these respective frequencies we would have:

Attitudes of Democrats Toward the Pardoning

| | |
|---|---|
| Wrong | 2 |
| Right | 1 |

This second approach to the construction of a bivariate distribution should look familiar. You can imagine two separate univariate distributions, one distribution for the attitudes of Republicans and the other distribution for the attitudes of Democrats. The two univariate distributions can then be juxtapositioned to arrive at a bivariate table that can be compared and contrasted.

Before advancing some guidelines for comparison and contrasting purposes it is important to be cognizant that these two different approaches--1) two sets of observations collected from every sampling element and 2) distri-

bution of responses in two discrete samples--are different ways of comprehending the building of a bivariate table. A final point is that in the former approach both sets of measurements--political preference and attitude--are conceived of as variables whereas in the latter approach the political preference dimension is a constant and the attitude is a variable whose distribution is being examined. More importantly, though, is that these two approaches lead to the same bivariate data set.

Guidelines for Interpreting a Bivariate Table

Let us return to Table 8.4. Attitude toward the pardoning of Nixon is construed as the dependent variable and political preference is the independent variable. The analyst asks whether people's political party preference influences their attitude toward Ford's action regarding Nixon.

The table presents the number of cases (e.g., 270 Republicans said Ford was right while 602 Democrats said he was wrong). It should be obvious that a simple comparison of the cell frequencies is not adequate because, among other things, there are nearly twice (827÷426 = 1.94) as many Democrats in the sample as Republicans. To say that 446 (602 - 156 =446) more Democrats than Republicans (comparison of numbers by subtraction) said the pardoning was wrong, is not particularly helpful in the analysis. Moreover, because 45 (270 - 225 = 45) more Republicans said the action was right does not, prima facie mean that Republicans and Democrats' attitudes are similar. By visualizing the Republican and Democratic samples separately note that the total number, what we've called the marginal totals, in each aggregate is quite dissimilar. There are 426 (or 34% of the total sample) Republicans in comparison to 827 (or 66% of the total sample) Democrats. Therefore, one of the recommended steps for adequate comparison is: Compute percentages to adjust or control for the unequal sample sizes.

Although this procedure will, when qualified, control for the inequality,

the actual computations await the answer to a much more serious query: How are the percentages to be computed? The formula for a percentage is: P = fi/N (100), but what is to be the N or denominator? There are actually three possible ways that percentages in a contingency table could be computed:[11] 1) the grand total, 1253 in this case, could be used as the base (or denominator); 2) the row totals, 758 and 495 in the present could be used; and 3) the column totals, 426 and 827 in the present case. In order to decide which is most appropriate we must return to the original query: Does political party preference affect attitudes toward the pardoning? The question clearly contains an implicit cause-effect connection between the variables. The causal or independent variable is party preference and the effect or dependent variable is "attitude". With this framework in mind we may qualify the percentage rule advanced above to read.

Compute percentages in the direction of the independent variable.[12] This means we use the column totals (since the independent variable and its categories are located across the top or heading) as the denominator for calculating percentages. Hence, for the four cell frequencies in Table 8.4, we compute the percentages as follows:

$n_{11}$ = 156/426 x 100 = 36.6 ⟩ 100%
$n_{12}$ = 602/827 x 100 = 72.8 ⟩ 100%
$n_{21}$ = 270/426 x 100 = 63.4
$n_{22}$ = 225/827 x 100 = 27.2

Notice that the percentages for each univariate ($n_{11} + n_{21}$ and $n_{12} + n_{22}$) distribution sum to 100. To illustrate, for the Republicans the percent that responded affirmatively is 63.4 while the percent that respond negatively

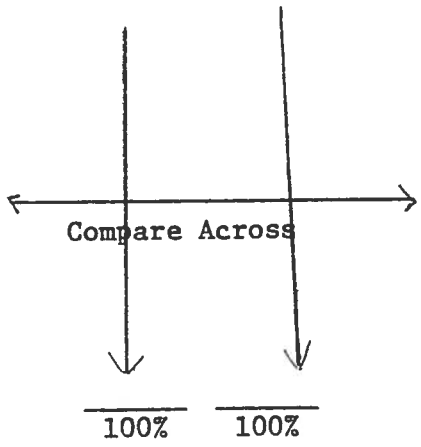
is 36.6. These two percentages when added yield 100 (63.4 + 36.6 = 100). Likewise the respective Democratic percentages are 27.2 and 72.8 which, when summated, total 100.

Referring to the original question, here is what the table reveals. Nearly 73% (72.8) of the Democrats regarded action as "wrong" in comparison to about 37% (36.6) of the Republicans who responded that way. Or, to look at the other side of the "coin", about 63% (63.4) of the Republicans vis-a-vis 27% (27.2%) of the Democrats viewed the pardon as "right." Substantively speaking, Democrats were twice as likely to see action as "wrong" while more than twice as many Republicans view his judicial decision as "right".

Note how the comparisons are made. If the table has been percentaged up and down (i.e., using the column totals as the denominator) the comparison is made across categories of the independent variable. This observation leads us to advance another rule of thumb for table interpretation:

> Compare percentages in the opposite direction that the percentages have been computed.[13]

Diagrammatically,

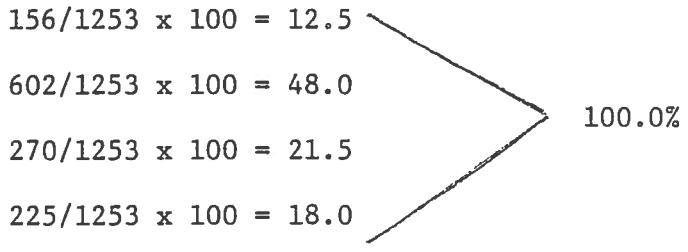



We compare the percentage of Republicans who said "wrong" to the percentage of Democrats who said "wrong" (36.6% to 72.8%) and the percentage of Republicans who said "right" to the percentage of Democrats who said right

(63.4% to 27.2%). One can also compare each univariate distribution separately. For example, 36.6% of the Republicans believed the decision was wrong in contrast to 63.4% who deemed it was right; 72.8% of the Democrats judged it to be wrong while 27.2% felt it was right. In short, Republicans were considerably more likely to see action as "correct" and Democrats were prone to see it as "incorrect."

A comparison of percentages computed in the same direction (e.g., 36.6 and 63.4 and 72.8 and 27.2), while interesting, does not answer the original query. Comparing the percentages computed in the same direction would not tell us whether party preference affects attitude toward the pardoning. That information per se does not tell us whether attitudes are influenced by party affiliation. Generally speaking, we compare the percentage of persons in the various independent variable categories who select a given response category of the dependent variable.

Since the novice is likely to go astray in such matters as percentaging a table correctly, it is heuristic to compute the percentages the other two ways, that is, (1) using grand total and (2) row totals, as the denominator. Let us compute the percentage of cases in each cell using the grand total as the denominator and then see what interpretation would follow suit. There are 1253 total subjects. If we divide each joint frequency by 1253 we have:




Note that the percentages total 100. What we have done is to express each cell frequency as a percentage of the total. With this information such things as nearly 50% (48.0%) of all cases in the table are Democrats who deem the decision to be wrong and less than a quarter (21.5%) of all obser-

vations are Republicans who judge the decision as correct. Ordinarily this information is not of interest and, more importantly, it would not provide us with statistical data to answer the initial research question. All this information does is to express each cell frequency as a percentage of the total.

The final mode of percentaging would be to use the row totals as the base of the percentage. The row total for those who believe the decision to be "wrong" is 758 while the row total for those who believe the decision to be "right" is 495. Speaking in percentage terms, of the entire sample 60.5% thought it was "wrong" while 39.5% judged it to be "right." If a ratio between the two quantities is computed it can be said that about 1.53 persons deemed the decision wrong for every one person who deemed it right:

$$758/495 = 1.53$$

or

$$60.5/39.5 = 1.53$$

If we percentage the cells using the row totals we have:

$$156/758 \times 100 = 20.6$$
$$602/758 \times 100 = 79.4 \quad \Big\rangle \quad 100.0\%$$

and

$$270/495 \times 100 = 54.5$$
$$225/495 \times 100 = 45.5 \quad \Big\rangle \quad 100.0\%$$

Notice that the percentages for each row sum to 100 (20.6 + 79.4 = 100.0 and 54.5 + 45.5 = 100). Looking at these latter four figures, here is what they tell us. About 21% (20.6%) of those who said the decision was wrong were Republican in contrast to nearly 80% (79.4) of the Democrats who responded the same. Likewise, slightly more than half (54.5%) of those who believed the decision was right were Republicans in comparison to less than half (45.5% of the Democrats who responded identically). Computing

percentages in this latter way does not answer the question: Does party preference affect attitudes toward the pardon? Instead it tells us if attitudes affect party preference which does not make much sense in the present context.

## Interpreting a More Complex Bivariate Table

Bivariate tables come in many different r x c dimensions. However, no matter how elaborate the rows or columns in a contingency table the logic of analysis is the same, albeit somewhat more tedious. For illustrative purposes consider the 2 x 3 table in Table 8.5. It contains two variables (making it bivariate) with the dependent variable, belief in an afterlife, dichotomized and the independent variable, religious affiliation, trichotomized. The question we wish to address is: Does religious affiliation affect one's attitude toward an afterlife? A bivariate table such as this contains an enormous amount of information. Some of the questions which are meaningful to ask along with the process by which the answers can be arrived at will now be explicated.

(Table 8.5 here)

1. How was the table constructed? These data were collected by NORC (National Opinion Research Center at the University of Chicago) and reflect two sets of measurements collected from each sample element. Recall that this table could also be visualized as having stemmed from three samples--Protestant, Catholic, and Jew--and represent univariate distributions.

2. How many cases are there in the entire sample? By adding the six cell frequencies (i.e., 715 + 272 + 7 + 144 + 84 + 31) we arrive at the figure of 1253.

3. Of the total sample of 1253 how many were Protestant, Catholic, and Jew? Since the categories of the independent variable are located along

TABLE 8.5

ATTITUDE TOWARD AN AFTERLIFE BY RELIGIOUS AFFILIATION

| | | Religious Affiliation (X) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Protestant | | Catholic | | Jew | | Totals |
| | | N | % | N | % | N | % | |
| Belief in Afterlife (Y) | Yes | 715 | 83.2 | 272 | 76.4 | 7 | 18.4 | 994 |
| | No | 144 | 16.8 | 84 | 23.6 | 31 | 81.6 | 259 |
| | Totals | 859 | 100.0 | 356 | 100.0 | 38 | 100.0 | 1253 |

the heading we answer this query by summing separately each of the vertically placed univariate distributions. Therefore, the number of Protestants is 715 + 144 = 859; the number of Catholics is 272 + 84 = 356; and the number of Jews is 7 + 31 = 38. As a check note that the column marginal totals of 859 + 356 + 38 = 1253; the original total sample size.

4. Since absolute frequencies are not particularly valuable, we ask what percent of the total sample is Protestant, Catholic, and Jew? We divide the number in each of the respective religious groupings by the total sample size:

$$859/1253 \times 100 = 68.5\%$$

$$356/1253 \times 100 = 28.4\%$$

$$38/1253 \times 100 = 3.0\%$$

Consequently, we can see that nearly 70% of the entire sample claim a Protestant affiliation, nearly 30% endorse Catholicism, and there is a very small percentage (3%) of Jews.

5. Of the total sample how many and what percent believe in an afterlife? By summing all those who responded affirmatively (i.e., 715 + 272 + 7) we discover that 994 of the total believe in an afterlife. To calculate the percentage of the entire sample who believe in a hereafter we divide this figure by the total sample size or

$$994/1253 \times 100 = 79.3\%$$

6. Of the total sample how many and what percent do not believe in an afterlife? Adding across categories of the dependent variable, we have 144 + 84 + 31 or 259. To determine the percent of the sample who do not believe in an afterlife we divide this number by the entire sample size or

$$259/1253 \times 100 = 20.7\%$$

7. What is the ratio of believers in an afterlife to non-believers? Using the appropriate formula ratios we have

994/259 = 3.84

or

79.3%/20.7% = 3.83

Substantively this means there were nearly four times as many believers as non-believers. Saying it differently but with the same thrust, for every non-believer there were 3.84 or nearly four believers.

8. Now to our original query, does religious affiliation affect belief in an afterlife? To answer this question we first have to decide which variable is independent and which is dependent and then percentage the table using the independent variable totals as the denominator and finally compare across categories of the independent variable(in the opposite direction that the percentages were computed).

Since the causal variable is located across the heading and the effect variable along the stub the column totals are used as the denominator of the percentages. There are 859 Protestants (715 + 144 = 859) and we ask, what percent of them believe and do not believe in an afterlife? Since 715 believe and 144 do not we divide these numbers by the total number of Protestants. Therefore:

715/859 x 100 = 83.2%

144/859 x 100 = 16.8%

Similarly, there are 356 Catholics and the percent who believe in an afterlife is 76.4 (272/356 x 100 = 76.4) in contrast to 23.6 (84/356 x 100 = 23.6) who do not. Finally, of the 38 Jews, 18.4% (7/38 x 100 = 18.4) believe in a hereafter while 81.6% (31/38 x 100 = 81.6) do not.

By this time you can see that comparing numbers by subtraction is not very meaningful since the sample sizes are grossly unequal. To say that 708 (715 - 7 = 708) more Protestants believe in an afterlife than Jews tells us virtually nothing since there are about 23 times as many Protestants as Jews in our sample (859/38 = 22.6).

Having percentaged the table correctly we may now directly answer the question by comparing the percentages across categories of the dependent variable. Whereas 83% of the Protestants believe in an afterlife, 76% and 18% of the Catholics and Jews believe accordingly. Whereas the percentage of Protestant and Catholic believers is fairly similar, both of these statistics are significantly larger than that for Jews. Comparing the "no" responses across the divisions of the independent variable provides one with a "mirror" image, namely, significantly more Jews than either Protestants or Catholics do not believe in an afterlife.

## SUMMARY

Since data analysis usually, but not always, implies statistical treatment of the collected observations, this chapter commences with a discussion of the role of statistics in social research. Since the function of statistics resides in assembling, describing, and inferring salient statistical properties of data distributions we considered how data are organized in tables. The manner in which univariate and bivariate qualitative variables can be represented in tabular form was highlighted. In doing this both single and cross-classification schemes were addressed.

Once data are meaningfully assembled it is incumbent on the researcher to make sense out of them. Absolute numbers--frequencies--are not only difficult to interpret in vacuo but are also of limited value in summarizing the essence of the data. In order to meaningfully interpret statistical information it is necessary to make relevant comparisons of numbers. When numbers are compared it generally means subtracting one from the other (comparison by subtraction) or dividing one number by another (comparison by division).

Because of the problems inherent in comparison by subtraction, most of the exposition was devoted to techniques for comparison by division. Among

the most salient procedures for comparing numbers by division are through computing ratios, proportions, percents, and rates. Each of the statistical indices was computed and interpreted for some real data. Additionally, when two pieces of information are collected from the same case it is possible to compute the percentage change index. This, too, was calculated and interpreted.

Moving from univariate to bivariate data we considered the structure of contingency tables. The key ingredients of such tables include the: heading, stub, independent variable (X), dependent variable (Y), cells (or cell frequencies), marginals (or marginal totals), and the grand total. Two different ways of building a contingency table were also identified. When data are in contingency table formats and the analyst desires to make meaningful comparisons the cell frequencies are percentaged in the direction of the independent variable and compared across categories of the dependent variable. Specific illustrations of this procedure illuminated this process.

Finally, in contingency tables other than 2 x 2, an enormous amount of information can be extracted. Some of the ways this information can be extracted were explored.

## IMPORTANT CONCEPTS DISCUSSED IN THIS CHAPTER

| | | | |
|---|---|---|---|
| Statistics | Sex Ratio | Crude Rates | Dependent Variable |
| Univariate Table | Proportions | Percentage Change | Cells (cell frequencies) |
| Absolute Numbers | Percents | Bivariate Table | |
| Comparing Numbers | Relative Numbers | Tabulation | Marginals (marginal totals) |
| Criterion Variable | Rates | Heading | Grand Total |
| Norming Variable | Birth Rates | Stub | Rules for Percentagir tables |
| Ratios | Refined Rates | Independent Variable | |