9
Descriptive and Multivariate Statistics

Jamie Price
Donald W. Chamberlayne[*]

Statistics is the science of collecting and organizing data and then drawing conclusions based on data. There are essentially three types of statistics: descriptive, multivariate, and inferential. Descriptive statistics summarize large amounts of information in an efficient and easily understood manner. Multivariate statistics allow comparisons among factors by isolating the effect of one factor or variable from others that may distort conclusions. Inferential statistics (covered in Chapter 13) suggest statements about a population based on a sample drawn from that population.

This chapter will familiarize readers with the principles of descriptive and multivariate statistics. Students and practitioners need to know the basic foundations of research and statistics so that they are better producers as well as consumers of research and police data. A fundamental understanding of descriptive and multivariate statistics is essential to job performance and evaluation.

Analysts should be able to perform a number of simple, though powerful, analyses to describe data and reach conclusions based on these data. Many people are intimidated by mathematics and statistics. However, the role of statistics is very important with regard to reports, publications, policy, and the general understanding of information that we process every day.

[*] Editor's Note: Jamie Price and Donald Chamberlayne worked independently on their respective sections of this chapter. Mr. Price wrote the material on descriptive statistics; Dr. Chamberlayne's contribution begins with the heading "Multivariate Statistics."

Descriptive Statistics

In the following sections, readers will learn the following major goals: (a) summarizing large and small data sets, (b) examining the integrity of large and small data sets, (c) determining which statistics best portray the data, (d) comparing more than one variable to others, and ultimately, (e) applying descriptive statistics to problem solving and data driven decision-making.

Levels of Measurement

One concept that applies to both research and statistics is "levels of measurement." Indeed, measurement is the process of assigning numbers or labels to units of analysis or items under study. In other words, numbers are assigned to the "who" or "what" that are being studied. Numbers are assigned to make the data amenable to statistical analysis. There are four levels of measurement: nominal, ordinal, interval, and ratio. These distinctions are important because how data are analyzed depends on how data were collected. Each level conveys a different amount of information.

Nominal

The **nominal** level of measurement is the process of classifying data into categories. It is the lowest level of measurement and all categories must be *exhaustive*, thus covering all observations that may exist. In addition, the categories must be *mutually exclusive*: each observation can only be classified in one way. Nominal measures merely provide names or labels for distinguishing observations. For example, we could classify respondents to a

survey by race or gender. Each respondent's race could be coded as "African American," "Asian," "Caucasian," or "Other." For gender, each respondent is coded as "Male" or "Female." Each respondent falls into only one classification and there is an appropriate category for each respondent according to their race and gender. Though we could assign a numeric code to each category (e.g., "1" for male and "2" for female), the code is still nominal data—there is no logical way to rank-order or perform calculations with it.

Ordinal

The **ordinal** level of measurement consists of the characteristics of the nominal level—exhaustive and mutually exclusive—but also exhibits a degree of difference between the categories on a scale. This degree of difference indicates order or ranking between categories. The categories are ordered in some way, but the actual distance between these orderings would not have any meaning. Examples are opinion of police, crime seriousness, levels of fear. Response scales such as "good, better, best," "agree, neutral, disagree," or "very unlikely, unlikely, undecided, likely, very likely" are common ordinal scale measures.

Interval

The **interval** level of measurement consists of all the characteristics of nominal and ordinal levels of measurements. In addition, the interval level assumes that all the items on a scale have equal units or intervals of measurement between them. In contrast to ordinal scales, the distance between categories would have meaning. There are logical distances between categories expressed in meaningful standard intervals. Examples of interval measurement would be temperature readings and IQ.

Ratio

In addition to all the characteristics of the previous three levels of measurement, the **ratio** level of measurement contains a true zero point. A true zero point allows for measuring the total absence of the concept under measure. Income, weight, time, and age are examples of ratio level measurements.

Table 9-1 summarizes the different information conveyed by the four levels of measurement.

There are three implications regarding the level of measurement. First, ratio is the highest level of measurement because it contains all the characteristics of the other three. Second, researchers and practitioners should seek the highest level of measurement possible, within reasonable time, effort, and cost constraints of the study. Lower levels of measurement cannot be converted to higher levels of measurement, but higher-level measurement can be converted to a lower level. Third, and most important, the statistical technique to be applied will determine the level of measurement needed. Specific analytic techniques require minimum levels of measurement.

	Levels Of Measurement			
Characteristic	Nom.	Ord.	Int.	Rat.
Exclusive & Exhaustive	X	X	X	X
Rank Order		X	X	X
Equal Intervals			X	X
Absolute Zero Point				X

Table 9-1: A comparison of levels of measurement

Ideally, many concepts can be measured on different levels of measurement depending on how data are collected. For example, the concept "age" could be measured using the following response sets for each level of measurement:

Nominal: "Young" or "Old"
Ordinal: "0–6," "7–13," "14–20," "30+"
Interval: "1–20," "21–40," "41–60," "61+"
Ratio: 0,1,2,3,4,5,6,7,8,9,10,11,12,13,etc.

Based on the implications stated earlier, several points can be demonstrated regarding the different levels of measurement pertaining to the "age" example above.

1) Regardless of a person's age, there is a response set that applies to each individual (exhaustive).
2) Including the subjective nominal level, a person's age falls into only one classification (exclusive).
3) Although the ordinal level contains a zero, the intervals between each category are not equal; thus they cannot be interval or ratio.
4) The interval method does not begin with zero; thus it cannot be ratio.
5) The ratio level is based on the assumption that individuals under study could be less than one year of age (infant less than 365 days for example).
6) Working highest to lowest, a person's age collected on the ratio level could be converted into lower levels of measurement. Although, the interval level could not be converted using the ordinal level above, a new ordinal level could be created from the interval level such as 0 to 20, 21 to 40, 41 to 60, 61 and over. All three levels of measurement, ordinal, interval, and ratio could be converted to nominal.
7) Working lowest to highest, the data could not be converted into higher levels. For example, if a person's age is classified as only young or old, no determination could be made as to whether that individual was 0 to 6, 7 to 13, and so on. The last category of the ordinal level, 30+, could not be converted into the more precise intervals such as 21 to 40, 41 to 60 and so on. Moreover the interval does not begin at zero. (For this reason, it is common for the interval and ratio levels to be treated as one level).
8) The age should be collected on the level necessary to the techniques the analyst is going to apply.

Data collection is very important. Likewise, two areas of great interest in descriptive statistics are describing the "average case" and describing how the "average case" compares to all the cases as a whole. These two areas complement each other. Indeed, measures of central tendency describe the characteristics of the average case and measures of dispersion indicate just how typical or average this case is.

Distributions

Before discussing measures of central tendency and dispersion, analysts should understand the distribution of data sets. Some data sets can be very large and cumbersome. Data should be described in a manner that is easily understood. One of the first steps in analyzing data is to construct a **frequency distribution**, which lists the number or frequency of scores or labels for each individual case. The frequency distribution allows for a basic description of the data set and for graphical representation.

Assume that we have collected data on the number of prior arrests for a group of 39 offenders and that we are interested in how many of them have more than three prior arrests. The **raw numbers** of prior arrests are hypothetically as follows:

```
1, 3, 7, 2, 5, 2, 4, 2, 1, 5, 6, 2, 3, 1, 4,
2, 3, 4, 3, 3, 5, 6, 1, 2, 3, 4, 6, 1, 3, 2,
2, 1, 3, 2, 4, 3, 12, 3, 1
```

If the data set is small, a visual inspection of the raw data set may reveal the answer relating to three or more arrests. However, if the data set is large it will probably be too difficult or impractical. For simplicity, we will use a small data set but assume it requires summation. Table 9-2 tabulates the number of prior arrests for 39 individuals.

Prior Arrests of Current Offenders

x (number of prior arrests)	*f* (frequency)	*f* x
12	1	12
7	1	7
6	3	18
5	3	15
4	5	20
3	10	30
2	9	18
1	7	7
	N = 39	$\sum fx = 127$

Table 9-2: A frequency distribution

From the frequency distribution above, the x column indicates that the data set of 39 offenders have between one and 12 prior arrests. Using the frequency column, it is easily determined that 13 offenders had more than three prior arrests. And among them, the group of offenders had 127 prior arrests.

In constructing a frequency distribution, the first step is to create an array or set of numerical values arranged in order from highest to lowest. This array is the first column or *x*. The second column is the frequency column (*f*). This column indicates how many times that particular score occurred in the data set. The third column, *f*x, indicates the total number of values by multiplying the first and second columns.

Once a frequency distribution has been created, analysts may want to further condense it to allow for easier, more meaningful graphing.

The **range** is a summary statistic that provides limited information but allows for condensing a frequency distribution. The range is the highest score minus the lowest score (H – L). From the example above, the range is: 12 – 1 = 11.

To condense the distribution, the range is used to create groupings of information into class intervals.

$$\text{Class Interval } i = \frac{\text{Range}}{N \text{ of desired intervals}}$$

Assume we want to group the number of prior arrests into four intervals:

$$i = \frac{12-1}{4} = \frac{11}{4} = 2.75$$

* 2.75 is rounded to 3

Thus, each interval will have three cases. Grouping the prior arrest data, the new **grouped distribution** would be:

Interval	*f* (frequency)
10–12	1
7–9	1
4–6	11
1–3	26
	n = 39

Table 9-3: A grouped frequency distribution

Although the data have been condensed and the distribution has changed, the nature of the data remains the same. Progressing from raw numbers to frequency distributions to grouped distributions allows for large data sets to be more easily managed and analyzed.

Data in frequency distributions or grouped distributions lend themselves to easy graphing. The purpose of **charts** and **graphs** is to portray the distribution of data for a quick and meaningful understanding. The following charts and graphs were created using Microsoft

Excel; however, other graphing software is available.

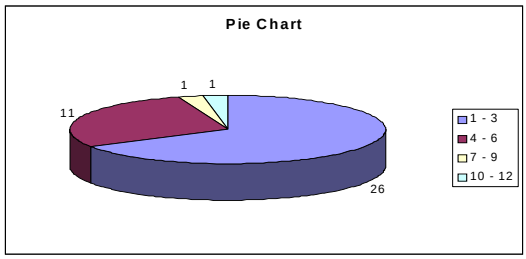

Figure 9-1: Pie chart illustrating the number of prior arrests, based on grouped frequency distribution in Table 9-3.

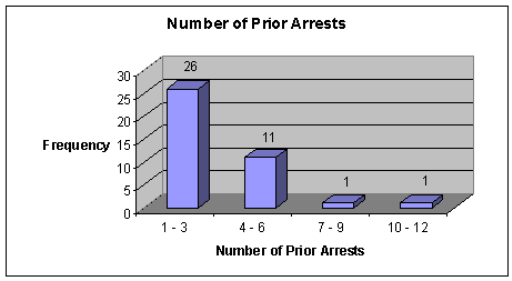

Figure 9-2: Column chart illustrating the number of prior arrests, based on grouped frequency distribution in Table 9-3.

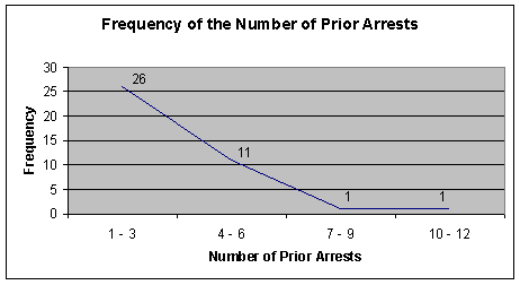

Figure 9-3: Line chart illustrating the number of prior arrests, based on grouped frequency distribution in Table 9-3.

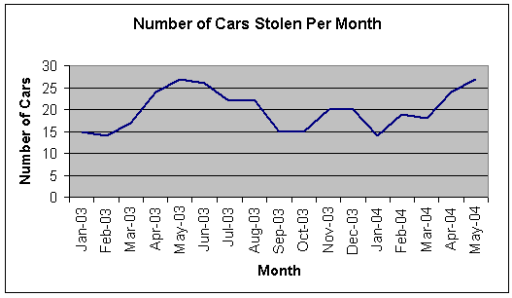

Figure 9-4: Line chart illustrating the number of stolen cars per month.

Graphing provides not only greater visibility of the distribution but also quick and simple interpretations. Pie charts (Figure 9-1) can present percentages or counts while comparing the relative size of various segments. Column graphs (Figure 9-2) also present percentages or counts while comparing the relative size of various segments. Histograms are special column graphs with continuous data. Line graphs (Figure 9-3) reveal plotted values and may reveal **skewness** or **trends** (Figure 9-4) over time.

In addition to frequency distributions and grouped frequency distributions, **percentages** and **cumulative percentages** can be calculated. Using the grouped frequency distribution for prior arrests, two additional columns have been added to indicated the percent of cases pertaining to each interval or segment as well as the cumulative percentage of each segment in combination with others.

Interval	*F* (freq.)	Percent %	Cumulative %
10–12	1	2.6	2.6
7–9	1	2.6	5.2
4–6	11	28.2	33.4
1–3	26	66.7	100.0
	N = 39	100.0*	

Table 9-4: A frequency distribution with percentages and cumulative percentages

The percent column is calculated by dividing the frequency of each interval by the total number of cases. For example, for the interval of ten to twelve prior arrests, one divided by thirty-nine equals 2.6% or approximately three percent of all cases. The cumulative percentage is determined by adding the percent column for each class interval. By adding the first two intervals, the cumulative percentage would be 5.2 percent (2.6 + 2.6). Adding the second and third intervals, the cumulative percentage would be 33.4 percent (28.2 + 5.2). And so on. One question that often comes up is: what should you do with **missing data**? It is important because depending on how you treat