

Chapter 1

Statistical Inference I: Descriptive Statistics

This chapter discusses and examines methods and techniques for summarizing and interpreting data. It starts by examining numerical descriptive measures. These measures, commonly known as point estimators, enable inferences about a population by estimating the value of an unknown population parameter using a single value (or point). This chapter also overviews graphical representations of data. Relative to graphical methods, numerical methods provide precise and objectively determined values that can easily be manipulated, interpreted and compared. They permit a more careful analysis of the data than more general impressions conveyed by graphical summaries. This is important when the data represent a sample from which inferences must be made concerning the entire population.

While this chapter concentrates on the most basic and fundamental issues of statistical analyses, there are countless thorough introductory statistical textbooks that can provide the interested reader with greater detail. For example, Aczel (1993) and Keller and Warrack (1997) provide detailed descriptions and examples of descriptive statistics and graphical techniques. Tuckey (1977) is the classical reference on exploratory data analysis and graphical techniques. For readers interested in the properties of estimators (Section 1.7), Gujarati (1992) and Baltagi (1998) are excellent and fairly mathematically rigorous references.

1.1 Measures of Relative Standing

A set of numerical observations can be ordered from smallest to largest magnitude. This ordering allows the boundaries of the data to be defined and allows comparison of the relative position of specific observations. If an observation is in the 90th percentile, for instance, then

90% of the observations have a lower magnitude. Consider the usefulness of percentile rank in terms of a nationally administered test such as the scholastic aptitude test (SAT) or graduate record exam (GRE). An individual's score on the test is compared with the scores of all people who took the test at the same time, and the relative position within the group is defined in terms of a percentile. If, for example, the 80th percentile of GRE scores is 660, this means that 80% of the sample who took the test scored below 660 and 20% scored 660 or better. A percentile is defined as that value below which lies $P\%$ of the numbers in the remaining sample. For sufficiently large samples, the position of the P^{th} percentile is given by $(n+1)P/100$, where n is the sample size.

Quartiles are the percentage points that separate the data into quarters: first quarter, below which lies $\frac{1}{4}$ of the data, making it the 25th percentile; second quarter, or 50th percentile, below which lies half the data; third quarter, or 75th percentile point. The 25th percentile is often referred to as the lower or first quartile, the 50th percentile as the median or middle quartile, and the 75th percentile as the upper or third quartile. Finally, the interquartile range, a measure of the spread of the data, is defined as the difference between the first and third quartiles.

1.2 Measures of Central Tendency

Quartiles and percentiles are measures of the relative positions of points within a given data set. The median constitutes a useful point because it lies in the center of the data, with half of the data points lying above it and half below. Thus, the median constitutes a measure of the centrality of the observations.

Despite the existence of the median, by far the most popular and useful measure of central tendency is the arithmetic mean, or more succinctly the mean. The sample mean or expectation is a statistical term that describes the central tendency, or average, of a sample of

observations, and varies across samples. The mean of a sample of measurements x_1, x_2, \dots, x_n is defined as

$$MEAN(X) = E[X] = \bar{X} = \frac{\sum_{i=1}^n x_i}{n}, \quad (1.1)$$

where, n is the size of the sample.

When an entire population constitutes the set to be examined, the sample mean \bar{X} is replaced by μ , the population mean. Unlike the sample mean, the population mean is constant. The formula for the population mean is

$$\mu = \frac{\sum_{i=1}^N x_i}{N}. \quad (1.2)$$

where, N is the number of observations in the entire population.

The mode (or modes because it is possible to have more than one of them) of a set of observations is the value that occurs most frequently, or the most commonly occurring outcome, and strictly applies to discrete variables (nominal and ordinal scale variables) as well as count data. Probabilistically, it is the most likely outcome in the sample; it has occurred more than any other value.

It is useful to examine the advantages and disadvantages of each the three measures of central tendency. The mean uses and summarizes all of the information in the data, is a single numerical measure, and has some desirable mathematical properties that make it useful in many statistical inference and modeling applications. The median, in contrast, is the central most (center) point of ranked data. When computing the median, the exact location of data points on the number line are not considered, only their relative standing with respect to the central observation is required. Herein lies the major advantage of the median; it is resistant to extreme

observations or outliers in the data. The mean is, overall, the most frequently used measure of central tendency; in cases, however, where the data contain numerous outlying observations the median may serve as a more reliable measure of central tendency.

If the sample data are measured on the interval or ratio scale, then all three measures of centrality, mean, median and mode make sense, provided that the level of measurement precision does not preclude the determination of a mode. If data are symmetric and if the distribution of the observations has only one mode, then the mode, the median, and the mean are all approximately equal (the relative positions of the three measures in cases of asymmetric distributions is discussed in Section 1.4). Finally, if the data are qualitative (measured on the nominal or ordinal scales), using the mean or median is senseless, and the mode must be used. For nominal data, the mode is the category that contains the largest number of observations.

1.3 Measures of Variability

Variability is a statistical term used to describe and quantify the spread or dispersion of data around its center, usually the mean. In most practical situations, knowing the average or expected value of a sample is not sufficient to get an adequate understanding of the data. Sample variability provides a measure of how dispersed the data are with respect to the mean (or other measures of central tendency). Figure 1-1 illustrates two distributions of data, one that is highly dispersed and another that is more tightly packed around the mean.

Figure 1-1: Examples of High and Low Variability Data

There are several useful measures of variability, or dispersion. One measure previously discussed is the interquartile range. Another measure is the range, which is equal to the difference between the largest and the smallest observations in the data. The range and the

interquartile range are measures of the dispersion of a set of observations, with the interquartile range being more resistant to outlying observations. The two most frequently used measures of dispersion are the variance and its square root, the standard deviation.

The variance and the standard deviation are more useful than the range because, like the mean, they use the information contained in all the observations. The variance of a set of observations, or sample variance, is the average squared deviation of the individual observations from the mean, and varies across samples. The sample variance is usually used as an estimate of the population variance, and is given by

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}. \quad (1.3)$$

When a collection of observations constitute an entire population, the variance is denoted by σ^2 . Unlike the sample variance, the population variance is constant, and is given by

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}, \quad (1.4)$$

where \bar{X} in Equation 1.3 is replaced by μ .

Because calculation of the variance involves squaring the original measurements, the measurement units of the variance are the square of the original measurement units. While variance is a useful measure of the relative variability of two sets of measurements, it is often preferable to express variability in the same units as the original measurements. Such a measure is obtained by taking the square root of the variance, yielding the standard deviation. The formulas for the sample and population standard deviations are given respectively as

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}} \quad (1.5)$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad (1.6)$$

Consistent with previous results, the sample standard deviation S^2 is a random variable, while the population standard deviation σ^2 is a constant.

A mathematical theorem attributed to Chebyshev establishes a general rule which states that at least $(1-1/k^2)$ of all observations in a sample or population will lie within k standard deviations of the mean, where k is not necessarily an integer. For the approximately bell-shaped normal distribution of observations, an empirical rule-of-thumb suggests that the following approximate percentage of measurements will fall within 1, 2, or 3 standard deviations of the mean. These intervals are given as

$(\bar{X} - s, \bar{X} + s)$	contains approximately 68% of the measurements,
$(\bar{X} - 2s, \bar{X} + 2s)$	contains approximately 95% of the measurements, and
$(\bar{X} - 3s, \bar{X} + 3s)$	contains approximately 99% of the measurements.

The standard deviation is an absolute measure of dispersion; it does not take into consideration the magnitude of the values in the population or sample. On some occasions, a measure of dispersion that accounts for the magnitudes of the observations (relative measure of dispersion) is needed. The coefficient of variation is such a measure. It provides a relative measure of dispersion, where dispersion is given as a proportion of the mean. For a sample, the Coefficient of Variation (CV) is given as

$$CV = \frac{s}{\bar{X}} \quad (1.7)$$

If, for example, on a certain highway section vehicle speeds were observed with mean $\bar{X} = 45$ mph and standard deviation $s = 15$, then the CV is: $s/\bar{X} = 15/45 = 0.33$. If, on another highway section, the average vehicle speeds is $\bar{X} = 60$ mph and standard deviation $S = 15$, then

the CV is equal to $s/\bar{x} = 15/65 = 0.23$, which is smaller and conveys the information that, relative to average vehicle speeds, the data in the first sample are more variable.

Example 1-1

Using the speed data contained in the “speed data” file that can be downloaded from the publisher’s web site (www.crcpress.com), the basic descriptive statistics are sought for the speed data, regardless of the season, type of road, highway class, and year of observation. Any commercially available software with statistical capabilities can accommodate this type of exercise. Table 1-1 provides descriptive statistics for the speed variable.

The descriptive statistics indicate that the mean speed in the sample collected is 58.86 mph, with little variability in speed observations (s is low at 4.41, while the CV is 0.075). The mean and median are almost equal, indicating that the distribution of the sample of speeds is fairly symmetric. The data set contains more information, such as the year of observation, the season (quarter), the highway class, and whether the observation was in an urban or rural area, which could give a more complete picture of the speed characteristics in this sample. For example, Table 1-2 examines the descriptive statistics for urban versus rural roads.

Interestingly, while some of the descriptive statistics may seem to differ from the pooled sample examined in Table 1-1, it does not appear that the differences between mean speeds and speed variation in urban versus rural Indiana roads is important. Similar types of descriptive statistics could be computed for other categorizations of average vehicle speed.

1.4 Skewness and Kurtosis

Two additional attributes of a frequency distribution that are useful are skeweness and kurtosis. Skeweness is a measure of the degree of asymmetry of a frequency distribution. It is given as the average value of $(x_i - \mu)^3$ over the entire population (this is often called the third moment around the mean, or third central moment, with the variance being the second moment). In general, when the distribution stretches to the right more than it does to the left, it can be said that the distribution is right-skewed, or positively skewed. Similarly, a left-skewed (negatively skewed) distribution is one that stretches asymmetrically to the left (Figure 1-2). When a distribution is right-skewed, the mean is to the right of the median, which in turn is to the right of the mode. The opposite is true for left-skewed distributions. To make the measure $(x_i - \mu)^3$ independent of the units of measurement of the variable, it is divided by σ^3 . This results in the population skeweness parameter often symbolized as γ_1 . The sample estimate of this parameter, (g_1) , is given as

$$g_1 = \frac{m_3}{(m_2 \sqrt{m_2})} \quad (1.8)$$

where:

$$m_3 = \sum_{i=1}^n (x_i - \bar{X})^3 / n$$

$$m_2 = \sum_{i=1}^n (x_i - \bar{X})^2 / n$$

If a sample comes from a population that is normally distributed, then the parameter g_1 is normally distributed with mean 0 and standard deviation $\sqrt{6/n}$.

Figure 1-2: Skeweness of a Distribution

Kurtosis is a measure of the “flatness” (versus peakedness) of a frequency distribution and is shown in Figure 1-3; it is the average value of $(x_i - \bar{x})^4$ divided by s^4 over the entire population or sample. Kurtosis (γ_2) is often called the fourth moment around the mean or fourth central moment. For the normal distribution the parameter γ_2 has a value of 3. If the parameter is larger than 3 there is usually a clustering of points around the mean (leptokurtic distribution), while if the parameter is lower than 3 the curve demonstrates a “flatter” peak than the normal distribution (platykurtic).

Figure 1-3: Kurtosis of a Distribution

The sample kurtosis parameter g_2 , is often reported as standard output of many statistical software packages and is given as

$$g_2 = \gamma_2 - 3 = (m_4 / m_2^2) - 3 \quad (1.9)$$

where

$$m_4 = \sum_{i=1}^n (x_i - \bar{X})^4 / n$$

For most practical purposes, a value of 3 is subtracted from the sample kurtosis parameter so that leptokurtic sample distributions have positive kurtosis parameters, while platykurtic sample distributions have negative kurtosis parameters.

Example 1-2

Revisiting the speed data from example 1-1, there is interest in determining the shape of the distributions for speeds on rural and urban Indiana roads. Results indicate that when all roads are examined together their skeweness parameter is -0.05 , while for rural roads the parameter

has the value of 0.056 and for urban roads the value of -0.37 . It appears that, at least on rural roads, the distribution of speeds is symmetric, while for urban roads the distribution is left-skewed.

While the skeweness parameter is similar for the two types of roads, the kurtosis parameter varies more widely. For rural roads the parameter has a value of 2.51, indicating a distribution close to normal, while for rural urban roads the parameter has a value of 0.26, indicating a relatively flat (platykurtic) distribution.

1.5 Measures of Association

Up to this point the discussion has focused on measures which summarize a large set of raw data. These measures are effective for providing information regarding individual variables. The mean and the standard deviation of a variable, for example, convey useful information regarding the nature of the measurements related to that variable. However, the measures reviewed thus far do not provide information regarding possible relationships between variables. The correlation between two random variables is a measure of the linear relationship between them. The correlation parameter ρ , is a commonly used measure of linear correlation and gives a quantitative measure of how well two variables move together.

The correlation parameter is always in the interval $[-1,1]$. When $\rho = 0$ there is no linear association, meaning that a linear relationship does not exist between the two variables examined. It is possible, however, for two variables to be nonlinearly related and yet have $\rho = 0$. When $\rho > 0$, there is a positive linear relationship between the variables examined, such that when one of the variables increases the other variable also increases, at a rate given by the value of ρ (see Figure 1-4). In the case when $\rho = 1$, there is a “perfect” positively sloped straight line

relationship between two variables. When $\rho < 0$ there is a negative linear relationship between the two variables examined, such that an increase in one variable is associated with a decrease in the value of the other variable, at a rate given by the value of ρ . In the case when $\rho = -1$ there is a perfect negatively sloped straight line relationship between two variables.

Figure 1-4: Positive and Negative Correlations between Two Variables

The concept of correlation stems directly from another measure of association, the covariance. Consider two random variables, X and Y , both normally distributed with population means μ_X and μ_Y , and population standard deviations σ_X and σ_Y respectively. The population and sample covariances between X and Y are defined respectively as follows:

$$COV_p(X, Y) = \frac{\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)}{N} \quad (1.10)$$

$$COV_s(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1} \quad (1.11)$$

As can be seen from Equations (1.10) and (1.11), the covariance of X and Y is the expected value of the product of the deviation of X from its mean and the deviation of Y from its mean. The covariance is positive when the two variables move in the same direction, it is negative when the two variables move in opposite directions, and it is zero when the two variables are not linearly related.

As a measure of association, the covariance suffers from a major drawback. It is usually difficult to interpret the degree of linear association between two variables from the covariance because its magnitude depends on the magnitudes of the standard deviations of X and Y . For

example, suppose that the covariance between two variables is 175. What does this say regarding the relationship between the two variables? The sign, which is positive, tells you that as one increases, the other also generally increases. However, the degree to which the two variables move together cannot be ascertained. But, if the covariance is divided by the standard deviations, a measure that is constrained to the range of values $[-1,1]$, as previously discussed, is obtained. This measure, called the Pearson product-moment correlation parameter, or correlation parameter for short, conveys clear information about the strength of the linear relationship between the two variables. The population ρ and sample r correlation parameter of X and Y are defined respectively as

$$\rho = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \quad (1.12)$$

$$r = \frac{COV(X,Y)}{s_X s_Y} \quad (1.13)$$

where s_X and s_Y are the sample standard deviations.

Example 1-3

Using data contained in the “aviation 1” file, the correlations between annual US revenue passenger enplanements, per capita US gross domestic product, and price per gallon for aviation fuel are examined. After deflating the monetary values by the Consumer Price Index (CPI) to 1977 values, the correlation between enplanements and per capita GDP is 0.94, and the correlation between enplanements and fuel price -0.72 .

These two correlation parameters are not surprising. One would expect that enplanements and economic growth go hand-in-hand, while enplanements and aviation fuel price (often

reflected by changes in fare price) move in opposite directions. But, a word of caution is necessary. The existence of a correlation between two variables does not necessarily mean that one of the variables causes the other. The determination of causality is a difficult question that cannot be directly answered by looking at correlation parameters. To this end, consider the correlation parameter between annual US revenue passenger enplanements and annual ridership of the Tacoma-Pierce Transit System in Washington State. The correlation parameter is considerably high (-0.90) indicating that the two variables move in opposite directions in nearly straight line fashion. Nevertheless, it is safe to say that neither of the variables causes the other or that the two variables are even remotely related. In short, it needs to be stressed that correlation does not imply causation.

Up to this point, the discussion on correlation has focused solely on continuous variables measured on the interval or ratio scale. In some situations however, one or both of the variables may be measured on the ordinal scale. Alternatively, two continuous variables may not satisfy the requirement of approximate normality assumed when using the Pearson product-moment correlation parameter. In such cases the *Spearman rank correlation parameter*, an alternative (non-parametric method), should be used to determine whether a linear relationship exists between two variables.

The Spearman correlation parameter is computed by first ranking the observations of each variable from smallest to largest. Then, the Pearson correlation parameter is applied to the ranks; that is, the Spearman correlation parameter is the usual Pearson correlation parameter applied to the ranks of two variables. The Equation for the Spearman rank correlation parameter is given as

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (1.14)$$

where $d_i, i = 1, \dots, n$, are the differences in the ranks of $x_i, y_i : d_i = R(x_i) - R(y_i)$.

There are additional nonparametric measures of correlation between variables, including Kendall's Tau. Its estimation complexity, at least when compared to Spearman's rank correlation parameter, makes it less popular in practice.

1.6 Properties of Estimators

The sample statistics computed in previous sections such as the sample average \bar{X} , variance S^2 , and standard deviation S and others, are used as estimators of population parameters. In practice population parameters (often called parameters) such as the population mean and variance are unknown constants. In practical applications, the sample average \bar{X} is used as an estimator for the population mean μ_X , the sample variance s^2 for the population variance σ^2 , and so on. These statistics, however, are random variables and, as such, are dependent upon the sample. "Good" statistical estimators of true population parameters satisfy four important properties: unbiasedness, efficiency, consistency, and sufficiency.

Unbiasedness

If there are several estimators of a population parameter, and if one of these estimators of the average coincides with the true value of the unknown parameter, then this estimator is called an unbiased estimator. An estimator is said to be unbiased if its expected value is equal to the true population parameter it is meant to estimate. That is, an estimator, say the sample average \bar{X} , is an unbiased estimator of μ_X if

$$E(\bar{X}) = \mu_X. \quad (1.15)$$

The principal of unbiasedness is illustrated in Figure 1-5. Any systematic deviation of the estimator away from the population parameter is called a bias, and the estimator is called a biased estimator. In general, unbiased estimators are preferred to biased ones.

Figure 1-5: Biased and Unbiased Estimators of a Population's Mean Value μ_X

Efficiency

The property of unbiasedness is not, by itself, adequate, because there can be situations in which two or more parameter estimates are unbiased. In these situations, interest is focused on which of several unbiased estimators is superior. A second desirable property of estimators is efficiency. Efficiency is a relative property in that an estimator is efficient relative to another; it means that an estimator has a smaller variance than an alternative estimator. An estimator with the smaller variance is more efficient. As can be seen from Figure 1-6, both X_1 and \bar{X} are unbiased estimators of μ_1 and μ_2 respectively, however $VAR(\hat{\mu}_1) = \sigma^2$ while $VAR(\hat{\mu}_2) = \sigma^2/n$, yielding a relative efficiency of $\hat{\mu}_2$ relative to $\hat{\mu}_1$ of $1/n$, where n is the sample size.

Figure 1-6: Comparing Efficiencies

In general, the unbiased estimator with minimum variance is preferred to alternative estimators. A lower bound for the variance of any unbiased estimator $\hat{\theta}$ of θ is given by the Crame'r-Rao lower bound that can be written as (Gurajati, 1992)

$$VAR(\hat{\theta}) \geq 1/\{nE(\partial \log f(X;\theta))/\partial \theta\}^2 = -1/\{nE(\partial^2 \log f(X;\theta))/\partial \theta^2\}. \quad (1.16)$$

It should be noted that the Cram er-Rao lower bound is only a sufficient condition for efficiency. Failing to satisfy this condition does not necessarily imply that the estimator is not efficient. Finally, unbiasedness and efficiency hold true for any finite sample n , and when $n \rightarrow \infty$ they become asymptotic properties.

Consistency

A third asymptotic property is that of consistency. An estimator $\hat{\theta}$ is said to be consistent if the probability of being closer to the true value of the parameter it estimates (θ) increases with increasing sample size. Formally, this says that as $n \rightarrow \infty \lim \Pr\left[|\hat{\theta} - \theta| > c\right] = 0$, for any arbitrary constant c . For example, this property indicates that \bar{X} will not differ from μ as $n \rightarrow \infty$. Figure 1-7 graphically depicts the property of consistency, showing the behavior of an estimator X^* of the population mean μ with increasing sample size.

Figure 1-7: The Property of Consistency

It is important to note that a statistical estimator may not be an unbiased estimator; however, it may be a consistent one. In addition, a sufficient condition for an estimator to be consistent is that it is asymptotically unbiased and that its variance tends to zero as $n \rightarrow \infty$ (Hogg and Craig, 1994).

Sufficiency

An estimator is said to be sufficient if it contains all the information in the data about the parameter it estimates. In other words, \bar{X} is sufficient for μ if \bar{X} contains all the information in the sample pertaining to μ .

1.7 Methods of Displaying Data

While many times the different measures described in the previous sections provide much of the information necessary to describe the nature of the data set being examined, it is often useful to utilize graphical techniques for examining data. These techniques provide ways of inspecting data to determine relationships and trends, identify outliers and influential observations, and quickly describe or summarize data sets. Pioneering methods frequently used in graphical and exploratory data analysis stem from the work of Tuckey (1977).

Histograms

Histograms are most frequently used when data are either naturally grouped (gender is a natural grouping for example), or when small sub-groups may be defined to help uncover useful information contained in the data. A histogram is a chart consisting of bars of various heights. The height of each bar is proportional to the frequency of values in the class represented by the bar. As can be seen in Figure 1-8, a histogram is a convenient way of plotting the frequencies of grouped data. In the Figure frequencies on the first (left) y-axis are absolute frequencies, or counts of the number of city transit buses in the State of Indiana belonging to each age group (data were taken from Karlaftis and Sinha (1997)). Data on the second y-axis are relative frequencies, which are simply the count of data points in the class (age group) divided by the total number of data points.

Figure 1-8: Histogram for Bus Ages in the State of Indiana (1996 data)

Histograms are useful for uncovering asymmetries in data and, as such, skeweness and kurtosis are easily identified using histograms.

Ogives

A natural extension of histograms is ogives. Ogives are the cumulative relative frequency graphs. Once an ogive like the one shown in Figure 1-9 is constructed, the approximate proportion of observations that are less than any given value on the horizontal axis can be read directly from the graph. Thus, for example, it can be estimated from Figure 1-9 that the proportion of buses that are less than 6 years old is approximately 60%, while the proportion less than 12 years old is 85%.

Figure 1-9: Ogive for Bus Ages in the State of Indiana

Box Plots

When faced with the problem of summarizing essential information of a data set, a box plot (or box-and-whisker plot) is a pictorial display that is extremely useful. A box plot illustrates how widely dispersed observations are, where the data are centered, and the dispersion of the data. This is accomplished by providing, graphically, five summary measures of the distribution of the data: the largest observation, the upper quartile, the median, the lower quartile, and the smallest observation (Figure 1-10).

Figure 1-10: The Box Plot

The Box plots can be very useful for identifying the central tendency of the data (through the median), identifying the spread of the data (through the Inter-Quartile Range (IQR) and the length of the whiskers), identifying possible skewness of the data (through the position of the

median in the box), identifying possible outliers (points beyond the 1.5(IQR) mark), and for comparing data sets.

Scatter Diagrams

Scatter diagrams are most useful for examining the relationship between two continuous variables. As examples, assume that transportation researchers are interested in the relationship between economic growth and enplanements, or the effect of a fare increase on travel demand. In some cases, when one variable depends (to some degree) on the value of the other variable, then the first variable, the dependent, is plotted on the vertical axis. The pattern of the scatter diagram provides information about the relationship between two variables. A linear relationship is one that can be approximately graphed by a straight line (Figure 1-4). A scatter plot can show a positive correlation, no correlation, and a negative correlation between two variables (section 1.5 and Figure 1-4 analyzed this issue in greater depth). Nonlinear relationships between two variables can also be seen in a scatter diagram, and typically will be revealed as curvilinear. Scatter diagrams are typically used to uncover underlying relationships between variables, which can then be explored in greater depth with more quantitative statistical methods.

Bar and Line Charts

A common graphical method for examining nominal data is a pie chart. The Bureau of economic analysis of the US Department of Commerce in their 1996 Survey of Current Business reported the percentages of the US GDP accounted by various social functions. As shown in Figure 1-11, transportation is a major component of the economy, accounting for nearly 11 percent of Gross Domestic Product in the US. The data are nominal since the “values” of the variable, *major social function*, include the six categories: transportation, housing, food,

education, health care, and other. The pie graph illustrates the proportion of expenditures in each category of major social function.

Figure 1-11: U.S. Gross Domestic Product by Major Social Function (1995) [Source: US DOT (1997)]

The Federal Highway Administration (1997) completed a report for congress that provided information on highway and transit assets, trends in system condition, performance, and finance, and estimated investment requirements from all sources to meet the anticipated demands in both highway travel and transit ridership. One of the interesting findings of the report was the pavement ride quality of the Nation's urban highways as measured by the International Roughness Index. The data are ordinal because the "values" of the variable, pavement roughness, include the five categories: very good, good, fair, mediocre, and poor. This scale, although it resembles the nominal categorization of the previous example, possesses the additional property of natural ordering between the categories (without the increments between the categories being uniform). A reasonable way to describe these data is to count the number of occurrences of each value and then to convert these counts into proportions.

Figure 1-12: Percent Miles of Urban Interstate by Pavement Roughness Category [Source: FHWA (1997)]

Bar charts are a common alternative to pie charts. They graphically represent the frequency (or relative frequency) of each category as a bar rising from the horizontal axis; the height of each bar is proportional to the frequency (or relative frequency) of the corresponding category. Figure 1-13, for example, presents the motor vehicle fatal accidents by posted speed

limit for 1985 and 1995, while Figure 1-14 presents the percent of on-time arrivals for some US Airlines for December, 1997.

Figure 1-13: Motor Vehicle Fatal Accidents by Posted Speed Limit [Source: US DOT (1997)]

Figure 1-14: Percent of On-Time Arrivals for December, 1997 [Source: BTS]

The last graphical technique considered in this section is the line chart. A line chart is obtained by plotting the frequency of a category above the point on the horizontal axis representing that category and then joining the points with straight lines. A line chart is most often used when the categories are points in time (time-series data). Line charts are excellent for uncovering trends of a variable over time. For example, consider Figure 2-15, which represents the evolution of the US Air-Travel market. A line chart is useful for showing the growth in the market over time. Two points of particular interest to the air-travel market, the deregulation of the market and the Gulf War, are marked on the graph.

Figure 1-15: US Revenue Passenger Enplanements 1954-1999 [Source: BTS]

Table 1-1: Descriptive Statistics for Speeds on Indiana Roads

Statistic	Value
N (number of observations)	1296
Mean	58.86
Std. Deviation	4.41
Variance	19.51
CV	0.075
Maximum	72.5
Minimum	32.6
Upper Quartile	61.5
Median	58.5
Lower Quartile	56.4

Table 1-2: Descriptive Statistics for Speeds on Rural versus Urban Indiana Roads

Statistic	Rural Roads	Urban Roads
N (number of observations)	888	408
Mean	58.79	59.0
Std. Deviation	4.60	3.98
Variance	21.19	15.87
CV	0.078	0.067
Maximum	72.5	68.2
Minimum	32.6	44.2
Upper Quartile	60.7	62.2
Median	58.2	59.2
Lower Quartile	56.4	56.15