1 Statistical Terminology and Descriptive Statistics

Like many sciences, statistics involves jargon. Learning the terminology can be challenging because the definition of one term often assumes some working knowledge of other terms, not all of which can be defined at the same time. For example, to explain a boxplot, one must already know the meaning of a mean, median, a quartile, and an outlier.

In this chapter, we introduce some of the key terms in statistics, and in the process overview what is termed descriptive statistics. In some cases, only the definitions are given, leaving a more complete explanation of the phenomena to a later section of the text. This approach can create some confusion for the student, but we feel that it is best to obtain at least a partial appreciation for the language first instead of reiterating a concept and then providing it with a different name.

After developing some mathematical tools, we reconsider the data set and then briefly discuss the concept of a distribution. We then define of parametric and nonparametric statistics and give a brief introduction to the two major phases of statistical processes—estimation and inference. We end by explaining descriptive statistics.

1.1 Statistical and Mathematical Tools

With the two exceptions notes below, all the mathematics needed to understand most of this text may be found in a high school or lower college division algebra course. A rudimentary knowledge of calculus is helpful to understand fully some of the advanced—but optional—sections. The two exceptions to this rule are the summation operator and the product operator.

1.1.1 The Summation Operator

The summation operation is designated by a symbol resembling the Greek upper case letter sigma (Σ). The generic form of the summation operator is

\[ \sum_{\text{start value}}^{\text{end value}} \text{algebraic expression}, \]

and the operator means take the sum of the results of the algebraic expression that follows it. For example,

\[ \sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 \]

and

\[ \sum_{i=2}^{4} (X_i - \bar{X}) = (X_2 - \bar{X}) + (X_3 - \bar{X}) + (X_4 - \bar{X}) . \]

Be careful of mathematical operators in summation notation. Operators placed before the summation sign denote that the summation occurs first and then the mathematical operation is performed. For example,

\[ \log \sum_{i=1}^{3} Y_i = \log(Y_1 + Y_2 + Y_3) . \]
Operators placed after a summation sign are part of the algebraic expression that gets added. Hence, these operations must be performed first and then the results added. For example,

\[ \sum_{i=1}^{3} \log(Y_i) = \log(Y_1) + \log(Y_2) + \log(Y_3). \]

In statistics, the most frequent use of the summation operator is to denote a summation that occurs over the observations in the data set (or over one or more groups within the data set). Here, the notation \( n \) or \( N \) (for Number of observations) is used as the “end value” of summation. For example,

\[ \sum_{i=1}^{N} Y_i \]

implies that one would take the sum all the individual values of variable \( Y \) in the data set. The algebraic indices most often used for the summation operator are the lower case Roman letters \( i, j, k, \) and \( l \). In some statistical texts, the algebraic index, its starting value and its ending value are omitted when the summation is over all observations. For example,

\[ \sum Y \text{ implies } \sum_{i=1}^{N} Y_i. \]

1.1.2 The Product Operator

The product operator behaves similarly to the summation operator except that the mathematical operation is multiplication instead of addition. The product operator is similar to the Greek upper case letter pi (\( \Pi \)). Examples of product notation are

\[ \prod_{i=1}^{3} X_i = X_1 \cdot X_2 \cdot X_3 \]

and

\[ \prod_{i=3}^{5} (p_i - k) = (p_3 - k)(p_4 - k)(p_5 - k). \]

The product operator is most often used in probability theory and problems involving probability.

1.2 Introducing The Data Set

Mathematically, a data set is a two-dimensional matrix of rows and columns, much akin to a modern spreadsheet. The rows of the data set are defined by independent observations and the columns are defined by the attributes or variables measured on those observations. One of the major problems encountered by the student just learning statistics is the recording of data in a lab book or work sheet so that the rows of the data are not independent observations. Usually, this takes the form of entering the same observational unit (i.e., the same cell culture, rat, or person) multiple times in the data set so that a single observation has multiple rows.

For example, consider a study that measures rats at three different time points. It is tempting to enter data similar to that in Table X.X where there are two variables, time
and score. We strongly recommend, however, that one enter the original data by assigning a row to each rat and then entering three variables—the scores for the four time points. One way to insure this type of entry is to always assign each observational unit a unique identifying number or name (*always* a recommended practice). An example of the latter type of entry is given in Table X.X.

**Table X.X. An example of non-independent rows in a data set.**

<table>
<thead>
<tr>
<th>Time</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.3</td>
</tr>
<tr>
<td>2</td>
<td>15.7</td>
</tr>
<tr>
<td>3</td>
<td>9.8</td>
</tr>
<tr>
<td>1</td>
<td>20.2</td>
</tr>
<tr>
<td>2</td>
<td>19.7</td>
</tr>
<tr>
<td>3</td>
<td>13.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>10.8</td>
</tr>
<tr>
<td>2</td>
<td>9.4</td>
</tr>
<tr>
<td>3</td>
<td>7.6</td>
</tr>
</tbody>
</table>

**Table X.X. An example of independent rows in a data set.**

<table>
<thead>
<tr>
<th>Rat ID:</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1J7</td>
<td>12.3</td>
<td>15.7</td>
<td>9.8</td>
</tr>
<tr>
<td>C23K8</td>
<td>20.2</td>
<td>19.7</td>
<td>13.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>D23A</td>
<td>10.8</td>
<td>9.4</td>
<td>7.6</td>
</tr>
</tbody>
</table>

The major danger of having non-independent rows in a data set is that a student may erroneously analyze the data using a statistical procedure that assumes that the rows are independent. The results of this error are unpredictable. In some circumstances, the stat program may be tricked into thinking that there are more observations than there really are. For example, if there were 12 rats in Table X.X, then a correlation between time and score using this data arrangement would be based on 36, not 12, rats. Usually this will result is false positive conclusions (i.e., thinking that there is a significant finding when, in fact, there is not one). In other circumstances (e.g., testing for a mean difference between time 1 and time 2), failing to account for the possibility that variables are correlated can result in false negative findings (i.e., failing to detect a significant finding when, in fact, there is one). Because the vast majority of canned statistical
procedures make the assumption of independent rows, errors are most easily avoided by ensuring that data entry conforms to the rules for Table X.X and not for Table X.X\(^1\).

### 1.2.1 An Example Data Set

Here, we present a real data set that will be used in other sections of this chapter to illustrate statistical terminology and descriptive statistics. Table X.X gives a listing of data based on Bowers et al. (19xx) that examined the effect of the protein kinase C-gamma gene on anxiety in mice. These mice in this study all have the same genetic background but differ in genotypes for the PKC-g gene. Transgenic methods were used to knock out the gene and then cross breeding used to produce three genotypes—those completely deficient in PKC-g (the “--“ genotype in Table 1), the heterozygote with one knockout allele and one normal allele (the “+-“ genotype), and the homozygote with two normal alleles (the “++“ genotype). The other variable in the data set is called “openarm.” It measures the percent of total time in an elevated plus-maze spent in the open (versus the closed) arm of the maze. A low percent of time in the open arm is associated with high levels of anxiety. Note that the actual data set has one and only one observation per row, giving a matrix of 45 rows and two columns. The data in Table X.X are organized differently only to save space.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Openarm</th>
<th>Genotype</th>
<th>Openarm</th>
<th>Genotype</th>
<th>Openarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>- -</td>
<td>15.8</td>
<td>+ -</td>
<td>5.2</td>
<td>++</td>
<td>10.6</td>
</tr>
<tr>
<td>- -</td>
<td>16.5</td>
<td>+ -</td>
<td>8.7</td>
<td>++</td>
<td>6.4</td>
</tr>
<tr>
<td>- -</td>
<td>37.7</td>
<td>+ -</td>
<td>0.0</td>
<td>++</td>
<td>2.7</td>
</tr>
<tr>
<td>- -</td>
<td>28.7</td>
<td>+ -</td>
<td>22.2</td>
<td>++</td>
<td>11.8</td>
</tr>
<tr>
<td>- -</td>
<td>5.8</td>
<td>+ -</td>
<td>5.5</td>
<td>++</td>
<td>0.4</td>
</tr>
<tr>
<td>- -</td>
<td>13.7</td>
<td>+ -</td>
<td>8.4</td>
<td>++</td>
<td>13.9</td>
</tr>
<tr>
<td>- -</td>
<td>19.2</td>
<td>+ -</td>
<td>17.2</td>
<td>++</td>
<td>0.0</td>
</tr>
<tr>
<td>- -</td>
<td>2.5</td>
<td>+ -</td>
<td>11.9</td>
<td>++</td>
<td>16.5</td>
</tr>
<tr>
<td>- -</td>
<td>14.4</td>
<td>+ -</td>
<td>7.6</td>
<td>++</td>
<td>9.2</td>
</tr>
<tr>
<td>- -</td>
<td>25.7</td>
<td>+ -</td>
<td>10.4</td>
<td>++</td>
<td>14.5</td>
</tr>
<tr>
<td>- -</td>
<td>26.9</td>
<td>+ -</td>
<td>7.7</td>
<td>++</td>
<td>11.1</td>
</tr>
<tr>
<td>- -</td>
<td>21.7</td>
<td>+ -</td>
<td>13.4</td>
<td>++</td>
<td>3.5</td>
</tr>
<tr>
<td>- -</td>
<td>15.2</td>
<td>+ -</td>
<td>2.2</td>
<td>++</td>
<td>8.0</td>
</tr>
<tr>
<td>- -</td>
<td>26.5</td>
<td>+ -</td>
<td>9.5</td>
<td>++</td>
<td>20.7</td>
</tr>
<tr>
<td>- -</td>
<td>20.5</td>
<td>+ -</td>
<td>0.0</td>
<td>++</td>
<td>0.0</td>
</tr>
</tbody>
</table>

\(^1\) Having said this, we offer a *caveat*. Some computer procedures (mostly graphical ones but also a few statistical ones) are most easily used when data are arranged as non-independent observations. In such cases, we recommend entering the data as independent observations but then creating a temporary data set that rearranges the data for such procedures.
Variable openarm is measured on a ratio scale. Because the total testing time was the same for each animal, a mouse that spend 20% of the time in the open arm spent twice as time there than a mouse with a score of 10.0. Because this variable is a percent, it is bounded.

Variable genotype is nominal, but this is not the best way to deal with this variable. Once again, ask whether one could present results for the three groups in any order. If the answer is “yes,” then the variable can be treated as categorical. All geneticists, however, would present results for the heterozygotye (genotype “+-”) between the results for the two homozygotes. Hence, instead of treating genotype as categorical, it would be better to create a new variable that would reflect the ordering of the data. There are several ways of doing this—use the ordinal number of 1, 2, and 3 to respectively denote genotypes “- -”, “+ -” and “++”, create a variable for the number of “-” alleles, or one for the number of “+” alleles.

1.2.2 The Distribution of Scores

Much statistical terminology talks of a distribution of scores. We deal with this topic more fully later in X.X Graphing Data, but here introduce the concept of a histogram to provide a brief overview of a distribution.

There are two types of distributions. The first is the empirical distribution that gives a figure and/or summary statistics for the observed data at hand. The second is a theoretical distribution or a mathematical model of the way in which the data are assumed to be distributed in an ideal population. A classic example of a theoretical distribution is the normal curve. We will have much more to say about theoretical distributions later—we just want to introduce the terminology here.

The easiest way to examine an empirical distribution is through graphical methods. Figure X.X illustrates a histogram of the variable openarm in the pkc data. In a histogram, one groups the data into intervals of equal size. In Figure X.X. there are eight intervals, the midpoint of each being presented in the labels of the horizontal axis. For example, the first interval consists of scores between –2.5 to 2.49999\(^2\) (with a midpoint of 0), the second from 2.5 to 7.49999 with a midpoint of 5, and so on.

The vertical axis gives the number of observations with scores that fall into the interval. For example, in the pkc data, six observations fell into the interval of –2.5 to 2.5, seven into the interval between 2.5 and 7.5, and so on. In some computer programs, the vertical axis may also be expressed in terms of the percentage or the proportion of observations that fall into an interval.

\(^2\) Strictly speaking, a number just less than 2.5.
1.2.2.1 Attributes of Distributions

Statisticians speak of several major attributes of distributions. The first of these is location, or the score(s) around which the distribution is centered. The second is spread or the variability or diversity of scores around the location. The third is symmetry or the extent to which the right-hand side of a distribution is a mirror image of the left-hand side. The fourth is peakedness or the extent to which a distribution, given its spread, is flat versus peaked. We discuss these later under descriptive statistics, but in order to understand descriptive stats, it is first necessary to digress into an explanation of parametric and nonparametric statistics.

1.3 Parametric and Nonparametric Statistics:

1.3.1 Populations, Parameters, Samples, and Statistics

Later, we will provide in-depth treatment of the concept of parametric statistics and probability theory. In this section, we merely provide definitions that will be useful in the interim. The branch of statistics known as parametric statistics deals with mathematical equations that apply to a population. The population may be humans, schizophrenics, rats, an inbred strain of mice, or aged nematodes. The equations may include those assumed to represent a distribution of scores (e.g., a normal curve) or those used to predict one or more variables from knowledge of other variables (e.g., regression, dealt with in Chapter XX).

A parameter is defined as an unknown in a mathematical equation that applies to a population. For example, the mathematical equation for a normal curve contains two unknowns—the mean of the curve and the standard deviation of the curve. It is customary to use Greek letters to denote parameters. The most frequently encountered parameters and their customary symbols are the mean (μ), the standard deviation (σ) along with its cousin, the variance (σ²), and the correlation coefficient (ρ).

With rare exception, it is not possible to study all members of a population. Hence, most empirical research begins by selecting a small number of observations (e.g., mice or cell lines) from the population, then gathers data on these observations, and finally uses mathematical inference to generalize to the population. The small number of observations is termed a sample. Most samples are random samples in the sense that the observations are randomly selected from all potential observations in the population. In other cases, observations are selected from the population because they meet some predefined criterion(a). These may be called nonrandom or selected samples.

Within parametric statistics, a statistic is defined as a mathematical estimate of a population parameter derived from sample data. For example, one could calculate the mean reaction time in a sample of schizophrenics and use that number as an estimate of the mean reaction time for the population of schizophrenics. Statistics are usually denoted by Roman letters—e.g., the mean (\( \bar{X} \)), the standard deviation (s) and variance (s²), and the correlation coefficient (r or R). A second type of notation for a statistic is to use the Greek letter for the population parameter with a “hat” (^) over it. For example, \( \hat{\mu} \) denotes an estimate of the population mean and \( \hat{\sigma}^2 \) denotes a statistic that estimates the population variance.
Because statistics are estimates of unknown population parameters, a great deal of parametric statistics deals with error or the discrepancy between an observed statistic and the population parameter. At this point, the astute student may ask—rightfully—how one can deal with error when the numerical value of the population parameter is unknown. The answer is that statisticians do not deal with error directly. Instead, they have developed a number of mathematical models based on probability theory and then use these models to make probabilistic statements about error.

As a result, parametric statistics involves a level of abstraction one step beyond the calculation of, say, mean reaction time in 12 schizophrenics. To understand this level of abstraction, one must realize that an observed statistic is regarded as being sampled from a population of potential observed statistics. A simple thought experiment can assist us in understanding these rather obscure concepts.

Suppose that we sampled 12 schizophrenics, calculated their mean on a reaction time paradigm, wrote down that mean on a piece of paper, and then dropped the paper into a big hat. Now repeat this process on a second sample of 12 schizophrenics. Finally, continue repeating the exercise until we have something approaching an infinite number of pieces of paper, each with its own sample mean, in that (very, very) big hat. The means in the hat can be treated as a “population of means” and the mean observed in any single reaction-time study of 12 schizophrenics can be regarded as randomly reaching into that big hat and picking one piece of paper.

The concept of error in parametric statistics deals with the variability of the means in the hat. If there is low variability, then there will probably be little discrepancy between the mean of the single piece of paper that we selected and the overall mean of all the means in the hat (i.e., the population parameter). If there is high variability, then there will probably be a greater discrepancy between the selected mean and the overall mean. Later, we will learn how a famous mathematical theorem—the central limit theorem—is used to estimate the variability of means in the hat. For the present time, realize that the mathematics of probability theory can be used to provide probabilistic statements about the discrepancy between an observed statistic and its unknown population parameter value. Hence, in parametric statistics, the concept of “population” and “sample” applies not only to observations in the physical world (something that is rather easily understood) but also to statistics (a concept that takes some getting used to).

1.3.2 Nonparametric Statistics

Parametric statistics start with a mathematical model of a population, estimate statistics from a sample of the population, and then make inferences based on probability theory and the values of the observed statistics. Nonparametric statistics, in contrast, do not begin with a mathematical model of how a variable is distributed in the population. They do, however, use probability theory to arrive at inferences about hypotheses.

The differences between parametric and nonparametric statistics are most easily seen in a simple example—do men and women differ in height? The approach used in parametric statistics would be to gather a sample of men and a sample of women, compute the mean height of the men and the mean height of the women, and then use probability theory to determine the likelihood of these two means being randomly
sampled from a single hat of means versus two different hats of means—one for men and
the other for women.

One nonparametric solution to this problem would involve the following. Line up
all the subjects—men and women together—according to height, from smallest to largest.
Now walk down this line until one reaches the point that separates the smallest 50% from
the tallest 50%. Finally, count the number of men in the bottom 50% and the number of
men in the top 50%. Do the same for the ladies. If there are no differences in height, then
there should be just as many men in the “small group” as there are in the “tall group” and
the number of women in the two groups should also be equal. Probability theory can
then be used to test whether the observed numbers differ from the theoretical 50-50 split
implied by this hypothesis.

There has been considerable debate over the relative merits of parametric versus
nonparametric statistics. The consensus answer is that it depends upon the problem at
hand. For most of the problems treated in this book, however, statisticians favor the
parametric approach because it is more powerful. That is, they are more likely than
nonparametric statistics to reject a hypothesis when in fact that hypothesis is false.

1.4 Descriptive Statistics

Descriptive statistics are numbers that “describe” a distribution. Given that
distributions have the four attributes of location, spread, symmetry, and peakedness, there
are statistics (or sets of statistics) that describe each of these.

1.4.1 Measures of Location

Measures of location, also called measures of central tendency, answer the
following question: around which number are the scores located? The three traditional
measures of location are the mode, the median, and the mean.

1.4.1.1 The Mode

The mode is the most frequent score in the set. In the pkc data, the most frequent
score for variable openarm is 0. This illustrates one problem with the mode—even
though it is the most frequent value, it need not be a number around which the rest of the
scores are clustered. The only time the mode is used in statistics is to verbally describe a
distribution of scores. A distribution with a single peak is called unimodel and one with
two peaks is termed bimodal. Sometimes the terms major mode and minor mode are used
to refer to, respectively, the larger and smaller peak in a bimodal distribution.

1.4.1.2 The Median

The median is defined as the “middle score.” It is the observed score (or an
extrapolation of observed scores) that splits the distribution in half so that 50% of the
remaining scores are less than that value and 50% of the remaining scores exceed that
value. The simplest way to compute the median is to sort the data from smallest to
largest value. If there are N observations and N is odd, then the median is the score of
observation \((N + 1)/2\). For example, if there are 13 observations, then the median is the score of the 7th sorted observation. If \(N\) is even, then the median is the usually defined as the average score of the two middle observations. For example, if \(N = 14\), then the median is the average of the 7th and 8th ordered scores. More elaborate formulas may sometimes be used when there are several tied scores around the median (see Zar, 1999).

Because data set pkc has 45 observations, the median will be the score associated with the 23rd ordered observation—11.1 in this case.

### 1.4.1.3 The Mean

There are three types of means: (1) the arithmetic mean; (2) the geometric mean; and (3) the harmonic mean. The arithmetic mean is simply the average—add up all the scores and divide by the number of observations. It is usually denoted in algebra by placing a bar above a symbol. The mathematical equation for the arithmetic mean is

\[
\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}.
\]

By far, the arithmetic mean is the most frequently used mean in statistics.

The geometric mean is sometimes used when an observed score is a result of a complex multiplicative process. (A multiplicative process will be described more fully later). The algebraic definition of the geometric mean is the \(N\)th root of the product of the scores, or

\[
\text{GM}(X) = \sqrt[N]{\prod_{i=1}^{N} X_i} = \sqrt[N]{X_1X_2\cdots X_N}.
\]

Note that if one of the scores is 0, then the geometric mean must also be 0. Note also that the geometric mean may be mathematically undefined when one or more of the scores are negative. These are not faults of the geometric mean. Instead, they represent cases where the geometric mean is not the appropriate measure of location.

The final mean is the harmonic mean. It is the inverse of the average of the reciprocals of the scores or

\[
\text{HM}(X) = \left(\frac{\sum_{i=1}^{N} 1/X_i}{N}\right)^{-1}.
\]

The harmonic mean is very seldom used in statistics.

### 1.4.2 Measures of Spread

Measures of spread index the variability of the scores. When spread is small, then the scores will be tightly clustered around their location. When spread is large, then scores are widely dispersed.
1.4.2.1 The Range

The range is defined as the difference between the largest and the smallest score. In the pkc data set, the largest score is 37.7 and the smallest is 0, so the range is $37.7 - 0 = 37.7$.

Although the range is sometimes reported, it is rarely used in statistical inference because it is a function of sample size. As sample size increases, there is a greater likelihood of observing an extreme score at either end of the distribution. Hence, the range will increase as sample size increases.

1.4.2.2 Quantiles, Percentiles and Other “Tiles”

The $p^{th}$ percentile is that score such that $p$ percent of the observations score lower or equal to the score. For example the $85^{th}$ percentile is the score that separates the bottom 85% of observations from the top 15%\footnote{In some areas of research, the term percentile is also applied to an observation. In education or clinical psychology, for example, one might read that a student was in the 63rd percentile on a standardized test.}. Note that the median, by definition, is the $50^{th}$ percentile. By definition, the minimum score is sometimes treated as $0^{th}$ percentile and the maximum score is the $100^{th}$ percentile\footnote{These definitions are not universal. Some algorithms begin with the $1^{st}$ percentile.}.

Other terms ending in the suffix “tile” are used to divide the distribution into equal parts. Quartiles, for example, divide the distribution into four equal parts—the lowest one-quarter of scores (termed the first quartile), those between the $26^{th}$ and $50^{th}$ percentile (the second quartile), those between the $51^{st}$ and $75^{th}$ percentile (the third quartile), and the one-quarter comprising the highest scores (the fourth quartile). Similarly, quartiles divide the distribution into five equal parts (the $20^{th}$, $40^{th}$, $60^{th}$, $80^{th}$, and $100^{th}$ percentiles), and deciles divide the distribution into ten equal parts. The generic term quantile applies to any “tile” that divides the distribution into equal parts.

The measure of spread most frequently associated with percentiles is the inter quartile range, sometimes called the semi inter quartile range. This is the difference in raw scores between the $75^{th}$ percentile and the $25^{th}$ percentile. For example, in the pkc data set, the score at the $75^{th}$ percentile (i.e., the third quartile) is 16.5 and the score at the $25^{th}$ percentile (the first quartile) is 5.8, so the inter quartile range is $16.5 - 5.8 = 9.7$. The inter quartile range is seldom used in statistical inference.

1.4.2.3 The Variance and the Standard Deviation

The variance and the standard deviation are the most frequent measures of spread. To understand the meaning of these statistics, let us first explore them in a population and then consider them in a sample from that population.

1.4.2.3.1 The Population Variance and Standard Deviation

The population variance is defined as the average squared deviation from the mean and in many statistical procedures is also called a mean square or MS. To
understand the variance, it is necessary to first spend a few moments discussing squared deviations from the mean. Table X.X gives hypothetical data on the five observations that we will consider as the population. The column labeled $X$ gives the raw data, and it is easily verified that the mean of the five scores is 8.

<table>
<thead>
<tr>
<th>Observation</th>
<th>$X$</th>
<th>$X - \mu$</th>
<th>$(X - \mu)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>Sum:</td>
<td>40</td>
<td>0</td>
<td>46</td>
</tr>
<tr>
<td>Average:</td>
<td>8</td>
<td>0</td>
<td>9.1</td>
</tr>
</tbody>
</table>

The column labeled $X - \mu$ gives the deviation from the mean. (A Greek lower case mu, $\mu$, is the traditional symbol for a population mean). One simply takes the raw score and subtracts the mean from it, so the value for the first observation is $12 - 8 = 4$. Note that the sum of the deviations from the mean is 0. This is not coincidental—it will always happen.

The final column gives the squared deviation from the mean—just take the number in the previous column and square it. The sum of the squared deviations from the mean—in this example, 46—is a very important quantity in statistics. It is referred to as the sum of squares and is abbreviated as $SS$. By definition, the average of the squared deviations is the population variance, almost denoted as $\sigma^2$. Hence, the formula for the population variance is

$$\sigma^2 = MS = \frac{SS}{N} = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}.$$

The population standard deviation is simply the square root of the variance, and hence is almost always denoted by $\sigma$. Thus,

$$\sigma = \sqrt{\sigma^2} = \sqrt{MS} = \sqrt{\frac{SS}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}.$$

It is because of this simple relationship between the standard deviation and the variance that we treat the two statistics together.
1.4.2.3.2 The Sample Variance and Standard Deviation

The sample variance and standard deviation are usually denoted as \( s^2 \) and \( s \), respectively. With rare exception, the purpose of calculating a sample variance is to estimate the population variance. One might intuit that the estimate could be obtained by simply plugging the observed mean into Equation X.X, and indeed that idea is not a bad starting point. The problem, however, is that the observed mean is an estimate of the population mean, so some correction must be made to account for the error introduced by replacing a population parameter with a fallible estimate.

The mathematics behind this correction are too complicated to explain here, so let us simply state that the correction involves dividing the sums of squares by its degrees of freedom (df) instead of the sample size, \( N \). The degrees of freedom for estimating a population variance from sample data using the sample mean equal \( N - 1 \), so the formula for the sample variance is

\[
s^2 = MS = \frac{\sum (X_i - \bar{X})^2}{df} = \frac{SS}{N - 1}.
\]

The sample standard deviation is simply the square root of this quantity.

Most introductory statistics texts treat the degrees of freedom as a mystical quantity and merely provide formula to calculate the df for various situations. We avoid that here by providing a short digression and before continuing discussion of the variance.

1.4.2.3.2.1 Degrees of Freedom

The degrees of freedom for a statistic may be loosely viewed as the amount of information in the data needed to figure everything else out. That is not much of a definition, so let us explore some examples. Suppose that a sample of 23 patients with seizure disorder had 11 males. It is perfectly obvious then that the sample has 12 females. Variable “sex” in this case has one degree of freedom even though it has two categories. Why? Because given the sample size, we only need to know the number of one sex—either the number of males or the number of females—before we can “figure everything else out.”

Similarly, given a sample mean based on \( N \) observations, we need to know only \( N - 1 \) raw scores before we can “figure everything else out.” The “everything else” in this case would be the value of the remaining raw score. For example, suppose there are three observations with scores of 6, 8, and 13. The mean is 9, so we have the equation

\[
9 = \frac{X_1 + X_2 + X_3}{3}.
\]

From this equation, one needs to know any two of the scores before being able to calculate the third.

In these two examples, it was quite easy to calculate the degrees of freedom. In many other problems, however, the degrees of freedom can be very difficult to derive, so the recommended method to determine the degrees of freedom for a problem is to “look it up in the text.”
1.4.2.3.3 The Variance Again

We are now at the point where we can give a generic definition and formula for estimating a variance from sample data. The estimate of a population variance from sample data, also called a mean square, is the sum of squared deviations from the estimated mean, also called the sum of squares, divided by the appropriate degrees of freedom, or

\[ \hat{\sigma}^2 = MS = \frac{SS}{df} = \frac{\sum_{i=1}^{N} (X_i - \hat{\mu})^2}{df} . \]  \hspace{1cm} (X.X)

This is a very important definition, so it should be committed to memory.

At this point, we note that the degrees of freedom will be \( N - 1 \) when we want to estimate the population variance of a single variance by using the sample mean as an estimate of \( \mu \). In statistical procedures like multiple regression and the analysis of variance that will be discussed later, different pieces of information from the data are used to estimate \( \mu \). Hence, the degrees of freedom can be something other than \( N - 1 \). Regardless, Equation X.X will always apply. This is another good reason for committing it to memory.

The variance has several important properties. First, if all scores are the same, then the variance will be 0. This makes intuitive sense because it tells us that there is no “spread” in the data. Second, when at least one score differs from the others, then the variance will be a positive number greater than 0. This will always be true even if all the raw scores are negative. The formula for the variance involves taking a deviation from the mean and then squaring it, giving a number that must be equal to or greater than 0. The most important property of the variance is that it can be partitioned. One can figuratively regard the variance of a variable as a pie chart that can have different “slices” reflecting the extent to which other variables predict the first variable. For example, a certain proportion of the variance in open field activity in mice may be explained in terms of mouse genotype.

An undesirable property of the variance is that it is expressed in terms of the square of the original unit of measurement. For example, if the original unit was milligrams, then the variance is expressed in milligrams squared. The standard deviation, on the other hand, is expressed in the same measurement units as the original variable.

1.4.2.4 The Coefficient of Variation

The coefficient of variation or CV equals the standard deviation divided by the mean, or

\[ CV = \frac{s}{\bar{X}} . \]  \hspace{1cm} (X.X)

It is not unusual to multiply the above quantity by 100 in order to express it as a percentage.

The CV should only be used on data measured by ratio scales. Despite this limitation, the statistic is seriously underutilized in neuroscience because it has the very desirable property of being unaffected by the units of measurement. Because the standard deviation and the mean are both expressed in the same units, dividing the first by the
latter removes the measurement units to permit direct comparison across metrics, across organ systems, or across organisms. For example, the weight of a sample of rats will have the same CV regardless of whether weight is measures in grams, kilograms, ounces, or pounds. Similarly, one could compare the CV for weight in rats to the CV for weight in humans to see if the variability in rat weight (relative to the size of rats) is greater than or less than the variability of human weight (relative to the size of humans).

One specific application of the CV would be to compare the variability of, say, a neurotransmitter in different areas of brain. For example, the variability of dopamine might be expected to be greater in areas like the basal ganglion that have a high concentration of dopaminergic cell bodies than areas with lower levels. Using the CV to compare brain regions avoids this difficulty.

1.4.3 Measures of Symmetry

The customary measure of symmetry is called skewness. Strictly speaking, the skewness statistic measures the lack of symmetry in a curve because a skewness of 0 implies that the distribution is symmetric. Figure X.X depicts two skewed distributions. The panel to the left depicts a negatively skewed distribution where the scores are drawn out on the side of the distribution pointing toward the negative side of the number line. The panel on the right illustrates a positively skewed distribution—the scores trail off on the positive side of the number line.

There are several different formulas used to compute skewness, so one must always consult the documentation for a computer program to know which one is being reported. The most common formula takes the form

$$\text{skewness} = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \frac{(X_i - \overline{X})^3}{s^3}$$

where $s^3$ is the cube of the standard deviation. A positive value of skewness denotes a positively skewed distribution and a negative value, a negatively skewed distribution.

Figure for skewed distributions:
(not completed yet)

1.4.4 Measures of Peakness/Flatness

The final attribute of a distribution is called kurtosis and is the extent to which scores, relative to the variance of the distribution, are concentrated close to the mean (a “peaked” distribution) or are disproportionately present in the tails (a “flat” distribution). The three curves in Figure X.X illustrate kurtosis. The curve with the solid line is a normal curve. The curve with the short dashed lines illustrates a peaked distribution known in statistical jargon as a leptokurtotic distribution after the Greek word leptos, meaning “thin.” A leptokurtotic distribution has relative more scores close to the mean and relatively fewer scores in the tails of the distribution that would be expected given the variance of the distribution.

The curve in Figure 1 with the longer dashed lines is a platykurtotic curve, named from the Greek word platys, meaning broad or flat. Here, there are relatively more scores
in the tail and relatively few scores close to the mean than would be expected by the variance of the distribution.

Like skewness, there are several formulas for kurtosis. The most frequent one is

$$kurtosis = \frac{N(N + 1)}{(N - 1)(N - 2)(N - 3)} \left( \frac{3(N - 1)^2}{(N - 2)(N - 3)} \right) \frac{\sum_{i=1}^N (X_i - \bar{X})^4}{s^4}. \quad (X.X)$$

A negative value of kurtosis denotes platykurtosis while a positive value signifies leptokurtosis.

Figures for kurtotic distributions:
(not completed yet)

1.4.5 Measures of Covariation

In addition to statistics that describe a distribution or estimate parameters associated with a distribution, there is a whole class of statistics that measure how well scores on one variable as associated with scores on a second variable. We call these measures of covariation. When both variables are measured on an interval or ratio scale, then the most customary measure is the covariance. Denoting the two variables as $X$ and $Y$, the covariance is defined as

$$cov(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}. \quad (X.X)$$

A second measure of covariation is the correlation coefficient. There are several types of correlation coefficients but the most often used one is called the Pearson Product Moment (PPM) correlation. The PPM correlation estimates the population correlation that is defined in the mathematical equation for two variables that have a bivariate normal distribution. The formula is

$$corr(X, Y) = r_{XY} = \frac{cov(X, Y)}{s_X s_Y}, \quad (X.X)$$

or the covariance of the two variables divided by the product of their standard deviations. Most statistical procedures use the covariance whereas most reports of covariation give the correlation. There is a very good reason for this—the covariance is much more difficult to interpret than is a correlation. To examine this, let us first discuss the two major properties of a measure of covariation: direction and magnitude.

1.4.5.1 Direction of Covariation

The direction of a measure of covariation is given by the sign of the coefficient. A positive sign denotes a positive or direct relationship. Here, high scores on one variable are associated with high scores on the second variable and low scores on the first variable predict low scores on the second variable (see Panel A of Figure X.X).

A negative sign denotes a negative or inverse relationship. Here high scores on one variable are associated with high scores on the second variable and low scores on the
first variable predict high scores on the second variable. Panel B of Figure X.X depicts a negative or inverse relationship.

1.4.5.2 Magnitude of Covariation

The magnitude of covariation measures the strength of the relationship between the two variables. The problem with using the covariance as an index of magnitude is that its value depends on the scale of the two variables. Consequently, given two covariances, say .012 and 12,000, it is impossible to way which of them indicates a stronger relationship.

The correlation coefficient, on the other hand, is independent of the scales of the two variables. Hence, it is the preferred statistic for reporting covariation. The correlation has a natural mathematical lower bound of –1.0 and an upper bound of 1. A correlation of 0 implies no statistical relationship between the two variables. That is, neither of the variables can predict the other better than chance. A correlation of 1.0 denotes a perfect, positive relationship. If we know an observation’s score on one variable, then we can predict that observation’s score on the second variable without error. A correlation of –1.0 denotes a perfect, but negative relationship. Here, we can also predict an observation’s score on the second variable given the score on the first variable—it is just that a high score on the first variable will predict a low score on the second.

The best index of magnitude is the square of the correlation coefficient. Note that squaring removes the sign so that correlations of .60 and -.60 both have the same estimate of magnitude (.36). The square of the correlation measures the proportion of variance in one variable statistically explained by the other variable. In less jargonish terms, it gives the extent to which knowledge of individual differences in one variable predict individual differences in the second variable. Note that these statements have been deliberately phrased is such cautious terms as “statistically explained” and “predict.” This is because a correlation coefficient has no causal implications. If \( X \) is correlated with \( Y \), then: (1) \( X \) may cause (or be one of several causal factors of) \( Y \); (2) \( Y \) may cause (or be one of several causal factors of) \( X \); (3) \( X \) and \( Y \) because are both caused by the same (or some of the same) factors; and (4) any combination of situations 1, 2, and 3 may occur.

1.5 Other Terms

1.5.1 Outliers

An outlier is defined as a data point that is well separated from the rest of the data points. Velleman (19xx) divides outliers into two types, blunders and rouges. A blunder is a data point resulting from instrument, measurement, or clerical error. For example, IQ, a staple of the neuropsychological test battery, is typically scaled so that the population mean is 100 and the population standard deviation is 15. A transposition error in data entry may record an IQ of 19 in a database instead of the correct value of 91. Note that a blunder does not always result in an outlier—transposing an IQ score of 98, for instance, will not give a score disconnected from the IQ distribution. Blunders, however, are easily detected and corrected by following recommended practice and
having data entered twice. A comparison of the results from the first entry with those in the second entry will almost always catch clerical data-entry errors.

A rouge, on the other hand, is a legitimate data value that just happens to be extreme. In clinical neuropsychology, for example, it is not unusual to encounter someone with an extremely low IQ score because of a gross genetic or environmental insult to the central nervous system. What should be done with rouge data points depends entirely on the purpose at hand. On the one hand, the observation(s) should never be deleted from the data set because they can provide very important information. For example, in a meta-analysis\(^5\) of data on the genetics of human aggression, Miles & Carey (199x) found that the two rouge values were the only two studies that measured aggression in the laboratory—all other studies used self-report questionnaires. On the other hand, the presence of a single rouge value in some statistical procedures can result in very misleading inference. Here, there are three possible solutions: (1) temporarily set the rouge value to missing value for the analysis; (2) transform the data; or (3) use a nonparametric procedure that is not sensitive to outliers. We discuss these later in the context of statistical procedures.

Although we have spoken of rouges as applying to a single variable, it is possible to have a multivariate rouge. This is a data point that might not be detected as a rouge when each variable is examined individually but emerges as an outlier when the variables are considered together. The classic example of a multivariate rouge would be an anorexic woman who is 5’10” tall and weighs 95 pounds. She is not remarkable in either height taken alone or weight taken alone, but she would be an outlier in terms of the joint distribution of height and weight.

1.5.2 Effect Size  
(not yet written)

1.5.3 Statistical Power  
(not yet written)

---

\(^5\) A meta-analysis is a systematic statistical analysis of a series of empirical studies.