In the previous chapter we discussed measurement and the various levels at which we can use measurement to describe the extent to which an individual observation possesses a particular theoretical construct. Such a description is referred to as a datum. An example of a datum could be how many conversations a person initiates in a given day, or how many minutes per day a person spends watching television, or how many column inches of coverage are devoted to labor issues in The Wall Street Journal. Multiple observations of a particular characteristic in a population or in a sample are referred to as data.

After we collect a set of data, we are usually interested in making some statistical summary statements about this large and complex set of individual values for a variable. That is, we want to describe a collective such as a sample or a population in its entirety. This description is the first step in bridging the gap between the “measurement world” of our limited number of observations, and the “real world” complexity. We refer to this process as describing the distribution of a variable. There are a number of basic ways to describe collections of data.
Describing Distributions

**Description by Enumeration**

One way we can describe the distribution of a variable is by enumeration, that is, by simply listing all the values of the variable. But if the data set or distribution contains more than just a few cases, the list is going to be too complex to be understood or to be communicated effectively. Imagine trying to describe the distribution of a sample of 300 observations by listing all 300 measurements.

**Description by Visual Presentation**

Another alternative that is frequently used is to present the data in some visual manner, such as with a bar chart, a histogram, a frequency polygon, or a pie chart. Figures 8-1 through 8-5 give examples of each of these, and the examples suggest some limitations that apply to the use of these graphic devices.

![Bar Chart Example](image)

**FIGURE 8-1** Representation of Data in a Bar Chart

The first limitation that can be seen in Figure 8-1 is that the data for bar charts should consist of a relatively small number of response categories in order to make the visual presentation useful. That is, the variable should consist of only a small number of classes or categories. The variable “CD Player Ownership” is a good example of such a variable. Its two classes (“Owns a CD Player” and “Does not own a CD Player”) lend themselves readily to presentation via a bar chart.

Figure 8-2 gives an example of the presentation of data in a histogram. In a histogram the horizontal axis shows the values of the variable (in this case the number of CD discs a person reports having purchased in the previous year) and the vertical axis shows the frequencies associated with
FIGURE 8-2  Representation of Data in a Histogram

FIGURE 8-3  Representation of Data in a Frequency Polygon
these values, that is, how many persons stated that they purchased, for instance, 8 CDs.

In histograms or bar charts, the shape of the distribution can convey a significant amount of information. This is another reason why it is desirable to conduct measurement at an ordinal or interval level, as this allows you to organize the values of a variable in some meaningful sequence. Notice that the values on the horizontal axis of the histogram are ordered from lowest to highest, in a natural sequence of increasing levels of the theoretical concept (“Compact Disc Purchasing”). If the variable to be graphed is nominal, then the various classes could be arranged visually in any one of a large number of sequences. Each of these sequences would be equally “natural”, since nominal categories contain no ranking or ordering information, and each sequence would convey different and conflicting information about the distribution of the variable. The shape of the distribution would convey no useful information at all. Bar charts and histograms can be used to compare the relative sizes of nominal categories, but they are more useful when the data graphed are at the ordinal or higher level of measurement.

Figure 8-3 gives an alternative to presenting data in a histogram. This method is called a frequency polygon, and it is constructed by connecting the points which have heights corresponding with the frequencies on the vertical axis. Another way of thinking of a frequency polygon is as a line which connects the midpoints of the tops of the bars in the histogram.

Notice that the number of response categories that can be represented in the histogram or frequency polygon is limited. It would be very difficult to accommodate a variable with many more classes. If we want to describe a variable with a large number of classes using a histogram or a frequency polygon, we would have to collapse categories, that is, combine a number of previously distinct classes, such as the classes 0, 1, 2, etc. into a new aggregate category, such as 0 through 4, 5 through 9, 10 through 14, etc. Although this process would reduce the number of categories and increase the ease of presentation in graphical form, it also results in a loss of information. For instance, a person who purchased 0 CDs would be lumped together with a person who purchased as many as 4 CDs in the 0-4 class, thereby losing an important distinction between these two individu-
als. Figure 8-4 illustrates the results of such a reclassification or recoding of the original data from Figure 8-3.

Yet another way of presenting data visually is in the form of a pie chart. Figure 8.5 shows a pie chart which presents the average weekly television network ratings during prime time.

Pie charts are appropriate for presenting the distributions of nominal variables, since the order in which the values of the variable are introduced is immaterial. The four classes of the variable as presented in this chart are: tuned to ABC, tuned to NBC, tuned to CBS and, finally, tuned to anything else or not turned on. There is no “one way” in which these levels of the variable can or should be ordered. The sequence in which these shares are listed really does not matter. All we need to consider is the size of the “slice” associated with each class of the variable.

Descriptive Statistics

Another way of describing a distribution of data is by reducing the data to some essential indicator that, in a single value, expresses information about the aggregate of all observations. Descriptive statistics do exactly that. They represent or epitomize some facet of a distribution. Note that the name “descriptive statistics” is actually a misnomer—we do not limit ourselves to sample distributions.

These descriptive statistics allow us to go beyond the mere description of a distribution. They can also be used for statistical inference, which permits generalizing from the limited number of observations in a sample to the whole population. We explained in Chapter 5 that this is a major goal of scientific endeavors. This fact alone makes descriptive statistics preferable to either enumeration or visual presentation. However, descriptive statistics are often used in conjunction with visual presentations.

Descriptive statistics can be divided into two major categories: Measures of Central Tendency; and Measures of Dispersion or Variability. Both kinds of measures focus on different essential characteristics of distributions. A very complete description of a distribution can be obtained from a relatively small set of central tendency and dispersion measures from the two categories.

Measures of Central Tendency

The measures of central tendency describe a distribution in terms of its most “frequent”, “typical” or “average” data value. But there are different ways of representing or expressing the idea of “typicality”. The descriptive statistics most often used for this purpose are the Mean (the average), the Mode (the most frequently occurring score), and the Median (the middle score).

FIGURE 8-5  Pie Chart Showing Average Rating for Three Networks
The Mean

The mean is defined as the arithmetic average of a set of numerical scores, that is, the sum of all the numbers divided by the number of observations contributing to that sum.

\[
\text{MEAN} = \overline{E} = \frac{\text{sum of all data values}}{\text{number of data values}}
\]

or, more formally,

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

which is the formula to be used when the data are in an array, which is simply a listing of a set of observations, organized by observation number. An example of data in an array can be found in Table 8-1(a).

Before we proceed, a few words about notation. The subscript “i” (in \(X_i\)) in the above formula represents the fact that there will be a number of values of the variable \(X\), one for each observation: \(X_1\) is the value of \(X\) for observation one (the first subject or respondent), \(X_2\) is the value of \(X\) for the second observation, etc., all the way to \(X_N\), which is the value of \(X\) for the last observation (there are \(N\) observations in the sample). If we observe 25 college students and ask them how many compact discs they have bought over the last year, the response of the first person will be \(X_1\); the second person’s response will be \(X_2\); and the last, or Nth response, will be \(X_{25}\). The symbol \(E\) in the formula instructs us to sum all these values of \(X\), beginning with \(X_1\) (i = 1) and continuing to do this until the last observation (\(N\), or 25 in the example) has been included. We will encounter this notation regularly in the chapters to come.

The reason why the sum of all the observations’ values is divided by \(N\) (the number of observations) is probably intuitively clear. Computing the Mean is something you have likely done innumerable times. It may, however, be informative to explain why the simple sum of a set of observations does not have much utility for describing distributions. The reason is that its value is dependent upon two factors: the values of the individual observations and the number of observations.

A given sum can be obtained in many different ways: four individuals who each watch 5 hours of television per day, watch, among them, 20 hours of television. The same total would also be obtained from 20 individuals who each watch 1 hour per day. These are the same sums, but they are obtained from very different data distributions. By knowing that the group as a whole watched 20 hours, we cannot draw any conclusions about how much each individual watched without knowing how many individuals were in the group. When we compute the mean we standardize the simple sum by spreading it evenly across all observations. The mean then represents, in our example, the “average viewing hours per person” and the information contained in the mean can be interpreted without having to know the number of individual observations in the distribution.

The mean can also be interpreted as the “balance point” or the “value center” of the distribution. Referring to the mean this way has its basis in physics, as is illustrated in Table 8-2. If we look at a data distribution as if it were a seesaw, then the mean can be taken as that point where the fulcrum keeps the board perfectly balanced. In Table 8-2(a) the two observations (6 and 2) are both 2 units away from the mean of 4, and 4 is the point where the distribution is balanced. Another way of stating this is to say that 6 has a deviation of +2 from the mean, and that 2 has a deviation of -2 from the mean. In Table 8-2(b) the value 6 again has a deviation of +2, but now this deviation is offset by the two values 3 which are each 1 unit away (for a total of two) in the opposite direction. Here too the mean “balances out” the values on both sides of it within the data distribution. This characteristic of the mean is also known as the “Zero-Sum Principle”; in any distribution, regardless of size, shape or anything else, the deviations of scores from the mean will cancel one another out, or sum to 0.00, as can be seen in Table 8-2. We will encounter the Zero-Sum Principle again later on in this chapter.

In addition to being ordered in the form of a simple array, data can also be organized in a classified frequency distribution, as is also shown in Table 8-1. A classified frequency distribution is...
merely a convenient way of listing all the observations in a distribution by taking advantage of the
fact that the observed values usually have multiple occurrences. Instead of listing, for instance, the
value 5 eleven times, we create a table which lists the observed value as well as the number of times
(or f, the abbreviation for frequency) the value was observed.

When data are in the form of a classified frequency distribution, the formula for the mean
changes to take that into consideration:

**Table 8-1a  The Organization of a Data Set in a Data Array**

<table>
<thead>
<tr>
<th>Observation number</th>
<th>Observation value</th>
<th>Observation number</th>
<th>Observation value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>33</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>34</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>35</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>37</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>38</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>39</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>40</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>42</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>43</td>
<td>7</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>44</td>
<td>7</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>46</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>47</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>49</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
<td>50</td>
<td>9</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>51</td>
<td>7</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>52</td>
<td>6</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
<td>53</td>
<td>8</td>
</tr>
<tr>
<td>27</td>
<td>7</td>
<td>54</td>
<td>7</td>
</tr>
</tbody>
</table>

\[ N=54 \quad \sum_{i=1}^{N} X_i = 364 \]

\[ \bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{364}{54} = 6.74 \]
which instructs us to sum a set of products \((f \cdot X_i)\) and to divide that sum by the number of observations. Multiplying the value of a class in a classified frequency distribution by the class frequency has the same effect as adding it the same number of times when we compute the mean in the simple data array. For example, the value “5” appears 11 times in the data array. If we compute the mean by adding all the observation values in the array, we will add the value “5” to the total sum 11 times. Multiplying 5 by 11 and adding it to the total sum has the same result.

**The Mode**

The Mode is simply the most frequently occurring score in a distribution. The mode can be determined by observation, and its value is most easily determined in a classified frequency distribution. All we need to do is to find the highest frequency of occurrence in the frequency \(f\) column and note the observation value that occurred with that frequency. In the example shown in Table 8-1, the highest frequency is associated with the value 7; it is the mode for this distribution. A distribution such as this one is called a unimodal distribution as it only has one mode.

Suppose that the value 6 (or some other value) also had occurred 15 times. In this case there would have been two modes, and the distribution would then be called a bimodal distribution. It is in fact possible to have any number of modes in a distribution, although the usefulness of the mode as a descriptive measure decreases in distributions where there are many modes.

Since the mode represents the most frequently occurring score in a distribution, the mode is also often referred to as the “probability” center of the distribution. If we were to randomly select one single observation out of the distribution, the modal value is the one that we would be most likely to observe, since it occurs more often in the data than any other value.

**The Median**

The median is the middle score in a distribution, determined after all the values in a distribution have been rank-ordered from lowest to highest, or vice-versa. The median can also be defined as that point in a distribution above which and below which lie 50% of all the cases or observations.
We will use the data from Table 8-1 to show a method for determining the median for a classified frequency distribution. This classified frequency distribution shows a total of 54 observations of \( X_i \) divided over 5 different values. The median is to be located in such a way that 50% of the cases are below it (and 50% above). This means that the median will lie between the 27th and the 28th case, so there will be 27 cases below the median (cases 1 through 27) and 27 cases (28 through 54) above the median. This places the median somewhere between 6 and 7 in the distribution.

How do we determine the precise location? Consider the following logic. The value category 5 contains the eleven cases that are the lowest scores (20.37% of all observations). We find these eleven cases also in the column headed by “cumulative frequency”. In passing through the category 6 we find an additional 13 cases, for a cumulative frequency (so far) of 24 cases (which constitute 44.44% of all the cases in the distribution) — but we have not encountered the median yet; the median is not in the category 6. In passing through the category 7 we count an additional 15 cases, for a total of 39, or 72.22% of the total number of observations. Now we are above the 50% mark, indicating that the median is located somewhere after we get out of the class 6 but before we get out of the class 7.

Note that we have no data value exactly at the median in this distribution, as the median lies somewhere in the middle of class 7. But we can construct a value which represents the real meaning of the median as the center score, by estimating (through interpolation) a fraction within the class 7.

In order to do this we have to recognize that the categories we have labeled as 5, 6, 7... etc. actually contain a range of values. In fact, by rounding, we say that all the numbers between 6.50000... and 7.49999... will fall within the category we have called “7”. The real lower limit of the value “7” is then 6.5000.
We can interpolate a value for the median by using the following formula which is based on this lower limit:

\[ \text{Median} = L + \frac{N/2 - cf}{f(\text{class})} \]

where:
- \( L \) is the true lower limit of the class in which we expect to find the median (in this case, 6.5)
- \( N \) is the total number of observations in the distribution (here 54)
- \( cf \) is the cumulative frequency UP TO but NOT INCLUDING the class in which we expect the median to fall (in this case, 24)
- \( f(\text{class}) \) is the number of cases in the class in which we expect the median to fall (in this case, 15), and
- \( I \) is the width or size of the class or interval (in this case equal to 1.0, for instance, from 6.5 to 7.5)

For this data set the median is calculated to be:

\[ \text{Median} = 6.5 + \frac{54/2 - 24}{15} \]

\[ = 6.5 + \frac{27 - 24}{13} \]

\[ = 6.5 + \frac{3}{15} \]

\[ = 6.5 + 0.2 = 6.7 \]

If we look closely at the formula we can see its logic: \( N/2 \) is the number of cases at the 50% mark, so if we subtract from this number (27) the cumulative number of cases observed up to the lower limit of the interval (24), the difference gives the number of additional cases needed (3) in the interval in which the median occurs. Since this interval holds 15 cases (and has a range of 1), we need to go 3/15 of the range into the interval. Adding this 3/15, or 0.2, to the lower limit of the interval (6.5), gives us the final value of the median.

**Comparing the Mean, the Mode and the Median**

The information obtained from these three measures of central tendency in a data distribution is similar in the sense that all reflect some aspect of the data values which is “typical” of the whole distribution. But they differ in the kind of “typicality” which they report and in how sensitive they are to changes in the values of the observations.

The mean represents the balance point, or center of gravity of the distribution. Its value will change when there is a change in any of the data values in the distribution.

The mode represents the most frequent or probable single value in the distribution. If the value of a datum in the distribution changes from a non-modal value to the modal value, the value calculated for the mode remains the same, even though the mean would (and the median might) change.

The median represents the middle score of the distribution. If the value of a datum is changed so that its position relative to the magnitude of the other values is not changed, the median will remain the same, even though the mean would, and the mode might.

To illustrate the differences in these measures’ sensitivity to changes in the data, consider what would happen if the 6 observation of the value 9 were to change to 6 observations of the value 19. The effect on the mode and the median is nil, but the value of the mean increases to 7.85.
FIGURE 8-6  Example of a Symmetrical Distribution

FIGURE 8-7  Example of a Skewed Distribution
The Shape of the Distribution: Skewness

There is one important situation in which all three measures of central tendency are identical. This occurs when a distribution is symmetrical, that is, when the right half of the distribution is the mirror image of the left half of the distribution. In this case the mean will fall exactly at the middle of the distribution (the median position) and the value at this central point will be the most frequently observed data value, the mode. An example of such a distribution is shown in Figure 8-6.

If the values of the mean, the mode and the median are identical, a distribution will always be symmetrical. To the extent that differences are observed among these three measures, the distribution is asymmetrical or “skewed”.

Asymmetry will occur whenever the distribution contains one or more observations whose deviation from the mean is not matched by an offsetting deviation in the opposite direction. Figure 8-7 gives an illustration of a mildly skewed distribution, where we have simply changed some of the values shown in Figure 8-6.

This distribution contains some values on the high end of the distribution which are very far from the mean that are not matched by corresponding values on the low end of the distribution. Consequently, the value of the mean has increased (from 7.00 to 7.46) and is drawn away from the center of the distribution toward the side of the distribution where the extreme values occurred. Note that the value of the median and the mode have not changed. Had the extreme cases been located at the low end of the distribution, then the mean would have been drawn in that direction.

The degree of discrepancy between the median and the mean can then be interpreted as an indicator of skewness, or the lack of symmetry. If these two indices are identical, the distribution is symmetrical. If the mean is greater than the median, the extreme values (or “outliers”) are located at the high end of the distribution and the distribution is said to be “positively skewed”; when the outliers are at the low end of the distribution, the mean will be less than the median and the distribution is said to be “negatively skewed”.

![Figure 8-8 Distributions with Identical Measures of Central Tendency and Varying Degrees of Dispersion](image-url)
Measures of Dispersion

The measures of central tendency focus on what is typical, average or in the middle of a distribution. The information provided by these measures is not sufficient to convey all we need to know about a distribution. Figure 8-8 gives a number of examples of distributions which share the same measures of central tendency, but are radically different from one another in another aspect, that is, how the observations are distributed (or dispersed) around these measures of central tendency.

For simplicity’s sake we have presented these three distributions as symmetrical distributions, and from the preceding sections you will recall that the mean, median and mode are all equal in such a case. All similarity ends there, however. When we inspect distribution A, we see that the values in that distribution all equal 5 with the exception of single instances of the values 4 and 6. Distribution B shows a pattern of greater dispersion: the values 3 and 4 and 6 and 7 occur repeatedly in this distribution, but not as frequently as the value 5. The greatest degree of dispersion is observed in distribution C: here values as small as 1 and as large as 9 are observed.

The main point to remember from these examples is that knowing the mean, the median or the mode (or all of these) of a distribution does not allow us to differentiate between distributions A, B and C in Figure 8-8 above. We need additional information about the distributions. This information is provided by a series of measures which are commonly referred to as measures of dispersion. These measures of dispersion can be divided into two general categories: between-points measures of dispersion and around-point measures of dispersion.

Between-Points Measures of Dispersion

One way of conceptualizing dispersion is to think of the degree of similarity-dissimilarity in a distribution. There is dispersion when there is dissimilarity among the data values. The greater the dissimilarity, the greater the degree of dispersion. Therefore a distribution in which the values of all the observations are equal to one another (and where the value of each observation then is equal to the value of the mean) has no dispersion. If the values are not equal to one another, some will be larger than the mean and some will be smaller. Measuring the distance between the highest and lowest values will then provide us with a measure of dispersion. The description “between-points” is used, since the measure of dispersion refers to the end points of the distribution.

The Crude Range and the Range

Based on this logic are two commonly used measures of dispersion: the crude range and the range.

The former is defined as follows:

\[
\text{Crude range} = \text{Highest observed value} - \text{Lowest observed value}
\]

For the examples in Figure 8-8, the crude range for the three distributions is:

- Distribution A: 6 - 4 = 2
- Distribution B: 7 - 3 = 4
- Distribution C: 9 - 1 = 8

The range is defined as

\[
\text{Range} = \text{Highest observed value} - \text{lowest observed value} + 1
\]

which is therefore also equal to the crude range, + 1.

For the examples in Figure 8-8 this formula gives Distribution A a range of 3, Distribution B a range of 5, and Distribution C a range of 9.

The range can be interpreted as the number of different value classes between the highest and the lowest observed values, including the highest and lowest classes themselves. But this does not
mean that observations were in fact found in all the classes between these extremes. The weakness of the range (and of the crude range) is that its value is only sensitive to the highest and lowest valued observations in the distribution. This can be misleading.

Distribution C in Figure 8-8 has a range of 9, indicating that there were 9 classes of values, and, in fact, cases were observed in all nine classes. But consider the following distribution which also has a range of 9:

<table>
<thead>
<tr>
<th>Xi</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

The range for this distribution is also equal to 9 (9 - 1 + 1), but the dispersion of values is very different from that found in Figure 8-8. The distribution could more accurately be described as consisting of the values 4, 5 and 6, with two extreme values, 1 and 9. These infrequently occurring extreme values (sometimes called “outliers”), in fact determine the value of the range in this case. For this reason we should be very careful in interpreting the value of the range.

**Around-Point Measures of Dispersion**

The second category of measures of dispersion is based on the same conceptualization of dispersion as explained above, but with some different logic in its measurement. Instead of looking at the extreme instances of data values as we did in computing the range, the around-point measures of dispersion look at differences between individual data values and the mean of the set of values.

Because these methods are based on comparing data values (located around the mean) with the mean, we call them “around-point” measures of dispersion. Dispersion is measured by calculating the extent to which each data value is different from that point in the distribution where the mean is located. We call this difference the deviation from the mean. For this deviation we will use the following notation:

\[ \sum f(X_i - \bar{X}) = 0 \]

**Table 8-3** Deviations and Sums of Deviations from the Mean in Two Distributions

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_i)</td>
<td>(f)</td>
<td>((X_i - \bar{X}))</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>-2</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>+1</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>+2</td>
</tr>
<tr>
<td>(\sum f(X_i - \bar{X}) = 0)</td>
<td>(\sum f(X_i - \bar{X}) = 0)</td>
<td></td>
</tr>
</tbody>
</table>
Table 8-3 gives an example of two distributions with varying degrees of dispersion. For each value of \( X \), the deviation from the mean has been computed. To determine the degree of dispersion in each distribution, it would appear that we would merely need to sum the individual deviations, like this:

\[
\sum f \cdot (X_1 - \overline{X})
\]

However, because of the “Zero-Sum Principle” we will quickly find that the sums of the deviations will equal zero for any distribution, regardless of the degree of dispersion. It follows that the average deviation also has to be zero. For this reason sums of deviations (and thus average deviations) are useless for comparing the dispersion of these distributions.

Squaring the Deviations. However, we can readily overcome the Zero-Sum Principle by squaring the deviations from the mean. This process is illustrated in Table 8-4. Squaring a deviation rids us of the sign of the deviation, because squaring a positive deviation yields a positive value (or, more precisely, an unsigned one); and squaring a negative deviation gives the same result. Summing the squared deviations will then produce a non-zero sum under all conditions but one: in a distribution with no dispersion (when all the values in a distribution are equal to one another and therefore equal to the mean), all the deviations from the mean will equal zero, will remain equal to zero when squared and will sum to zero.

The Sum of Squared Deviations. Below is the notation for computing the Squared Deviation from the Mean:

\[
(X_1 - \overline{X})^2
\]

Based on the squared deviation, the Sum of the Squared Deviations for a classified frequency distribution is then computed as follows:

\[
\sum f \cdot (X_1 - \overline{X})^2
\]

Table 8-4 illustrates the computation of the Sum of Squared Deviations (also often referred to as the Sum of Squares) for the two distributions A and B.

For Distribution A, the distribution with the smaller dispersion, the Sum of the Squared Deviations is equal to 64, for Distribution B the Sum of Squares equals 174.

The Variance

If we want to compare distributions with differing numbers of observations, the sum of squared deviations needs to be standardized to the number of observations contributing to that sum. To do this, we compute the average squared deviation, which is more frequently referred to as the variance. The formula for determining a distribution’s variance is as follows:

\[
\text{Variance} = \frac{\sum f \cdot (X_1 - \overline{X})^2}{N}
\]

Applying this formula to the sums of squares of distributions A and B yields variances of 1.68 and 3.48, respectively, as illustrated in Table 8-4. The value of the variance is interpreted as the average squared deviation of all observations and can be used to compare different-sized distributions. Our conclusion from comparing the variances of A and B is that the scores in Distribution B are more dispersed than are those in Distribution A.
The Standard Deviation

Remember that we squared an observation’s deviation from the mean in order to overcome the Zero-Sum Principle. The sum of squared deviations and the variance we obtained are therefore squared “versions” of the original deviations from the mean. To return to the magnitude of the original units of measurement we simply reverse the distortion introduced from squaring an observation’s deviation from the mean by taking the square root of the variance. This produces another measure of dispersion, which is referred to as the standard deviation. The formula for determining the value of the standard deviation is defined as follows:

\[
\text{Standard Deviation} = \sqrt{\frac{\sum f \cdot (X_i - \bar{X})^2}{N}}
\]

This description of dispersion can now be expressed in the original units of measurement. If the \(X_i\) values in Table 8-4 are the number of conversations with children per day, then we can say that the standard deviation of distribution A is 1.30 conversations, while that of distribution B is 1.87 conversations.

The Shape of a Distribution: Kurtosis

Earlier in this chapter we discussed skewness, the degree to which a distribution deviates from symmetry. Another way in which a distribution can be characterized is in terms of kurtosis, or whether a distribution can be described as relatively flat, or peaked, or somewhere in between.
Different shapes of distributions have different labels, and some of them are shown in Figure 8-9.

A leptokurtic distribution is a relatively tall and narrow distribution, indicating that the observations were tightly clustered within a relatively narrow range of values.

Another way of describing a leptokurtic distribution would be to say that it is a distribution with relatively little dispersion. A mesokurtic distribution reveals observations to be more distributed across a wider range of values, and a platykurtic distribution is one where proportionately fewer cases are observed across a wider range of values. We could also say that mesokurtic and platykurtic distributions are increasingly “flatter”, and have increasingly greater dispersions.

Measurement Requirements for Measures of Central Tendency and Measures of Dispersion

The choice of descriptive measures depends entirely on the level of measurement for a particular variable. To refresh your memory, the levels of measurement are:

- Nominal measurement: merely a set of mutually exclusive and exhaustive categories.
- Ordinal measurement: as in nominal, but with the addition of an underlying dimension which allows comparative statements about a larger or smaller quantity of the property being measured.
- Interval measurement: as in ordinal, with the addition of equal sized value intervals separating each of the value classes. Values are continuous, i.e. fractional values of intervals are meaningful.
- Ratio measurement: as in interval, but the scale includes an absolute zero class.

The measures of central tendency require different minimum levels of measurement. Table 8-5 below indicates whether a given central tendency statistic is appropriate for a given level of measurement.

As we mentioned in Chapter 7, the measurement decisions made while specifying the operational definition for a concept will have consequences. For instance, operationally defining a dependent variable at the ordinal level denies the possibility of computing a mean for the variable. This is important, because a statistical test which contrasts means can’t be used to establish a relationship between this variable and another one; a test which contrasts medians, however, would be appropriate. The most sensitive measure of central tendency is the mean, and it can be meaningfully computed only for interval or ratio-level data. This argues for making every attempt to obtain interval or ratio levels of measurement for variables. With interval/ratio data the widest range of statistics, and those most sensitive to subtle differences in data distributions, can be used.

The basis for computing the sum of squared deviations, the variance, and the standard deviation is an observation’s deviation from the mean. The level of measurement required for these measures of dispersion is therefore identical to the level of measurement required for computing means.

### Table 8-5 Appropriateness of Measures of Central Tendency and Dispersion for Different Levels of Measurement

<table>
<thead>
<tr>
<th></th>
<th>Mode</th>
<th>Median</th>
<th>Mean</th>
<th>Variance/Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nominal</strong></td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td><strong>Ordinal</strong></td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td><strong>Interval/Ratio</strong></td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Central Tendency and Dispersion Considered Together

There are a number of reasons why it is useful to consider the descriptions of central tendency and dispersion together.

The first one is the ability to fully describe distributions. The computation of both measures of central tendency as well as measures of dispersion allows us to describe distributions with sufficient detail to visualize their general dimensions; neither the measures of central tendency nor the measures of dispersion alone allow us to do this. For instance, Distribution A in Table 8-4 is characterized by a mean, mode and median of 5, as is Distribution B. This just tells us that the two distributions are symmetrical, and have the same central values. But adding the information that Distribution A has a standard deviation of 1.30, as compared to B's standard deviation of 1.87, tells us that the latter has a more dispersed distribution; that is, it is "wider" in appearance than Distribution A.

The z-score

A second, and extremely useful, way in which measures of central tendency and dispersion can be used together is in the standard score, also known as the z-score. A standard or z-score represents an observations' deviation from the mean, expressed in standard deviation units:

\[ z_i = \frac{X_i - \bar{X}}{St.Dev} \]

By dividing by the standard deviation, we can compare data values from distributions which vary widely in dispersion. For very dispersed distributions, the standard deviation will be large, and thus a very large deviation from the mean is required before the z-score becomes large. For distributions with data values tightly clustered around the mean (i.e., with very small standard deviations), a very small deviation from the mean can give the same standard score. Standard scores thus tell us how far an data observation is from the mean; relatively where it is positioned in the distribution.

Z-scores are useful in a number of different ways. First, from the formula above it can be readily seen that the z-score is a signed value, that is, a non-zero z-score has either a negative or a positive value. If the score to be "standardized" is larger then the mean, the deviation will be a positive one and the value of the z-score will be positive. If the score to be standardized is smaller than the mean, the resulting z-score will be negative. The sign of a z-score thus tells us immediately whether an observation is located above or below the mean. The position of an unstandardized (or "raw") data value relative to the mean can only be determined if we know the mean: a zscore communicates this information directly.

Second, the magnitude of the z-score communicates an observation's relative distance to the mean, as compared to other data values. A z-score of +.10 tells us that an observation has a positive, though relatively small, deviation from the mean, as compared to all the other data values in the distribution. It is only one tenth of a standard deviation above the mean. On the other hand, an exam grade which standardizes to a z-score of -2.30 indicates that the unfortunate recipient of that grade turned in one of the poorest performances. It's not only below the mean (a negative z-score), but it's also a very large deviation, compared to the deviation for the average student. This is indicated by the large z-score value obtained after dividing the student's deviation from the mean by the standard deviation, i.e., making the student's deviation from the mean proportional to the "average" deviation. Again, the same interpretation could be given to a raw score only if it was accompanied by both the mean and the standard deviation.

Third, we can use z-scores to standardize entire distributions. By converting each score in a distribution to a z-score we obtain a standardized distribution. Such a standardized distribution, regardless of its original values, will have a mean of 0.0 and a variance and standard deviation of 1.0. Table 8-6 below shows an example of a standardized distribution.
Table 8-6 indicates that the sum of the squared deviations in the standardized distribution is equal to $N$ (excluding rounding error), and hence that the variance and the standard deviation of a standardized distribution are equal to 1.00.

By standardizing entire distributions, we can compare very dissimilar distributions, or observations from dissimilar distributions. For instance, we can’t directly compare a student’s score on a 100-point exam, on which the class performance varied widely, to her score on a 20 point quiz, on which all the class members performed in very similar fashion. Such a comparison would be inappropriate because the central points of the two distributions are different, and so are the dispersions of the scores. By converting the exam score and the quiz score to $z$-scores, the two scores become comparable because the two distributions have become comparable: both have a mean of 0.00 and a standard deviation of 1.00. Assume that your score on a 100-point exam was 70 and that you got 9 points on a 20-point quiz. Did you perform differently on the two tasks? If we know, furthermore, that the class mean for that exam was 65, with a standard deviation of 15, and that the quiz class mean and standard deviation, respectively, were 12 and 4, we can answer that question by computing the $z$-scores $z_{\text{exam}}$ and $z_{\text{quiz}}$. These $z$-scores can be determined to be equal to

\[
z_{\text{exam}} = \frac{70 - 65}{15} = +.33
\]

\[
z_{\text{quiz}} = \frac{9 - 12}{4} = -.75
\]
The z-scores indicate although you scored a little above the mean on the exam, you were a substantial distance below the mean on the quiz.

As another example, imagine that we are interested in determining whether a person is more dependent on television than on newspapers for news, compared to the dependence of other respondents in a survey. We can contrast that person’s time spent with television news, converted to a z-score, to his newspaper-reading time z-score. If his television z-score were to be the greater, his dependence on television news would be thought to be higher.

The fourth reason why z-scores are useful is because they can be mathematically related to probabilities. If the standardized distribution is a normal distribution (a “bellshaped curve”) we can state the probability of occurrence of an observation with a given z-score. The connection between probability and z-scores forms the basis for statistical hypothesis testing and will be extensively discussed in the next three chapters.

Summary

In this chapter we have presented the most commonly used ways for describing distributions. Of these methods, visual representations of data appear to be relatively limited in their utility from a data-analytic point of view. Furthermore, the use of these methods is significantly constrained by the level of measurement of the data to be described.

Of particular interest to the communication researcher are the descriptive statistics which use mathematics to summarize a particular aspect of a data distribution. These aspects essentially center on what is typical or common in a distribution and on the amount of variability of data in the distribution. The first set of measures is called measures of central tendency; the second is called measures of dispersion.

Neither set of measures is sufficient by itself if we wish to have a complete visualization of a data distribution. For instance, contrasting the mean and the median will allow us to determine the degree of symmetry in a distribution, but we will also need such indicators as the variance or the standard deviation to fully describe the shape of the data distribution.

Of particular importance in this chapter is the z-score or standard score, which brings together both a measure of central tendency and a measure of dispersion. Not only does the z-score allow for the comparison of cases from dissimilar distributions (with differing mean and standard deviation) but it is also basic to the process of hypothesis testing and statistical inference, as it provides the link, based on probability, between a limited set of observations and other unobserved cases.

The next several chapters will provide an extensive introduction to the process of statistical inference by introducing the types of distributions we’ll need to consider and the role these distributions play in classical hypothesis testing. For these reasons it is imperative that you have a strong grasp of the materials that have been presented in this chapter.

References and Additional Readings


