Chapter four

Describing the Typical Case: Measures of Central Tendency

The mode

How is It Calculated?

What Information Does It Use?

What are Its Advantages and Disadvantages?

The median

How is It Calculated?

What Information Does It Use?

What are Its Advantages and Disadvantages?

The mean

How is It Calculated?

What Information Does It Use?

What are Its Advantages and Disadvantages?

What are Its Other Unique Properties?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_4,

THE NATURAL FIRST STEP in summarizing research is to provide a basic portrait of the characteristics of a sample or population. What is the typical case? If the researcher could choose one case to represent all others, which would it be? When a sample is very small, it is possible merely to show the array of cases and let the reader decide. However, as the number of cases grows, it becomes difficult to make a decision about typicality from the distribution as a whole. This is the function of measures of central tendency in statistics. They provide us with a simple snapshot of our data that can be used to gain a picture of the average case.

In this chapter, three commonly used measures of central tendency are discussed and compared. The first, the mode, is used primarily with nominal-level data. It is the simplest measure of central tendency, drawing information only about the frequency of events in each category. The second measure, the median, takes into account not only frequency but also the order or ranking of study subjects. Finally, the mean adds the additional factor of the exact scores associated with each subject studied. As in the discussion of levels of measurement, we emphasize in this chapter the benefits gained from statistics that use more information. But we also illustrate the importance of looking carefully at the distribution of cases in your study before deciding which measure of central tendency is most appropriate.

The Mode: Central Tendency in Nominal Scales

Faced with a nominal-scale measure, how would you define a typical case? Take as an example Table 4.1. Here you have a nominal scale of legal representation for a sample of offenders convicted of white-collar crimes in U.S. federal courts. Offenders were placed into one of five categories, indicating the type of legal representation they had: no attorney

Table 4.1

Legal Representation for White-Collar Crime

CATEGORY	FREQUENCY (*N*)
No Attorney	20
Legal Aid	26
Court Appointed	92
Public Defender	153
Private Attorney	380
Total (Σ)	671

present, legal-aid attorney, court-appointed attorney, public defender, and privately retained attorney. The number of individuals who fall in each category—or, in statistical language, the *N* of cases—is reported.

Clearly, you have very limited information in this example on which to base a choice about typicality. Here, as with other nominal-scale measures, you simply know how many cases fall into one category or another. You would probably choose the category "private attorney" as most representative of this sample, because it contains by far the most cases (380). And indeed, this is precisely how statisticians define typicality for nominal-level variables. We call the category with the largest *N,* or number of cases, the **mode.** In this sample of white-collar offenders, the modal category for type of representation is "private attorney."

By defining one category as the modal category, we are able to provide a summary of the type of case that is typical of our sample or population. Such statements are common in criminal justice research. We often are interested in the racial category that appears most often in our data or the type of offense that is most common. The modal category can also provide a basis for making comparisons among samples. For example, let's say that a sample of offenders convicted of nonviolent property crimes that would not ordinarily be defined as white collar was compared to this larger sample of offenders convicted of white-collar crimes. For the former group, as is apparent from Table 4.2, the modal category is not "private attorney" but rather "court-appointed attorney." Although this comparison of the two samples is not a complex one, it

Table 4.2

Legal Representation for Common Crime

CATEGORY	FREQUENCY (*N*)
No Attorney	40
Legal Aid	7
Court Appointed	91
Public Defender	22
Private Attorney	70
Total (Σ)	230

Table 4.3 Financial Harm for a Sample of Convicted Offenders

CATEGORY	FREQUENCY (*N*)
Less than $100	15
$101–$2,500	92
$2,501–$10,000	20
More than $10,000	19
Total (Σ)	146

illustrates the different backgrounds of the two groups. White-collar offenders are much more likely than common criminals to have the resources to pay for private legal representation.

In general, we do not use the mode to describe central tendency with ordinal or interval scales. The reason, in good part, is that the mode does not take advantage of the additional information that such scales provide. The average case should not be chosen simply on the basis of the frequency of events in a particular category, because higher-level scales also provide information on the order or nature of the differences between categories.

Nonetheless, there are cases where researchers choose to use the mode to describe ordinal- or interval-level measures. Generally this occurs when there is a very large group of cases in one particular category. Table 4.3, for example, provides an ordinal-level measure of the financial harm caused by a sample of convicted offenders. Because almost two-thirds of the individuals studied fall in the category "$101–$2,500," you might want to describe typicality in this case by saying that this category is the modal category. Similarly, if you were examining prior arrests and two-thirds of the offenders in your sample had no prior arrests, you might want to report no arrests as the modal category. Even though this measure is an interval measure, the mode provides a fairly good summary of the typical case in your sample.

The Median: Taking into Account Position

In constructing the **median,** we utilize information not only on the number of cases found in a particular category, but also on the positions of the categories. The median may be defined simply as the middle score in a distribution. For ordinal scales, it is the category in which the middle score lies. For interval scales, the median is the value that splits the distribution of scores in half.

There are two general steps in determining the median for a distribution of scores. First, the values need to be arranged from low to high

Table 4.4 Student Views on Public Drunkenness

CATEGORY	FREQUENCY (*N*)	CUMULATIVE *N*
Not serious at all	73	73
A bit serious	47	120
Somewhat serious	47	167
Serious	27	194
Very serious	26	220
Extremely serious	39	259
Most serious	22	281
Total (Σ)	281	

scores. As we saw in Chapter 3, a frequency distribution allows us to represent our data in this way. Table 4.4 presents a frequency distribution of views of public drunkenness, drawn from a survey of students. The students were presented with an ordinal-scale measure that allowed them to rate the seriousness of a series of crimes. The ratings ranged from "not serious at all" to "most serious."

Second, we need to determine which observation splits the distribution. A simple formula, Equation 4.1, allows us to define which observation is the median when the number of observations in the distribution is odd, as is the case with our example of views of public drunkenness.

$$\text{Median observation} = \frac{N+1}{2}$$

Equation 4.1

In this case, we add 1 to the total number of observations in the sample or population we are studying and then divide by 2. For the frequency distribution in Table 4.4, the median observation is the 141st score:

Working It Out

$$\begin{aligned}\text{Median observation} &= \frac{N+1}{2}\\ &= \frac{281+1}{2}\\ &= 141\end{aligned}$$

However, because our variable, student views on drunkenness, is measured on an ordinal scale, it does not make sense to simply state that the 141st observation is the median score. To give a substantive meaning to the median, it is important to define which category the median score falls in. The 141st observation in our distribution of ordered scores falls in the category labeled “somewhat serious.”

The advantage of the median over the mode for describing ordinal scales is well illustrated by our example of views of public drunkenness. If we used the mode to describe typicality in student assessments of the seriousness of public drunkenness, we would conclude that the typical student did not see drunkenness as at all serious. But even though the “not serious at all” category includes the largest number of cases, almost three-quarters of the students rate this behavior more seriously. The median takes this fact into consideration by placing the typical case in the middle of a distribution. It is concerned with not only the number of cases in the categories, but also their position.

If the number of observations or cases in your distribution is even, then you cannot identify a single observation as the median. While statisticians recognize that the median is ambiguously defined in this case, by convention they continue to use Equation 4.1 to identify the median for an ordinal-level distribution. In practice, this places the median score between two observations. For example, consider the distribution of 146 scores in Table 4.3, representing financial harm in a sample of offenders. Here the number of scores is even, and thus there is not a single observation that can be defined as the median. Using Equation 4.1, we can see that the median is defined as the halfway point between the 73rd and the 74th observation. This means that the median falls in the category defined as \$101 to \$2,500.[1]

Working It Out

$$\begin{aligned} \text{Median observation} &= \frac{N+1}{2} \\ &= \frac{146+1}{2} \\ &= 73.5 \end{aligned}$$

73rd observation: \$101–\$2,500

74th observation: \$101–\$2,500

The median is sometimes used for defining typicality with interval scales. For example, Table 4.5 presents the average number of minutes of public disorder (per 70-minute period) observed in a sample of 31

[1] With this method, it is possible that the defined median value will fall between two categories of an ordinally measured variable. In that case, you simply note that the median falls between these two categories.

Table 4.5 Hot Spots: Minutes of Public Disorder (A)

HOT SPOT SCORE	FREQUENCY (*N*)	CUMULATIVE (*N*)
0.35	1	1
0.42	1	2
0.46	1	3
0.47	1	4
0.52	1	5
0.67	1	6
1.00	1	7
1.06	1	8
1.15	1	9
1.19	2	11
1.48	1	12
1.60	1	13
1.63	1	14
2.02	1	15
2.12	1	16
2.21	1	17
2.34	1	18
2.45	1	19
2.66	1	20
3.04	1	21
3.19	1	22
3.23	1	23
3.46	1	24
3.51	1	25
3.72	1	26
4.09	1	27
4.47	1	28
4.64	1	29
4.65	1	30
6.57	1	31
Total (Σ)	31	31

"hot spots of crime," or city blocks with high levels of crime. The hot spots are arranged in ascending order on the basis of the number of minutes of disorder observed. In this case, the distribution has an odd number of observations, and thus the median is the score in the middle of the distribution, or the 16th observation, which has a value of 2.12.

Working It Out

$$\text{Median observation} = \frac{N + 1}{2}$$

$$= \frac{31 + 1}{2}$$

$$= 16$$

Accordingly, using the median, we would describe the average hot spot as having a little more than two minutes of disorder in each 70-minute period.

As noted above, when the number of observations in a distribution is even, the median is ambiguously defined. Let's, for example, delete the hot spot with a score of 6.57 from Table 4.5. In this case, there is no single middle value for the array of cases in the table. If we use Equation 4.1 to define the median observation, we get a value of 15.5.

Working It Out

$$\text{Median observation} = \frac{N+1}{2} = \frac{30+1}{2} = 15.5$$

But what is the value or score associated with an observation that lies between two scores in an interval-level scale? If both the 15th and the 16th observation are in the same category, then the solution is easy. You simply define the median as the score associated with both the 15th and the 16th observation. However, it will sometimes be the case with an interval-level variable that each of these observations will have a different value on the scale, as we find here. There is no true median value for this example. By convention, however, we define the median with interval-level measures as the midpoint between the observation directly below and the observation directly above the median observation. In our example, this is the midpoint on our scale between the scores 2.02 and 2.12. The median in this case is defined as 2.07.[2]

Working It Out

$$\text{15th case} = 2.02$$

$$\text{16th case} = 2.12$$

$$\text{Median} = \frac{2.02 + 2.12}{2} = 2.07$$

[2] Sometimes the median for ordinal-level variables is also calculated using this method. In such cases, the researcher should realize that he or she is treating the variable under consideration as an interval-level measure. Only for an interval-level measure can we assume that the units of measurement are constant across observations.

The median is generally more appropriate than the mode for assessing central tendency for both ordinal- and interval-level measures. However, the median does not take advantage of all the information included in interval-level scales. Although it recognizes the positions of the values of a measure, it does not take into account the exact differences among these values. In many cases, this can provide for a misleading estimate of typicality for interval-level measures.

For example, let's say that the distribution of disorder in hot spots is that represented in Table 4.6. In this case, the median is 1.83. But is 1.83 a good estimate of central tendency for this distribution? The 17th score is 3.34, which is not very close to 1.83 at all. The score of 1.83 is not an ideal estimate of typicality, as it is far below half the scores in the distribution. The median is not sensitive to the gap in our measure between the values of the 16th and 17th cases. This is because it looks only at

Table 4.6 Hot Spots: Minutes of Public Disorder (B)

HOT SPOT SCORE	FREQUENCY (*N*)	CUMULATIVE (*N*)	CUMULATIVE %
0.35	1	1	3.2
0.42	1	2	6.5
0.46	1	3	9.7
0.47	1	4	12.9
0.52	1	5	16.1
0.67	1	6	19.4
1.00	1	7	22.6
1.06	1	8	25.8
1.15	1	9	29.0
1.19	2	11	35.5
1.48	1	12	38.7
1.60	1	13	41.9
1.63	1	14	45.2
1.73	1	15	48.4
1.83	1	16	51.6
3.34	1	17	54.9
3.44	1	18	58.1
3.45	1	19	61.3
3.66	1	20	64.5
4.04	1	21	67.7
4.19	1	22	71.0
4.23	1	23	74.2
4.46	1	24	77.4
4.51	1	25	80.6
4.72	1	26	83.9
5.09	1	27	87.1
5.47	1	28	90.3
5.64	1	29	93.5
5.65	1	30	96.8
5.57	1	31	100.0
Total (Σ)	31	31	100.0

position and not at the size of the differences between cases. The median does not take advantage of all the information provided by interval-level measures.

Another way to describe the median in interval-level measures is to say that it is the 50th percentile score. A percentile score is the point or score below which a specific proportion of the cases is found. The 50th percentile score is the score below which 50% of the cases in a study lie. For the data in Table 4.6, if we defined the median in this way, we again choose 1.83 as the median minutes of disorder observed in the hot spots. In this case, if we add the percentage of cases for all of the scores up until the middle, or 16th, score, we come to a total (or cumulative percentage) of 51.6. At the 15th score, or 1.73, the cumulative percentage is only 48.4, less than 50%.

The Mean: Adding Value to Position

The **mean** takes into account not only the frequency of cases in a category and the positions of scores on a measure, but also the values of these scores. To calculate the mean, we add up the scores for all of the subjects in our study and then divide the total by the total number of subjects. In mathematical language, the mean can be written as a short equation:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \textbf{Equation 4.2}$$

Even though equations sometimes put students off, they are an important part of statistics. Indeed, equations are the language of statistics. They show how a statistic is constructed and the method we use to calculate it. Equations provide a short way of writing out what would often take a number of sentences to describe in English. One of our tasks in this text is to help you to translate such equations and to become more comfortable with them.

In the case of the mean, we introduce what are for most students of criminal justice some new symbols and concepts. First, to express the mean, statisticians provide us with a shorthand symbol, $\overline{X}$—in English, "X bar." The equation also includes the summation symbol, Σ. Under the symbol is $i = 1$, and above it is N. What this means is that you should start summing your cases with the first subject in the sample and end

Table 4.7 Total Number of Prior Arrests

TOTAL NUMBER OF ARRESTS	FREQUENCY (*N*)	CUMULATIVE (*N*)
0	4	4
1	1	5
2	2	7
4	3	10
5	3	13
7	4	17
8	2	19
10	1	20

with the last one (represented by *N* because, as we have already discussed, *N* is the number of cases in your sample). But what should you sum? *X* represents the measure of interest—in the case of our example, minutes of disorder. We use the subscript *i* to denote each of the observations of the variable *X*. If, for example, we wrote X_3, we would be referring only to the 3rd observation of the variable. So Equation 4.2 says that you should sum the scores for minutes of disorder from the first to the last case in your study. Then you should divide this number by the total number of cases.

Table 4.7 presents information about the total number of prior arrests for a sample of 20 individuals arrested for felony offenses. To calculate the mean, we first sum all of the scores, as shown in the numerator of Equation 4.2:

Working It Out

$$\begin{aligned}\sum_{i=1}^{N} X_i &= \sum_{i=1}^{20} X_i \\ &= 0 + 0 + 0 + 0 + 1 + 2 + 2 + 4 + 4 + 4 \\ &\quad + 5 + 5 + 5 + 7 + 7 + 7 + 7 + 8 + 8 + 10 \\ &= 86\end{aligned}$$

We then take the sum of the values, 86, and divide by the number of observations in the sample.

Working It Out

$$
\begin{aligned}
\overline{X} &= \frac{\sum_{i=1}^{N} X_i}{N} \\
&= \frac{86}{20} \\
&= 4.3
\end{aligned}
$$

The result, 4.3, tells us that in this sample the typical person arrested for a felony has, on average, 4.3 prior arrests.

As another example, let's take the data from Table 4.5 on minutes of disorder in crime hot spots. According to Equation 4.2, the first step is to sum all of the scores:

Working It Out

$$
\begin{aligned}
\sum_{i=1}^{N} X_i = \sum_{i=1}^{31} X_i
&= 0.35 + 0.42 + 0.46 + 0.47 + 0.52 + 0.67 + 1.00 + 1.06 \\
&\quad + 1.15 + 1.19 + 1.19 + 1.48 + 1.60 + 1.63 + 2.02 + 2.12 \\
&\quad + 2.21 + 2.34 + 2.45 + 2.66 + 3.04 + 3.19 + 3.23 + 3.46 \\
&\quad + 3.51 + 3.72 + 4.09 + 4.47 + 4.64 + 4.65 + 6.57 \\
&= 71.56
\end{aligned}
$$

We then take this number, 71.56, and divide it by *N*, or 31, the number of cases in our sample.

Working It Out

$$
\begin{aligned}
\overline{X} &= \frac{\sum_{i=1}^{N} X_i}{N} \\
&= \frac{71.56}{31} \\
&= 2.308387097
\end{aligned}
$$

The result, 2.308387097 (rounded to the ninth decimal place), brings up an issue that often arises in reporting statistics. Do you really need to provide your audience with the level of precision that is given by your statistic? In this case, for example, at what level of precision should minutes of disorder be presented?

A basic rule of thumb is to use your common sense in answering such questions. Don't provide statistics developed out to a large number of decimal places just to impress others. In making this decision, you should ask: What is the simplest presentation of my results that will provide the reader or listener with enough information to understand and evaluate my work? Overall, criminal justice researchers seldom report the mean to more than two decimal places. This is a good choice in our example. Rounding to the second decimal place gives a mean of 2.31. Providing a more precise representation of the mean here would not add important information for the reader.

In some cases, it is useful to develop estimates with much greater precision. In particular, if the values for the cases you are examining are very small in the first place, you will want to present a more precise mean. For example, Lawrence Sherman and his colleagues looked at the mean daily rate of reported domestic violence in a study that compared the impact of arrests versus warnings as a strategy for controlling spouse abusers.[3] Had they reported their findings only to the second decimal place, as recommended above, they would have ended up with a mean daily rate over the longest follow-up period (361–540 days) of 0.00 for short arrest and 0.00 for warning. The difficulty here is that individuals are unlikely to report cases of domestic violence on a very frequent basis. Sherman et al. needed a much higher degree of precision to examine the differences between the two groups they studied. Accordingly, they reported their results to the fourth decimal place. For arrests, the rate was 0.0019, and for warnings it was 0.0009. These differences, though small, were found to be meaningful in their research.

Comparing Results Gained Using the Mean and Median

Returning to the example from Table 4.5, we see that the mean for minutes of disorder, 2.31, is very similar to the median of 2.12 calculated earlier. In this case, adding knowledge about value does not change our portrait of the typical hot spot very much. However, we get a very different sense of the average case if we use the data from Table 4.6. Here, the median provided a less than satisfying representation of the average

[3] L. Sherman, J. D. Schmidt, D. Rogan, P. Gartin, E. G. Cohn, D. J. Collins, and A. R. Bacich, "From Initial Deterrence to Long-Term Escalation: Short-Custody Arrest for Poverty Ghetto Domestic Violence," *Criminology* 29 (1991): 821–850.

case. It was not sensitive to the fact that there was a large gap in the scores between the 16th and 17th cases. Accordingly, the median, 1.83, was very close in value to the first half of the cases in the sample, but very far from those hot spots with higher values. The mean should provide a better estimate of typicality here, because it recognizes the actual values of the categories and not just their positions. Let's see what happens when we calculate the mean for Table 4.6.

Following our equation, we first sum the individual cases:

W orking It Out

$$\sum_{i=1}^{N} X_i = \sum_{i=1}^{31} X_i$$

$$\begin{aligned} &= 0.35 + 0.42 + 0.46 + 0.47 + 0.52 + 0.67 + 1.00 + 1.06 \\ &\quad + 1.15 + 1.19 + 1.19 + 1.48 + 1.60 + 1.63 + 1.73 + 1.83 \\ &\quad + 3.34 + 3.44 + 3.45 + 3.66 + 4.04 + 4.19 + 4.23 + 4.46 \\ &\quad + 4.51 + 4.72 + 5.09 + 5.47 + 5.64 + 5.65 + 5.77 \\ &= 84.41 \end{aligned}$$

We then divide this number by the total number of cases:

W orking It Out

$$\begin{aligned} \overline{X} &= \frac{\sum_{i=1}^{N} X_i}{N} \\ &= \frac{84.41}{31} \\ &= 2.7229 \end{aligned}$$

Here, we gain an estimate of typicality of 2.72 (rounding to the second decimal place). As you can see, this score is much better centered in our distribution than is the median. The reason is simple. The median does not take into account the values of the categories. The mean does take value into account and thus is able to adjust for the gap in the distribution.

There are cases in which the sensitivity of the mean to the values of the categories in a measure can give misleading results. For example, let's say that one case in your study is very different from the others. As noted in Chapter 1, researchers call such a case an **outlier,** because it is very much outside the range of the other cases you studied. Taking the example of minutes of disorder from Table 4.5, let's say that the last case had 70 minutes of disorder (the maximum amount possible) rather than 6.57 minutes. When we calculate the mean now, the sum of the cases is much larger than before:

Working It Out

$$\sum_{i=1}^{N} X_i = \sum_{i=1}^{31} X_i$$

$$\begin{aligned} &= 0.35 + 0.42 + 0.46 + 0.47 + 0.52 + 0.67 + 1.00 + 1.06 \\ &\quad + 1.15 + 1.19 + 1.19 + 1.48 + 1.60 + 1.63 + 2.02 + 2.12 \\ &\quad + 2.21 + 2.34 + 2.45 + 2.66 + 3.04 + 3.19 + 3.23 + 3.46 \\ &\quad + 3.51 + 3.72 + 4.09 + 4.47 + 4.64 + 4.65 + 70.0 \\ &= 134.99 \end{aligned}$$

Dividing this sum by the total number of cases provides us with a mean of 4.35 (rounded to the second decimal place):

Working It Out

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

$$= \frac{134.99}{31}$$

$$= 4.3545$$

The mean we calculated with the original score was 2.31 (see page 76). Accordingly, merely by changing one score to an outlier, we have almost doubled our estimate of typicality. In this case, the sensitivity of the mean to an extreme value in the distribution led it to overestimate the average case. This illustrates the general principle that the mean is sensitive to outliers. Because the mean is used to develop many other more complex statistics, this principle is relevant not only to the mean itself but also to a number of other important statistical techniques used by researchers.

So what should you do if outliers lead to a misleading conclusion regarding typicality in your study? One solution is simply to exclude the outliers from specific analyses and let your readers or audience know that some cases have been excluded and why. If the number of extreme cases is large enough, you may want to analyze these cases separately. Another solution is to transform the outliers. That is, you may want to replace them with values closer to the rest of the distribution (e.g., the highest value that is not an outlier). In this way, you can include the cases, but minimize the extent to which they affect your estimate of typicality. However, you should be cautious in developing such transformations of your scores, keeping in mind that you are changing the character of the distribution examined in your study.

Other Characteristics of the Mean

Two other traits of the mean are important because they play a role in how we develop other statistics. The first concerns what happens when we look at **deviations** (or differences) **from the mean.** This will become an issue in the next chapter, when we discuss measures of dispersion. The second, often termed the **least squares property** of the mean, will become important to us in Chapter 15, when we discuss regression.

If we take each score in a distribution, subtract the mean from it, and sum these differences, we will always get a result of 0. In equation form, this principle is represented as follows:

$$\sum_{i=1}^{N} (X_i - \overline{X}) = 0 \qquad \textbf{Equation 4.3}$$

In English, this equation says that if we sum the deviations from the mean, from the first to the last case, we will always get a result of 0. This principle is illustrated in Table 4.8, using the data on minutes of public disorder from Table 4.5. Here we have taken the 31 scores and subtracted the mean from each one. We then added these differences. Because the positive scores balance out the negative ones, the result is 0. This will always happen when we use the mean.

The second trait, the least squares property, is very important for understanding regression analysis (introduced in Chapter 15), a technique commonly used for describing relationships among variables in criminal justice. For the moment, it is enough to note this fact and that the issues

Table 4.8 Deviations from the Mean for Minutes of Public Disorder (A)

SCORE (X)	DEVIATION FROM THE MEAN ($X_i - \overline{X}$)
0.35	0.35 − 2.31 = −1.96
0.42	0.42 − 2.31 = −1.89
0.46	0.46 − 2.31 = −1.85
0.47	0.47 − 2.31 = −1.84
0.52	0.52 − 2.31 = −1.79
0.67	0.67 − 2.31 = −1.64
1.00	1.00 − 2.31 = −1.31
1.06	1.06 − 2.31 = −1.25
1.15	1.15 − 2.31 = −1.16
1.19	1.19 − 2.31 = −1.12
1.19	1.19 − 2.31 = −1.12
1.48	1.48 − 2.31 = −0.83
1.60	1.60 − 2.31 = −0.71
1.63	1.63 − 2.31 = −0.68
2.02	2.02 − 2.31 = −0.29
2.12	2.12 − 2.31 = −0.19
2.21	2.21 − 2.31 = −0.10
2.34	2.34 − 2.31 = 0.03
2.45	2.45 − 2.31 = 0.14
2.66	2.66 − 2.31 = 0.35
3.04	3.04 − 2.31 = 0.73
3.19	3.19 − 2.31 = 0.88
3.23	3.23 − 2.31 = 0.92
3.46	3.46 − 2.31 = 1.15
3.51	3.51 − 2.31 = 1.20
3.72	3.72 − 2.31 = 1.41
4.09	4.09 − 2.31 = 1.78
4.47	4.47 − 2.31 = 2.16
4.64	4.64 − 2.31 = 2.33
4.65	4.65 − 2.31 = 2.34
6.57	6.57 − 2.31 = 4.26
Total (Σ)	0*

*Because of rounding error, the actual column total is slightly less than zero.

we address early on in statistics are often the bases for much more complex types of analysis. "Don't forget the basics" is a good rule. Many mistakes that researchers make in developing more complex statistics come from a failure to think about the basic issues raised in the first few chapters of this text.

The least squares property is written in equation form as follows:

$$\sum_{i=1}^{N} (X_i - \overline{X})^2 = \text{minimum} \qquad \textbf{Equation 4.4}$$

What this says in English is that if we sum the squared deviations from the mean for all of our cases, we will get the minimum possible result. That is, suppose we take each individual's score on a measure, subtract the mean from that score, and then square the difference. If we then sum all of these values, the result we get will be smaller than the result we would have gotten if we had subtracted any other number besides the mean. You might try this by calculating the result for minutes of disorder using the mean. Then try other values and see if you can find some other number of minutes that will give you a smaller result. The least squares property says you won't.

Using the Mean for Noninterval Scales

The mean is ordinarily used for measuring central tendency only with interval scales. However, in practice, researchers sometimes use the mean with ordinal scales as well. Is this wrong? In a pure statistical sense, it is. However, some ordinal scales have a large number of categories and thus begin to mimic some of the characteristics of interval-level measures.

This is particularly true in cases where the movements from one category to another in an ordinal scale can be looked at as equivalent, no matter which category you move from. Taking our example of student attitudes toward public drunkenness in Table 4.4, a researcher might argue that the difference between "somewhat serious" and "a bit serious" is about equivalent to that between "very serious" and "extremely serious," and so forth. Thus, the difference between these categories is not just a difference of position; it is also a movement of equal units up the scale. Taking this approach, we can say that this measure takes into account both position and value, although the values here are not as straightforward as those gained from true interval scales such as number of crimes or dollar amount stolen.

A researcher might argue that the mean is appropriate for presenting findings on views of public drunkenness because this ordinal-scale measure of attitudes is like an interval-scale measure. Although it is easy to see the logic behind this decision, it is important to note that such a decision takes a good deal of justification. In general, you should be very cautious about using the mean for ordinal-level scales, even when the above criteria are met.

Statistics in Practice: Comparing the Median and the Mean

The general rule is that the mean provides the best measure of central tendency for an interval scale. This follows a principle stated in Chapter 1: In statistics, as in other decision-making areas, more information is better than less information. When we use the mean, we take into account not only the frequency of events in each category and their position, but also the values or scores of those categories. Because more information is used, the mean is less likely than other measures of central tendency to be affected by changes in the nature of the sample that a researcher examines. It is useful to note as well that the mean has some algebraic characteristics that make it more easily used in developing other types of statistics.

The mean is generally to be preferred, but when the distribution of a variable is strongly **skewed,** the median provides a better estimate of central tendency than the mean. "Skewed" means that the scores on the variable are very much weighted to one side and that frequencies of extreme values trail off in *one* direction away from the main cluster of cases. A distribution that has extreme values lower than the main cluster of observations (i.e., there is a "tail" to the left in the distribution) is said to be negatively skewed, while a distribution that has extreme values greater than the main cluster of observations (i.e., there is a "tail" to the right in the distribution) is said to be positively skewed.[4]

A good example of a skewed distribution in criminal justice is criminal history as measured by self-reports of prisoners. Horney and Marshall, for example, reported results on the frequency of offending for a sample of prisoners.[5] As is apparent from Figure 4.1, most of the offenders in their sample had a relatively low offending rate—between 1 and 20 offenses in the previous year. But a number of offenders had rates of more than 100, and a fairly large group had more than 200. The mean for this distribution is 175.

Clearly, 175 offenses provides a misleading view of typical rates of offending for their sample. Because the mean is sensitive to value, it is inflated by the very high frequency scores of a relatively small proportion of the sample. One solution suggested earlier to the problem of outliers

[4] A formal statistic for measuring the degree of skewness of a distribution is given by the following equation:

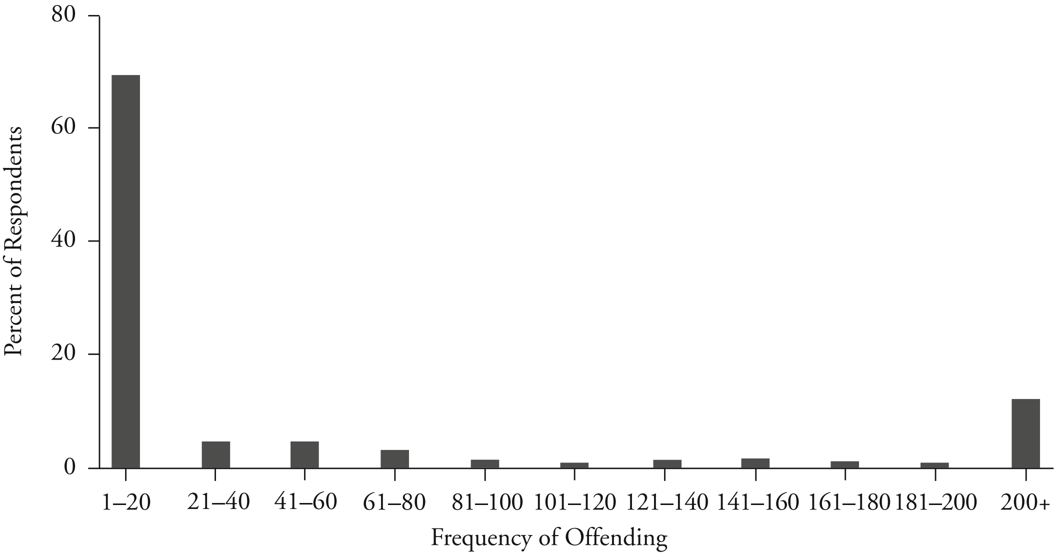
$$\text{skewness} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^3}{Ns^3}$$

In words, this equation tells us to take the deviation between a value and the mean and cube it, then sum these values over all observations; the sum of the cubed deviations is then divided by the sample size (*N*) multiplied by the standard deviation cubed. The measure of skewness will have a value of 0 if the distribution is symmetrical, a negative value if the distribution is negatively skewed, and a positive value if the distribution is positively skewed. The greater the value of the measure, the greater the degree of positive or negative skewness.

[5] J. Horney and I. H. Marshall, "An Experimental Comparison of Two Self-Report Methods for Measuring Lambda," *Journal of Research in Crime and Delinquency* 29 (1992): 102–121.

Figure 4.1

Individual Frequency of Offending for a Sample of Offenders: A Case Where the Mean Is a Misleading Measure of Central Tendency



was to exclude such cases. But here, this would mean excluding almost 30% of the sample. Thus, these are not outliers in the traditional sense. Another option mentioned earlier is to analyze the "outliers" separately. But again, there is quite a spread of scores even if we look at those above 50 or 100 separately, and the analysis of the outliers might in itself provide a misleading view of central tendency. A common solution used for describing this type of skewed interval-level distribution is to use the median rather than the mean to describe central tendency. The median for this distribution is 4, which is certainly more representative of the average case than is the mean. But even if you choose this solution, it is very important to note to your audience that the distribution is skewed and to tell them a bit about the nature of the distribution.

How should you decide when a distribution is so skewed that it is preferable to use the median as opposed to the mean? You should begin by comparing the mean and the median. When there is a very large difference between them, it may be the result of skewness. In such cases, you should look at the distribution of the scores to see what is causing the mean and median to differ widely. But there is no solid boundary line to guide your choice.

In extreme cases (such as that of criminal history in our example) or cases where the mean and median provide relatively close estimates, your choice will be clear. In the former case you would choose the median, and in the latter the mean. However, when your results fall somewhere in between, you will have to use common sense and the experiences of other researchers working with similar data as guidelines. What seems to make sense? What have other researchers chosen to do? One way of being fair to your audience is to provide results for both the mean and the median, irrespective of which you choose as the best measure of typicality.

Chapter Summary

The **mode** is calculated by identifying the category that contains the greatest number of cases. It may be applied to any scale of measurement. Because the mode uses very little information, it is rarely used with scales of measurement higher than the nominal scale. It can occasionally serve as a useful summary tool for higher-level scales, however, when a large number of cases are concentrated in one particular category.

The **median** is calculated by locating the middle score in a distribution and identifying in which category it falls. It is also known as the 50th percentile score, or the score below which 50% of the cases lie. The information used includes both the number of cases in a particular category and the positions of the categories. The median uses more information than does the mode and requires a scale of measurement that is at least ordinal in magnitude.

The **mean** is calculated by dividing the sum of the scores by the number of cases. The information used includes not only the number of cases in a category and the relative positions of the categories, but also the actual value of each category. Such information normally requires at least an interval scale of measurement. For this reason, the researcher should be cautious about using the mean to describe an ordinal scale. The mean uses more information than the mode and the median. It is, however, sensitive to extreme cases—**outliers.** Faced with the distorting effect of outliers, the researcher may choose to keep them, to transform them to other values, or to delete them altogether. If a distribution of scores is substantially **skewed,** then it may be more appropriate to use the median than to use the mean.

The sum derived by adding each score's **deviation from the mean** will always be 0. If the deviation of each score from the mean is squared, then the sum of these squares will be less than it would be if any number other than the mean were used. This is called the **least squares property.**

Key Terms

deviation from the mean The extent to which each individual score differs from the mean of all the scores.

least squares property A characteristic of the mean whereby the sum of all the squared deviations from the mean is a minimum—it is lower than the sum of the squared deviations from any other fixed point.

mean A measure of central tendency calculated by dividing the sum of the scores by the number of cases.

median A measure of central tendency calculated by identifying the value or category of the score that occupies the middle position in the distribution of scores.

mode A measure of central tendency calculated by identifying the score or category that occurs most frequently.

outlier(s) A single or small number of exceptional cases that substantially deviate from the general pattern of scores.

skewed Describing a spread of scores that is clearly weighted to one side.

Symbols and Formulas

X	Individual score
$\overline{X}$	Mean
N	Number of cases
Σ	Sum

To calculate the median observation:

$$\text{Median observation} = \frac{N + 1}{2}$$

To calculate the mean:

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

To show how the sum of the deviations from the mean equals 0:

$$\sum_{i=1}^{N} (X_i - \overline{X}) = 0$$

To express the least squares property:

$$\sum_{i=1}^{N} (X_i - \overline{X})^2 = \text{minimum}$$

Exercises

4.1 Drivers cited for moving violations are required by a state's laws to take a driving safety course taught by the local police department. The sign-in sheet asks individuals to note why they received a ticket. The 14 participants at a recent class noted the following:

Speeding, Running a red light, Running a stop sign, Speeding, Speeding, Running a red light, Tailgating, Speeding, Running a red light, Recklessness, Speeding, Running a red light, Speeding, Running a stop sign

a. Categorize these data and calculate the mode.

b. Explain why the median would not be an appropriate measure of central tendency for these data.

4.2 Calculate the mode, median, and mean for the following data:

a. Number of previous employments held by 25 convicts:

3 3 1 1 0 1 0 2 1 0 8 4 3
1 2 1 9 0 1 7 0 7 2 0 1

b. Weeks of training undergone by 20 prison guards:

10 16 12 16 16 16 10 8 10 12
16 18 12 16 16 8 0 12 10 16

c. Height (in meters) of 30 convicts:

1.72 1.78 1.73 1.70 1.81 1.64 1.76 1.72 1.75 1.74
1.88 1.79 2.01 1.80 1.77 1.79 1.69 1.74 1.75 1.66
1.77 1.73 1.72 1.91 1.80 1.74 1.72 1.82 1.86 1.79

4.3 A researcher checked the response times of police to ten emergency telephone calls. The data below record the number of minutes that elapsed from when the telephone call ended to when the police arrived:

24 26 14 27 198 22 27 17 19 29

a. Calculate the mode, the median, and the mean.

b. Which of these measures is the most suitable for this particular case? Explain your choice.

4.4 Airport officials wished to check the alertness of their security officers over the two busiest weeks of the summer. During this period, they sent out 50 undercover staff carrying suspicious items of hand luggage. Five of them were stopped at the entrance to the airport. Six made it into the airport, but were stopped at check-in. Thirteen more got into the airport and through check-in, only to be stopped at the hand-luggage inspection point. Two passed the airport entrance, check-in, and hand-luggage inspection, but were stopped when presenting their boarding cards at the gate. Four people made it past every one of these stages, only to be stopped when boarding the plane. Twenty of the undercover staff were not detected at all.

a. Categorize the data and calculate the median category.

b. Is the median a good measure of central tendency in this case? Explain your answer. If you think it is not, suggest an alternative and explain why.

4.5 On the first day of the term in a statistics course, the professor administered a brief questionnaire to the students, asking how many statistics courses they had ever taken before the current term. Of the 33 students who answered the question, 17 said none, 9 said one, 3 said two, 2 said three, 1 said four, and 1 said five.

a. Calculate the mode, median, and mean for number of prior statistics classes.

b. Which one of these measures of central tendency best measures the typicality of these data?

4.6 As part of her undergraduate thesis, a criminal justice student asked ten other criminal justice majors to rate the fairness of the criminal justice system. The students were asked to say whether they strongly agreed, agreed, were uncertain, disagreed, or strongly disagreed with the following statement: "The criminal justice system in our country treats all defendants fairly." The ten responses were

Strongly agree, Strongly agree, Strongly disagree, Strongly disagree, Uncertain, Disagree, Disagree, Agree, Strongly disagree, Uncertain

a. Categorize these data and calculate an appropriate measure of central tendency.

b. Explain why this measure of central tendency best represents the typicality of these data.

4.7 There are five prisoners in the high-security wing in a prison—Albert, Harry, Charlie, Dave, and Eddie. Only Eddie's biographical details have been lost. The information available is as follows:

	Age	Previous Convictions
Albert	23	1
Harry	28	4
Charlie	18	1
Dave	41	1
Eddie	?	?

a. Can we compute any of the following for previous convictions of the five prisoners: the mode, the median, or the mean? If any (or all) of these three measures may be calculated, what are their values?

b. If the mean number of previous convictions is 2.0, how many convictions does Eddie have?

c. If we know that the median age for the five prisoners is 28, what does this tell us about Eddie's age? Explain why.

d. If the mean age for the five prisoners is 28.2, how old is Eddie?

4.8 A researcher sat on a bench along the main shopping street of a city center on ten successive Saturdays from 11:00 A.M. to 2:00 P.M.—the three busiest shopping hours of the day—and recorded the number of times a police officer passed by. The results for the ten weeks are as follows:

Week No.:	1	2	3	4	5	6	7	8	9	10
No. of Police Officers Observed:	4	3	6	4	4	5	4	35	3	5

On week 8, local unionists held a demonstration in the city center, and the high number of observations for that week can be explained by the extra officers called in to police the rally.

a. Calculate the mode, median, and mean for the number of police officers observed.

b. Which measure of central tendency best represents typicality for these data? Discuss the issues involved in choosing the most appropriate means of describing the data.

c. Imagine that the unionists had decided to hold a regular demonstration in the city center on alternating weeks. The results recorded for the same study would be as follows:

Week No.:	1	2	3	4	5	6	7	8	9	10
No. of Police Officers Observed:	4	30	6	31	6	52	4	35	4	34

Would the measure of central tendency you recommended in part b still be the best measure of typicality? Explain why.

4.9 On a recent evening, a police crackdown on prostitution solicitation resulted in 19 arrests. The ages of the persons arrested were

17 18 24 37 32 49 61 20 21 21
25 24 24 26 30 33 35 22 19

a. Calculate an appropriate measure of central tendency.

b. Explain why this measure of central tendency best represents typicality for these data.

4.10 Using your answers from part a of Exercise 4.5, calculate

a. The sum of the deviations from the mean.

b. The sum of the squared deviations from the mean.

c. The sum of the squared deviations from the median.

d. The sum of the squared deviations from the mode.

e. Which of these sums of squared deviations has the smallest value?

Computer Exercises

Measures of central tendency are reported along with many other statistics in many of the software programs you might encounter. The commands that we describe below will be the same as those that we highlight at the end of Chapter 5.

SPSS

There are two primary ways of obtaining measures of central tendency in SPSS. The quickest way to obtain information on the mean for one or more variables measured at the interval level of measurement is to use the DESCRIPTIVES command

```
DESCRIPTIVES VARIABLES=variable_names.
```

The output window will contain quite a bit of information for each variable that you have named—much of it will make more sense after reading Chapter 5. It is important to note that in regard to measures of central tendency, the DESCRIPTIVES command will report only the mean, not the mode or the median. Since the only measure of central tendency this command will calculate is the mean, this command is generally useful only for interval-level data. This command is most useful when you are working with a data set that contains almost exclusively interval-level variables.

An alternative to obtaining measures of central tendency is to again use the FREQUENCIES command discussed at the end of Chapter 3. As you may recall from the previous chapter's computer exercises, this command will produce frequency distributions for the variables whose names are included in the list of variables. To obtain measures of central tendency for the variables of interest, we simply need to add an option requesting these values:

```
FREQUENCIES VARIABLES=variable_names
  /FORMAT=NOTABLE
  /STATISTICS=MEAN MEDIAN MODE.
```

Where the /STATISTICS = option lists the three measures of central tendency we are interested in. The /FORMAT = NOTABLE option suppresses the printing of the frequency distributions for all the variables included in the list. Should you want the frequency distributions, just omit this line. Also, the separation of the command across three lines is simply for ease of reading—all of this material can be included on a single line in a syntax file. The formatting on this page of the text would have made it difficult to read what was being done within SPSS.

After running this command, the output window will contain a box labeled "Statistics." Each column of this table refers to a separate variable. As you move down the rows, you should see the reported values for the mode, the median, and the mean for each variable.

Caution: The mode and the median are listed as numbers, even though the data may be nominal or ordinal and you have entered value labels. To report correctly the value of the mode or median, you need to report the *category* represented by that number. For example, suppose you had analyzed the variable labeled "gender," where males were coded as 1 and females as 2, the mode would be reported by SPSS as either 1 or 2, but it would be up to you to report correctly whether the modal category was male or female–not a 1 or a 2.

Specific examples of the use of each of these commands are provided in the accompanying SPSS syntax file for Chapter 4 (Chapter_4.sps).

Stata

In Stata, there are also two primary ways of obtaining median and mean—both methods are fairly straightforward. The **summarize** command is

```
summarize variable_names
```

The basic output will include the mean and other measures on the variables included in the list. To obtain the median, you will need to add the detail option:

```
summarize variable_names, detail
```

You should note that the output will label the median as the 50th percentile.

Alternatively, if we want to avoid looking at a variety of other statistics that may not be of interest, we can use the **tabstat** command and explicitly ask for only the median and the mean:

```
tabstat variable_names, statistics ( median mean)
```

The output will list the variables named across column and the median and the mean will appear in separate rows of the table.

The mode is not reported in any of the standard Stata output, but is easily determined by running the **tab1** command to obtain a frequency distribution that was described in the Computer Exercises at the end of Chapter 3.

Specific examples of the use of each of these commands are provided in the accompanying Stata do file for Chapter 4 (Chapter_4.do).

Recoding Variables

A common situation in statistical analysis is the need to recode the values for some variable included in our data file. There are many reasons for why we may need to recode, such as collapsing categories to simplify the categories of a variable or defining some values as "missing" or inappropriate for our analysis.

The recode commands in SPSS and Stata are quite similar. In SPSS, RECODE takes the following form:

```
RECODE variable_name (old=new)(old=new) (ELSE=Copy) INTO
new_variable_name.

EXECUTE.
```

The structure of RECODE will have a series of old and new values listed in parentheses and can refer to specific values, ranges of values as well as missing values. For the values that are not going to be changed, the (ELSE = COPY) ensures that they are copied to the new variable. We would also like to emphasize the importance of using new variable names to contain the recoded values—it is good practice and helps to protect the data file that you are working with. More than one researcher has made the mistake of sending the recodes back into the original variable, realizing a mistake was made, and just damaged the data file that was being used and the need to start over.

The EXECUTE command following the RECODE command is necessary to force SPSS to perform the recodes now, rather than waiting for a call to another procedure and performing the recodes at that time.

In Stata, the **recode** command takes the form

```
recode variable_name (old=new), gen(new_variable_name)
```

The recoding of old to new values occurs in however many parentheses are required, just as in SPSS. There is no need to refer to the other values in the original variable—they are automatically carried over to the new variable. Following all of the recodes in parentheses, the creation of a new variable is noted by adding a comma to the command and then the **gen**(new_variable_name) command to generate a new variable that contains the recoded values.

Of special use is the option of coding one or more values as missing. For example, a person responding to a survey writes down an incorrect number that is beyond the range of acceptable responses. In SPSS, system missing values are noted by SYSMIS in the RECODE command. In Stata, a period (.) denotes a missing value.

Problems

1. Open the NYS data file (nys_1.sav, nys_1_student.sav, or nys_1.dta). Consider the level of measurement for each variable, and then compute and report appropriate measures of central tendency for each variable.

2. Do any of the variables included in the data file appear to have potential outliers? (You may want to consult your histograms from the Chapter 3 computer exercises or create histograms now for the interval-level variables included in the data file.)

3. If you find one or more potential outliers, take the following steps to investigate their effect on the measures of central tendency.

   a. Use the recode command to create two new variables: one that recodes the outliers as "System Missing" values and one that recodes the outliers as the next smaller value.

      Hints:

      - You will need to look at frequency distributions to determine the maximum and next to maximum values.
      - Keep in mind that SPSS will look for SYSMIS in the RECODE command, while Stata will look for a period (.) in the recode command.

   b. Using your variable that recodes the outliers as missing, report how the values of the mean and the median change when potential outliers are removed from the analysis.

   c. Using your variable that recodes the outliers as the next smaller value, report how the values of the mean and the median change when potential outliers are made less extreme.

   d. Which approach for handling potential outliers do you think is more appropriate for analyzing these data? Explain why. Faced with potential outliers, which measure of central tendency would you report?