

Central Tendency and Dispersion

4

In this chapter, you can learn

- how the values of the cases on a single variable can be summarized using measures of central tendency and measures of dispersion;
 - how the central tendency can be described using statistics such as the mode, median, and mean;
 - how the dispersion of scores on a variable can be described using statistics such as a percent distribution, minimum, maximum, range, and standard deviation along with a few others; and
 - how a variable's level of measurement determines what measures of central tendency and dispersion to use.
-

Schooling, Politics, and Life After Death

Once again, we will use some questions about 1980 GSS young adults as opportunities to explain and demonstrate the statistics introduced in the chapter:

- Among the 1980 GSS young adults, are there both believers and nonbelievers in a life after death? Which is the more common view?
- On a seven-attribute political party allegiance variable anchored at one end by “strong Democrat” and at the other by “strong Republican,” what was the most Democratic attribute used by any of the 1980 GSS young adults? The most Republican attribute? If we put all 1980

GSS young adults in order from the strongest Democrat to the strongest Republican, what is the political party affiliation of the person in the middle?

- What was the average number of years of schooling completed at the time of the survey by 1980 GSS young adults? Were most of these twentysomethings pretty close to the average on schooling completed, or were there large differences in amounts of school completed?

Once the data analysis techniques have been explained, evidence to answer the following questions will be quickly presented at the chapter's end:

- If we put the 1980 GSS young adults in line from the most conservative to the most liberal, what was the political orientation of the person in the middle?
- How low was the lowest social class reported by 1980 GSS young adults? How high was the highest social class? Putting them in order by social class, what did the person in the middle of the line say his or her social class was?
- How important did they think hard work was for success in life? How important was luck?
- How happy were they with their life? How healthy did they feel they were?
- How did the years of schooling completed thus far by 1980 GSS young adults compare to their fathers' schooling? Their mothers' schooling? And for those currently married, their spouses' schooling?
- Yes, they were all in their 20s when they took part in the GSS, but what was their average age?
- They were baby boomers, so did they have lots of siblings? What's the largest number of siblings any of them had?
- How many children did they have so far?
- How much on average did 1980 GSS young adults earn in the year prior to the survey?
- How much TV did the 1980 GSS young adults typically watch each day?



SPSS TIP

To Duplicate the Examples in This Chapter

To duplicate the examples in this chapter, use the *fourGroups.sav* data set. Before doing any of the statistical procedures, set the select cases condition to `GROUP = 1`.

Overview

This chapter takes you beyond frequency tables. On many occasions, you will not have the time, the space, or the desire to present and discuss all of the information in a frequency table, particularly when variables have many attributes. You want a way to summarize the data.

Think about when you come to class the day after an exam and someone asks the instructor how the grades were. Do you really want to know what percent of the grades were 100s, what

percent 99s, what percent 98s, and so on? You want to know the average because that gives you a sense of the center of the grade distribution, and you might want to know the low grade and the high grade because they give you a sense of how spread out or concentrated the grades were. Those are the kinds of statistics this chapter discusses: measures of central tendency and measures of dispersion. **Central tendency** gets at the typical score on the variable, while **dispersion** gets at how much variety there is in the scores.

When describing the scores on a single variable, it is customary to report on both the central tendency and the dispersion. Not all measures of central tendency and not all measures of dispersion can be used to describe the values of cases on every variable. What choices you have depend on the variable's level of measurement. Table 4.1 gives you the "big picture" for this chapter. The statistics you gain as you move from nominal to ordinal to interval/ratio are in boldface in the table.

Table 4.1 Measures of Central Tendency and Dispersion by Level of Measurement

Level of Measurement	Measures of Central Tendency	Measures of Dispersion
nominal	mode	percent distribution
ordinal	median mode	minimum and maximum range percentiles percent distribution
interval/ratio	mean median mode	variance standard deviation minimum and maximum range percentiles percent distribution

Modes and percent distributions are relatively simple in nature and only require that the attributes that make up a variable be different. That is a property that the attributes of nominal, ordinal, and interval/ratio variables all have. Therefore, you can calculate modes and percentage distributions for all three levels of measurement.

Medians, minimums, maximums, ranges, and percentiles provide more information about the scores on variables, but these statistics only make sense if the attributes of a variable have rank order. Since rank order is a property possessed by the attributes of ordinal and interval/ratio variables but not nominal variables, you may only calculate medians, minimums, maximums, ranges, percentiles, and interquartile ranges for ordinal and interval/ratio variables.

Means, variances, and standard deviations provide still more information about the scores on variables, but these statistics require the attributes of the variable to form a numeric scale with a

fixed unit of measurement. Since only interval/ratio variables have this property, means, variances, and standard deviations may only be calculated for interval/ratio variables.

Statistics are often referred to by the lowest level of measurement for which they can legitimately be calculated. For example, the mode would be described as a nominal measure of central tendency or the range as an ordinal measure of dispersion. Understand, however, that any measure that can be calculated for a nominal-level variable can also be calculated for ordinal and interval/ratio and that any statistic that can be calculated for an ordinal-level variable can also be calculated for an interval/ratio level of measurement.

When you have several measures of central tendency or several measures of dispersion from which to choose, how do you decide which one to use? Usually, you should use a statistic that makes the fullest use of the information packed into the variable's attributes. That means, for example, that in describing scores on an interval/ratio-level variable, you would normally choose a mean over a median or a mode because the mean makes use of the fact that the attributes on the variable not only are different and rank-ordered but also constitute a numeric scale. Using a median or a mode ignores some of those properties of the attributes. As you will see in this chapter, however, there are times when a statistic that does not use all the information contained in the variable is the better statistic!

Most of the chapter will be devoted to describing the individual statistics and what information each conveys. Requesting SPSS to calculate the various statistics is easy and can be quickly described in the final pages of the chapter. Most of the SPSS output in this chapter was generated using the *Frequencies* procedure.

CONCEPT CHECK

Without looking back, can you answer the following questions:

- What is the difference between central tendency and dispersion?
- What are three measures of central tendency? What are three measures of dispersion?
- What measures of central tendency and dispersion become available only at the interval/ratio level of measurement?

If not, go back and review before reading on.

Nominal Measures

Participants in the GSS were asked to indicate with a yes or a no if they believe there is life after death. Figure 4.1 shows the answers given by 1980 GSS young adults.

Statistics				
POSTLIFE				
N	Valid	294		
	Missing	33		
Mode		1		

POSTLIFE					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 no	64	19.6	21.8	21.8
	1 yes	230	70.3	78.2	100.0
	Total	294	89.9	100.0	
Missing	8 don't know	31	9.5		
	9 no answer	2	.6		
	Total	33	10.1		
Total		327	100.0		

Figure 4.1 Frequencies Output for POSTLIFE for 1980 GSS Young Adults

Mode

The **mode** is the attribute of a variable that occurs most often in the data set. It is the most common valid answer. System-missing or user-defined missing values are usually not eligible to be the mode.

There are several ways of finding the mode for a variable. You could look at the valid percent column (or the frequency column or the percent column) in a frequency table and find the row among the valid values that has the highest percent (or frequency). The value that that row represents is the mode for that variable. In Figure 4.1, the most common answer to the question about life after death is yes. In other words, the mode is “yes.” This attribute was coded “1” when the data set was created, and it is technically correct to say that the mode is 1; however, it is more informative when an attribute has a label to report the label rather than the numeric code. After all, the value label was added because the numeric code was not self-explanatory.

If you ask SPSS to calculate the mode, then the value of the mode will be included in the box of statistics that precedes the frequency table. As you can see in Figure 4.1, SPSS unfortunately only reports the numeric code for the mode and not the value label.



RESEARCH TIP

The Mode Is an Attribute, Not a Frequency or a Percent

A common mistake when identifying the mode from a frequency table is to report the largest frequency or the largest percent as the mode. The correct mode for the variable “belief in life after death” (Figure 4.1) is “yes.” If a person reported the mode as 78.2% or as 230, she would be wrong! You only locate the largest valid percent or the largest frequency so that you can identify the attribute that is the mode. The mode is not the frequency or the percentage; the mode is the attribute.

It is always possible that two or more attributes may tie as the most common value in the data set. Watch for ties when you scan a frequency table looking for the mode. If SPSS is identifying the mode for you, it will tell you if there is more than one mode. Having informed you of that fact, it will report only the first mode it located. You would have to examine a frequency table to identify the other mode or modes.

Any distribution with more than one mode can be described as **multimodal**. A distribution with two modes, however, is usually referred to as **bimodal**. A distribution with three or more modes is just referred to as multimodal.



RESEARCH TIP

Multimodal Distributions That Are Not Really

Technically speaking, two or more attributes must exactly tie for most common in order for a distribution to be multimodal. In fact, however, the terms *bimodal* and *multimodal* are often used in a looser sense to describe when cases tend to cluster around two or more different attributes. For example, on a test in which there were 3 Fs, 13 Ds, 2 Cs, 20 Bs, and 5 As, the professor might describe the distribution as bimodal. Technically, it is not. The mode is B. But the instructor seeks to highlight the fact that many more students got Ds and Bs than got Fs, Cs, or As. This looser meaning of bimodal or multimodal is sufficiently common that you need to be aware of it

The mode is a nominal measure of central tendency, which means it can be legitimately reported for nominal, ordinal, or interval/ratio variables.

Percent Distribution

So, now you know that yes was the most common answer to the question about believing in life after death. But what other answers were given, and how frequently were each of the answers

given? These are questions about dispersion; they are asking not about central tendency but about variation in the data. To answer questions such as these, you need to see a **percent distribution**, specifically, a valid percent distribution.

You can see in Figure 4.1 that yes was by far the more common answer (75.2%) but that about a fourth (24.8%) of 1980 GSS young adults said no.

Percent distributions can be calculated for nominal, ordinal, or interval/ratio variables.

RESEARCH TIP

The Mighty Percent Distribution—in Moderation

Although percent distributions are one of the simplest types of statistics, they are often the most effective way of describing the dispersion of scores on a variable. But reporting a long list of percents can deaden even the most interested audience. If you want the reader or listener to be aware of the valid percents for more than four categories, include the percent table in the presentation or paper and refer the reader to it.



Ordinal Measures

The GSS asked respondents to politically identify themselves on a 7-point scale from strong Democrat to strong Republican. The 7 points of the scale and the answers given by 1980 GSS young adults appear in Figure 4.2. The variable is named PARTYID. Being politically independent was placed in the middle of the scale. There was a user-missing code for persons who belonged to a political party other than Democrat or Republican, and one 1980 GSS young adult used that category.

Because the seven attributes are different from one another and can be put in order from least Republican (which would be the strong Democrat category) to most Republican (the strong Republican category) but do not form a numeric scale (remember, assigning numeric codes doesn't make a variable a numeric scale), PARTYID is an ordinal variable. We can still use the mode ("Democrat but not strong") to describe the central tendency and the percent distribution (the largest categories were "Democrat but not strong" [28.0%] and "independent" [21.2%], and the smallest categories were "strong Republican" [3.1%] and "strong Democrat" [6.2%]) to describe the dispersion, but we can also use some statistics that make use of the natural order among the attributes.

The ordinal measures of central tendency and dispersion are easiest to understand if you imagine the cases in your data set put in line based on their attribute on the variable: First we have all the "strong Democrats," then the "Democrats but not strong," then the "independents but nearly Democrat," and so on until finally the "strong Republicans" join the line.

Median

The **median** is a measure of central tendency. It identifies the value of the middle case when the cases have been placed in order or in line from low to high. The middle of the line is as far from

Statistics		
PARTYID		
N	Valid	325
	Missing	2
Median		3.00
Mode		1
Range		6
Minimum		0
Maximum		6
Percentiles	25	1.00
	50	3.00
	75	4.00

PARTYID		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 strong Democrat	20	6.1	6.2	6.2
	1 Democrat but not strong	91	27.8	28.0	34.2
	2 independent, nearly Democrat	49	15.0	15.1	49.2
	3 independent	69	21.1	21.2	70.5
	4 independent, nearly Republican	37	11.3	11.4	81.8
	5 Republican but not strong	49	15.0	15.1	96.9
	6 strong Republican	10	3.1	3.1	100.0
	Total	325	99.4	100.0	
Missing	7 other party	1	.3		
	9 no answer	1	.3		
	Total	2	.6		
	Total	327	100.0		

Figure 4.2 Frequencies Output for PARTYID for 1980 GSS Young Adults

being extreme as you can get. There are as many cases in line in front of the middle case as behind the middle case. The median is the attribute used by that middle case. When you know the value of the median, you know that at least half the cases had that value or a higher value, while at least half the cases had that value or a lower value.

Be careful, here. The median is not about putting the *attributes* in order from low to high and identifying the middle attribute. It is about putting the *cases* in order from low to high based on their attributes and identifying the attribute used by the middle case. For an ordinal variable with seven attributes, the median won't necessarily be the fourth attribute. It might be, but it could be any of the seven attributes. It depends on how the cases are distributed across the attributes.

Figure 4.2 shows there were 325 valid answers on political party identification for the 1980 GSS young adults. The middle of a line with 325 cases is the 163rd case. Counting from the beginning of the line (“strong Democrats”), the 163rd case comes in the group of “independents.” Therefore, the median for this variable is “independent.” While “independent” happens to be the middle attribute on this seven-attribute variable, “independent” is the median because that was the attribute used by the middle case.

As with the mode, use the attached value label, if there is one, rather than the numeric code when reporting the value of the median. It is more informative to say the median is “independent” than to say the median is “3.”

If the exact middle had fallen between two cases in different categories, the best way to report the median is to say that the median falls between the attributes ____ and ____.

The cumulative percent column in a frequency table can be used to quickly identify the median. As long as the rows of the table correspond to the natural order of the attributes, simply go down the cumulative percent column, stopping at the first row with a cumulative percent of 50.0 or higher. The attribute whose row you stopped at is the median. Try it on the frequency table in Figure 4.2. The first cumulative percent of 50.0 or higher is 70.5. The attribute represented by that row is “independent,” so that is the median. (In the rare case where the cumulative percent is exactly 50.0, then the median is between the attributes represented by that row and the next row.)

An even easier way of getting the median is to ask SPSS to identify it. The box labeled “Statistics” in Figure 4.2 reports the median to be 3.00, which is the numeric code for the attribute “independent.” If the median falls between two different attributes, SPSS will report as the median the average of the numeric codes for the two attributes (e.g., 2.50). While that is acceptable if the variable is interval/ratio, it is not appropriate when the variable is ordinal. In that case, it is better to report the two attributes between which the median falls.

The median is an ordinal measure of central tendency. It can be legitimately reported for ordinal- and interval/ratio-level variables. It is not appropriate for nominal variables because the attributes of a nominal variable lack rank order. A word of caution: SPSS will report medians for nominal variables as long as they are numeric-type variables. It is up to you to recognize that a median makes no sense for nominal variables.

Minimum and Maximum

What was the lowest attribute used by any case? That value is the **minimum**. What was the highest attribute used by any case? That value is the **maximum**. The minimum is the category with the lowest numeric code that was actually used by at least one person, and the maximum is the category with the highest numeric code that was actually used by at least one person. If we put the cases in line from low to high, the minimum is the attribute used by the first person in line, and the maximum is the attribute used by the last person. For 1980 GSS young adults, the minimum was “strong Democrat,” and the maximum was “strong Republican.”

When you know the value of the minimum and the maximum, you know that all of the cases had scores somewhere between those two values. For the 1980 GSS young adults, the answers went from strong Democrat to strong Republican. If in some other group the answers went only from “Democrat but not strong” to “independent but nearly Republican,” then there would be less dispersion (less difference) in that group’s political affiliations than among the 1980 GSS young adults.

While SPSS can identify the minimum and maximum values for you, you can certainly spot them yourself from a frequency table. If the rows are organized in order of ascending attributes, the minimum is the attribute represented by the first row of the valid values, and the maximum is the attribute represented by the last row of the valid values.



RESEARCH TIP

What Minimums and Maximums Aren't

The minimum and the maximum will not necessarily be the lowest available attribute and the highest available attribute. On a test, for example, no one may get a 100 and no one, hopefully, will get a 0. The maximum would be the highest score anyone actually got, and the minimum would be the lowest score actually received.

Furthermore, the minimum and the maximum will not necessarily be the attributes with the lowest and the highest frequencies. On a test in which there were 3 Fs, 13 Ds, 2 Cs, 20 Bs, and 5 As, the value of the minimum is F and the value of the maximum is A. That C was the least common grade and B the most common does not matter in identifying the minimum and the maximum.

Minimum and maximum are ordinal-level measures of dispersion, so they can be reported for ordinal- and interval/ratio-level variables since variables at both those levels of measurement have attributes with a natural order.

Range

The distance between the minimum and the maximum is called the **range**. The larger the value of the range, the more dispersed the cases are on the variable; the smaller the value of the range, the less dispersed (the more concentrated) the cases are on the variable.

Since the calculation of the range makes use of the minimum and the maximum, the range is an ordinal measure of dispersion. The interpretation of the range for ordinal variables is slightly different than for interval/ratio variables.

For ordinal variables, the range indicates how many attributes apart the minimum and the maximum are. For example, in Figure 4.2, “strong Republican” is six attributes away from “strong Democrat.” Even if some of the intervening categories were empty, the range would still be 6. You count the number of attributes away the maximum is from the minimum, including both used and unused attributes. If consecutive numeric codes have been used to represent the attributes of the ordinal variable, you simply subtract the code representing the minimum from the code representing the maximum.

For these 1980 GSS young adults, the range on political party affiliation is 6. The maximum is six attributes from the minimum. If the minimum had been “Democrat but not strong” and the maximum had been “independent but nearly Republican,” the range would have been 3 because the maximum is just three categories away from the minimum. A smaller range means there is less

dispersion in the answers. If everyone in the data set had the same attribute on a variable, the range would be 0 (and the variable could correctly be described as a constant).

For interval/ratio variables, the range represents the distance on a numeric scale from the minimum to the maximum. You calculate the range by subtracting the minimum value from the maximum value.

$$\text{range} = \text{maximum} - \text{minimum}$$

If the maximum grade was 100 and the minimum was 55, the range would be 45.

If the range of final grade point averages is 0.50 for graduating criminal justice majors and 1.50 for graduating social work majors, then there is more dispersion in the final grade point averages of social work majors than of criminal justice majors. Note that this only tells you that there are greater differences among social work majors than among criminal justice majors. It does not tell you which group had the higher average GPA or which group included more students.

RESEARCH TIP

The Statistical Range Is a Single Number

Do not make the common mistake of reporting the range as “*the value of the minimum to the value of the maximum.*” Simply putting the word *to* between the minimum and the maximum values does not make it the range. The range is a single number. The range for the political party affiliation is not “strong Democrat” to “strong Republican.” The range is 6.



RESEARCH TIP

Ranges Are Never Negative

Remember that the range is *the maximum minus the minimum.* Do the subtraction in the wrong order and you will get a different answer—the wrong answer. If you get a negative number for the value of the range, you have done the subtraction incorrectly!



Percentiles

To understand percentiles, go back to the image of the cases in line from lowest to highest. To get the value of the median, you started at the low end of the line and walked 50% of the way down the line, turned to that case, and recorded its value. That value was the value of the median. Another name for the median is the 50th percentile. If you walked just 25% of the way down the line, turned to that case, and recorded its value, you would have the value of the 25th percentile. If you walked 98% of

the way down the line, turned to that case, and recorded its value, that would be the 98th percentile. **Percentiles** are like milestones that mark certain points in the distribution of cases on a variable.

The 50th percentile is important because it marks the middle of the distribution. Other percentiles are usually reported in sets. For example, a researcher might want the values of the 25th, 50th, and 75th percentiles. These are called quartiles. Knowing them allows the researcher to divide the cases into four equal-size groups. A fourth of the cases have values between the minimum and the 25th percentile, a fourth have values between the 25th and 50th percentiles, a fourth have values between the 50th and 75th percentiles, and a fourth have values between the 75th percentile and the maximum.

In generating the *Frequencies* output for political party affiliation in Figure 4.2, the values for the quartiles (the 25th, 50th, and 75th percentiles) were requested. The values appear in the statistics box. The 25th percentile is “Democrat but not strong.” The 50th percentile is “independent,” and the 75th percentile is “independent, nearly Republican.” So, what does that tell us? Since the 50th percentile is “independent,” we know the middle person in the group identifies himself or herself as an independent. But that is one person. What if we want to know about the whole middle part of the group? Sometimes researchers wish to exclude the more extreme respondents on a variable and look at where the middle part of the distribution is—where what you might call the middle-of-the-roaders or the nonextremists are. By looking at the values of the 25th and 75th percentiles together, you see where the middle 50% of the respondents stand on the question. For these 1980 GSS young adults, the middle 50% ranged from “Democrat but not strong” to “independent, nearly Republican.” No one in the middle half of the group identified as a Republican! If you compared the 20th and 80th percentiles, you would see where the middle 60% are; the 10th and 90th percentiles tell you about the middle 80%.

Percentiles require that the cases be capable of being put in order from least to most. This can be done with ordinal and interval/ratio variables.

Interval/Ratio Measures

The GSS asked respondents how many years of formal schooling they had completed thus far. Since this is an interval/ratio variable, you can still ask the nominal-level questions (what is the most common years of schooling completed, and what percent of respondents reported various years of schooling completed?) and the ordinal-level questions (what is the fewest years of schooling completed, what is the greatest, how much difference is there between the greatest and the fewest, and how many years of schooling did the person in the middle of the distribution have?). You can also ask some interval/ratio-level questions: What is the average years of schooling completed? Were most persons pretty close to or pretty far from the average number of years of schooling completed? Figure 4.3 shows the SPSS results.

Mean

Our third and final measure of central tendency is the **mean**. (The technical name is the arithmetic mean to distinguish it from other means that we will not cover here.) The mean is what in

Statistics				
EDUC				
N	Valid			327
	Missing			0
Mean				12.79
Median				12.00
Mode				12
Std. Deviation				2.139
Variance				4.576
Skewness				.265
Std. Error of Skewness				.135
Range				12
Minimum				7
Maximum				19
Percentiles	25			12.00
	50			12.00
	75			14.00

EDUC					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7	5	1.5	1.5	1.5
	8	4	1.2	1.2	2.8
	9	9	2.8	2.8	5.5
	10	14	4.3	4.3	9.8
	11	14	4.3	4.3	14.1
	12	154	47.1	47.1	61.2
	13	26	8.0	8.0	69.1
	14	36	11.0	11.0	80.1
	15	13	4.0	4.0	84.1
	16	37	11.3	11.3	95.4
	17	8	2.4	2.4	97.9
	18	5	1.5	1.5	99.4
	19	2	.6	.6	100.0
Total		327	100.0	100.0	

Figure 4.3 Frequencies Output for EDUC for 1980 GSS Young Adults

everyday conversation is called the average. It is calculated by simply adding the values of all the valid cases together and dividing by the number of valid cases.

$$\text{mean} = \frac{\sum X_i}{N}$$

In this formula, X represents the variable we are working with. The X with a subscript i refers to each case's value on X . Putting the uppercase sigma (\sum) in front of the X_i indicates to sum all the individual values. After summing the values, you divide by N , which is the number of valid cases.

For example, a data set consists of just three cases. One of the variables is exam grade. The three scores on exam grade are 60, 90, and 90. To get the mean, you add the individual values ($60 + 90 + 90 = 240$) and divide that sum by the number of cases ($240/3 = 80$). The mean is 80.

The statistics in Figure 4.3 show that 1980 GSS young adults had a mean of 12.79 years of schooling completed at the time of the survey. You could calculate the mean by adding the 327 valid answers and then dividing the sum by 327. (Thank heaven for computers!)

The mean is an interval/ratio measure of central tendency. Its calculation requires that the attributes of the variable represent a numeric scale.

Variance

If someone tells you that the mean annual income of university graduates 10 years after graduation is \$45,000, he is telling you about central tendency. But are all the graduates' incomes clustered right around \$45,000, or are they actually quite different from each other? After all, you could get a mean of \$45,000 in many different ways. If absolutely every graduate made \$45,000, that would make the mean \$45,000, but so would if half the grads were making \$90,000 and half were making nothing. When you want information about the variety of scores or their spread, you are asking about dispersion.

An important measure of dispersion is the **variance**. The calculation of the variance requires the attributes of a variable to form a numeric scale. Thus, it is an interval/ratio-level measure of dispersion.

The variance indicates how close to or far from the mean are most of the cases for a particular variable. The smaller the value of the variance, the more the cases are concentrated around the value of the mean; the larger the value of the variance, the more spread out away from the mean are the cases.

The variance is not the simplest or easiest to understand measure of dispersion for interval/ratio variables. The reason why statisticians prefer it is because it provides an excellent basis for some very important multivariate statistics. One of the things that make the variance a bit tricky is that it does not look at simple distance of cases from the mean; instead, it looks at the squared distance of cases from the mean.

The formula for the variance is

$$\text{variance} = \frac{\sum (X_i - \bar{X})^2}{N - 1}$$

To calculate the variance, you must first calculate the mean. (The mean is represented in the variance formula by \bar{X} . A bar over a variable is a common way of representing the mean.) Then, for each case in the data set, take the value of the case and subtract the mean. Take the result of each subtraction and square it (multiply it by itself). Now add up all those squared values. Finally, divide that sum by the number of cases in the data set minus 1. The result is the variance.

Table 4.2 illustrates the calculation of the variance using a data set of just three cases. The cases have the values 2, 4, and 6.

Table 4.2 Illustrating the Calculation of the Variance

Step 1	$\bar{X} = (2 + 4 + 6)/3 = 12/3 = 4$		
Step 2	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
	2	$2 - 4 = -2$	$-2 \times -2 = 4$
	4	$4 - 4 = 0$	$0 \times 0 = 0$
	6	$6 - 4 = 2$	$2 \times 2 = 4$
Step 3	$\sum (X_i - \bar{X})^2 = 4 + 0 + 4 = 8$		
Step 4	$\sum (X_i - \bar{X})^2 / (N - 1) = 8/(3 - 1) = 8/2 = 4$		

In Step 1, you calculate the mean. In Step 2, for each case, you subtract the mean from the value of the case and square that result. In Step 3, you add together the squared results from the previous step. In Step 4, you take the result from the previous step and divide by the number of cases minus 1.

What the variance tells you is the approximate average squared distance of cases from the mean. If that doesn't excite you, remember the big advantage of the variance is that it will provide a foundation for some very powerful multivariate statistics. The variance would be telling you the exact average squared distance of cases from the mean had it been divided by N rather than by $(N - 1)$. The reason for dividing by $(N - 1)$ is because social scientists so often work with sample data, which tends to underestimate the variance in the population. The $(N - 1)$ corrects for this underestimate. Technically, what we are calculating is the "sample variance," but, like SPSS, we will refer to it simply as the variance.

The important thing to remember is that the larger the variance is, the more spread out or dispersed the cases are; the smaller the variance is, the less spread out or dispersed the cases are. In the extreme situation that every case has the same value, the variance equals 0, which makes sense since there is no variability or dispersion in the data.

For 1980 GSS young adults, the variance for years of schooling completed is 4.576. The approximate average squared distance of cases from the mean is between 4 and 5 years of schooling. By itself, that is not very useful information. If you were comparing the years of schooling completed

by different generations, however, you could compare the variances to see which generations had more uniformity in schooling and which had more variability.

Standard Deviation

Because people do not usually think in terms of squared distances, many researchers choose to report the **standard deviation** instead of the variance. The standard deviation tells you the approximate average distance of cases from the mean. This is easier to comprehend than the squared distance of cases from the mean.

The standard deviation is directly related to the variance. If you know the value of the variance, you can easily figure out the value of the standard deviation. The reverse is also true. If you know the value of the standard deviation, you can easily calculate the value of the variance.

The standard deviation is the square root of the variance.

$$\text{standard deviation} = \sqrt{\text{variance}}$$

In turn, the variance is equal to the value of the standard deviation squared.

$$\text{variance} = \text{standard deviation}^2$$

If you know the value of the variance and want the value of the standard deviation, take the square root of the variance. If the variance is 25, the standard deviation must be 5. If you know the value of the standard deviation and want the value of the variance, just square the standard deviation. If the standard deviation is 3, the variance would be 9.

Just like for the variance, the larger the standard deviation is, the more dispersed the cases are; the smaller the standard deviation is, the more concentrated around the mean the cases are. If all the cases had the same value on a variable, the standard deviation, like the variance, would have a value of 0.

Recall that the mean years of schooling completed by 1980 GSS young adults was 12.79 years. From Figure 4.3, you know the standard deviation is 2.139. That tells you something about how the cases clustered around that mean. On average, cases were about 2.139 years away from the mean value. Some cases were closer to the mean than that; some were further away. Is that a lot of dispersion or a little? Well, it certainly suggests a lot of these twentysomethings must have had 10 to 15 years of schooling, but, like the variance, the standard deviation is most useful when you can compare the standard deviations for different groups.

CONCEPT CHECK

Without looking back, can you answer the following questions:

- If an ordinal variable has nine attributes, is the fifth attribute always the median?
- If the value of the 25th percentile for household income is \$20,000 and the value of the 75th percentile is \$80,000, what do you know about the distribution of household incomes?

- If the mean on an exam was 75 and the standard deviation was 5, what do you know about the distribution of exam grades around the mean?

If not, go back and review before reading on.

Which Measure to Use?

Researchers rarely report every possible measure of central tendency and measure of dispersion. So which ones should you report? Usually, you would use the measures that make the greatest use of the information packed into the attributes of the variable. For an interval/ratio variable, you use interval/ratio measures; for an ordinal variable, you use ordinal measures; and, of course, for a nominal variable, you can only use nominal measures.

If a researcher has time to describe only a few things about each variable, here are the standard approaches: For a nominal variable, report the mode and just a few valid percents. For an ordinal variable, report the median, the minimum, and the maximum. For an interval/ratio variable, report the mean and the standard deviation.

There are two exceptions to the standard reporting packages that occur frequently enough for you to be aware of them. Both exceptions deal with interval/ratio variables. First, it is often the case that the minimum and maximum will also be reported along with the mean and standard deviation. A standard deviation does not give most persons a sense of the dispersion of the cases the way a minimum and maximum do.

The second exception has to do with interval/ratio variables whose distributions are badly skewed. A skewed distribution occurs when there are a few extreme scores on a variable but only in one direction. For example, when all but a few students score in the 80s and 90s on an exam but those few others score below 50, the grade distribution is negatively skewed. Or when most employees in a company receive annual salaries between \$25,000 and \$50,000 but a few executives receive salaries in excess of \$250,000, the salary distribution is positively skewed. The extreme cases pull the mean in their direction. The result is that the mean, although mathematically correct, is not a good measure of central tendency. The median, however, is not so sensitive to extreme values. The addition of one or two extreme cases has only a minor effect on the median, but they have a major effect on the mean. Therefore, report the median instead of the mean when the distribution is badly skewed.

RESEARCH TIP

What Constitutes a Badly Skewed Distribution?

The **skewness coefficient** is a statistic that indicates how skewed an interval/ratio variable's distribution is. When most of the cases are clustered above the mean but a few cases are far below the mean, the skewness coefficient will be negative. The further below zero the skewness coefficient is, the more negatively skewed (the more out of balance) the distribution is.

(Continued)



(Continued)

When most of the cases are clustered below the mean but a few cases are far above the mean, the skewness coefficient will be positive. The further above zero the skewness coefficient is, the more positively skewed (the more out of balance) the distribution is. The closer to zero the value of the skewness coefficient is, the greater the balance in the distribution of cases above and below the mean.

Ask your instructor for his or her suggestion for identifying badly skewed distributions, but here is one possible guideline: If the skewness coefficient is equal to or less (lower on the number line) than -2.00 or equal to or greater than $+2.00$, report the median instead of or in addition to the mean. Not many distributions have skewness coefficients at or below -2 , and not many have skewness coefficients at or above $+2$, but those that do are pretty severely skewed. (The skewness coefficient is easy to obtain from SPSS.)

You might notice that government agencies often report the median on variables such as years of schooling, income, and age. Although these variables are typically interval/ratio and could support the calculation of a mean, they are often positively skewed in the population. That is the reason government statisticians prefer reporting medians. Housing prices are also often positively skewed and, therefore, better summarized by the median than the mean.

Table 4.3 summarizes the standard packages for describing the distribution of cases on a variable. As always, you have to know the variable's level of measurement before proceeding.

Table 4.3 Usual Statistics Reported When Describing the Distribution of Cases on a Variable

Variable's Level of Measurement	Usual Measure of Central Tendency	Usual Measures of Dispersion
Nominal	Mode	1 to 4 valid percents
Ordinal	Median	Minimum, maximum
Interval/ratio (not seriously skewed)	Mean	Standard deviation, minimum, maximum
Interval/ratio (seriously skewed)	Median	Standard deviation, minimum, maximum

Getting SPSS to Calculate Statistics

Several SPSS procedures will calculate the measures of central tendency and dispersion discussed in this chapter. Here are three. You have already seen *Frequencies*. The two new procedures are *Descriptives* and *Explore*.

Frequencies

The *Frequencies* procedure was discussed in Chapter 3. In addition to frequency tables and charts, this procedure can provide a variety of univariate statistics. The initial *Frequencies* procedure dialog (see Figure 3.1) includes a “Statistics” button. When it is clicked, a dialog similar to Figure 4.4 appears with its menu of statistics.

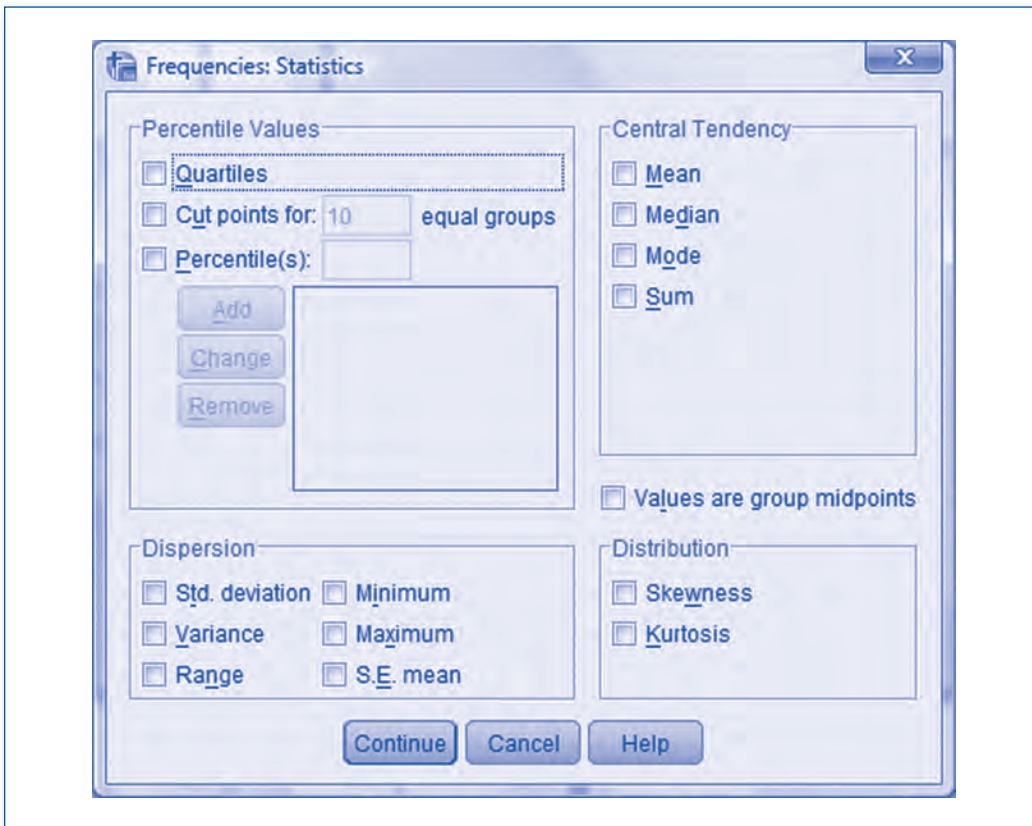


Figure 4.4 *Frequencies* Dialog for Selecting Univariate Statistics

Check as many or as few statistics as you want to appear in your output. When you click the “Continue” button, this dialog will disappear, and you will be back at the initial dialog for the *Frequencies* procedure.

Sometimes, all you want from the *Frequencies* procedure are the statistics and not the frequency table itself. For continuous variables with a large number of attributes, suppressing the actual

frequency table is a good idea. You can tell SPSS not to print the frequency table by unchecking the “Display frequency tables” message on the initial *Frequencies* dialog (see Figure 3.1).

Descriptives

The *Descriptives* procedure is a simple method for displaying univariate statistical information for interval/ratio variables. To select the procedure, pull down the *Analyze* menu, move the cursor to *Descriptive Statistics*, and click *Descriptives*.

Analyze | *Descriptive Statistics* | *Descriptives*

A dialog similar to Figure 4.5 appears.

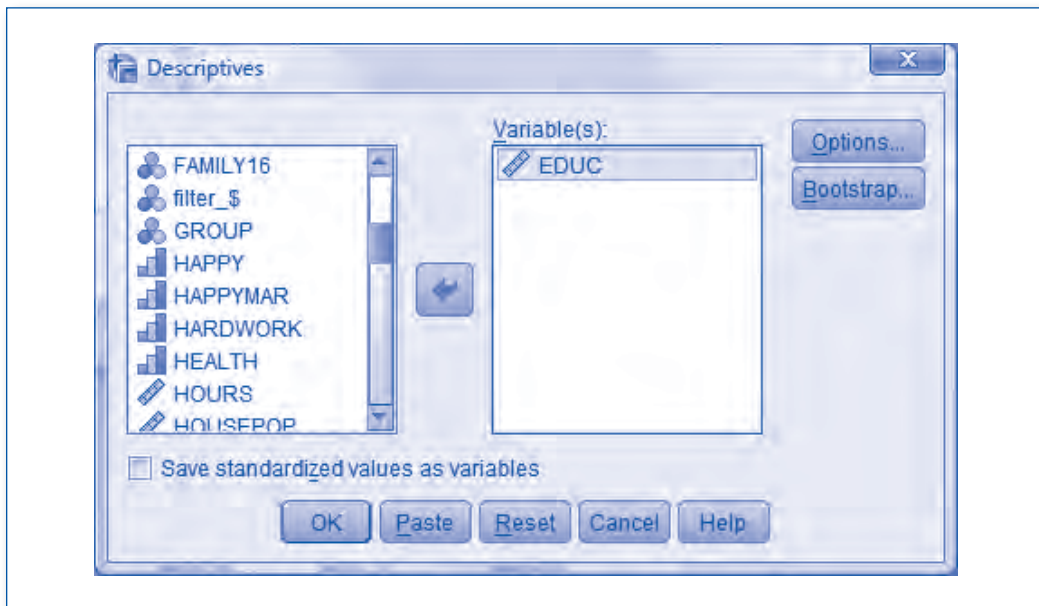


Figure 4.5 Dialog for *Descriptives* Procedure

Move the variables you want the statistical information for into the “Variable(s)” list, and then click “Options” to see the menu of available statistics. A dialog similar to Figure 4.6 will appear.

Make your choices and click “Continue” to return to the primary dialog. Click OK to generate the output. The output is straightforward and compact in format. An example appears in Figure 4.7. The

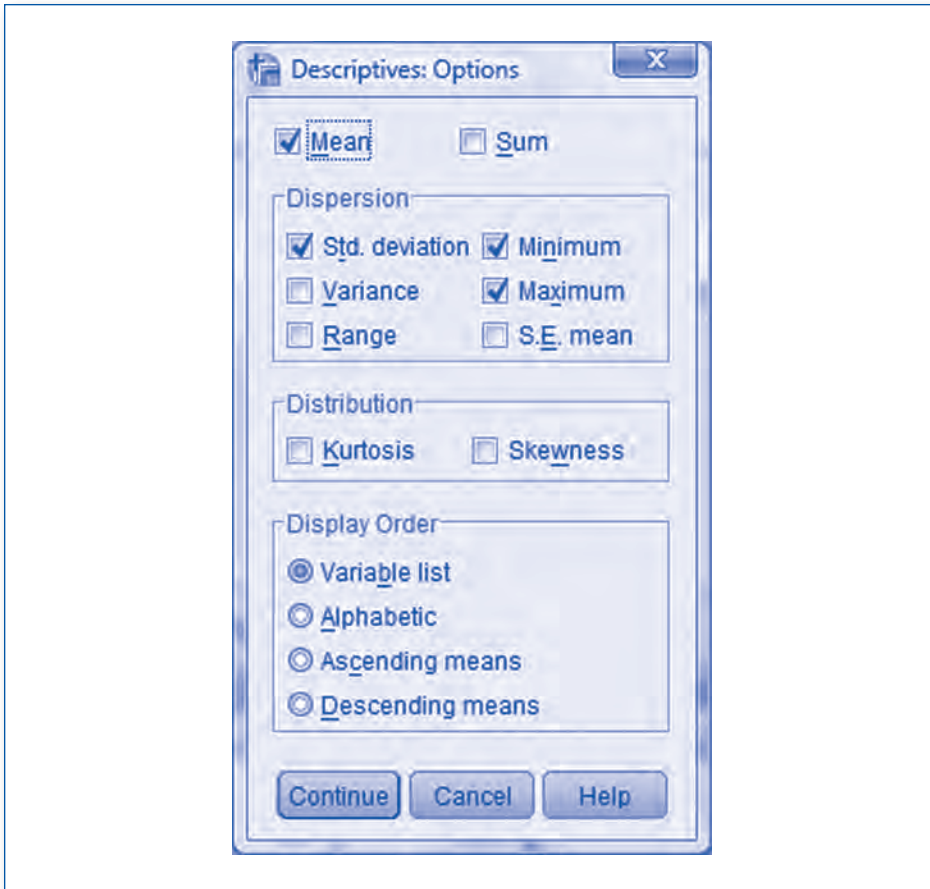


Figure 4.6 Descriptives Dialog for Selecting Univariate Statistics

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
EDUC	327	7	19	12.79	2.139
Valid N (listwise)	327				

Figure 4.7 Descriptives Output for EDUC for 1980 GSS Young Adults

final line, which is labeled “Valid N (listwise),” tells you the number of cases with valid information on all of the variables appearing in the table. When you run the *Descriptives* procedure on several variables and those variables have differing amounts of missing data, it is sometimes useful to know how many cases had complete data, that is, valid values on all the variables included in the procedure. That is what that last line of the output is telling you.

Explore

The *Explore* procedure gives you a package of frequently used univariate statistics. The only statistical choice you have is whether or not you want percentiles included in the package. *Explore*, like *Frequencies*, provides charts to visually display your data. If you choose to have charts, you can get a boxplot, a stem-and-leaf plot, and a histogram for each variable. All these charts assume your data are interval/ratio. Although neither boxplots nor stem-and-leaf plots are discussed in this book, the *SPSS Help* functions can help you with them.

To access the *Explore* procedure, pull down the *Analyze* menu, move your cursor to *Descriptive Statistics*, and click on *Explore*.

Analyze | *Descriptive Statistics* | *Explore*

A dialog similar to Figure 4.8 appears.

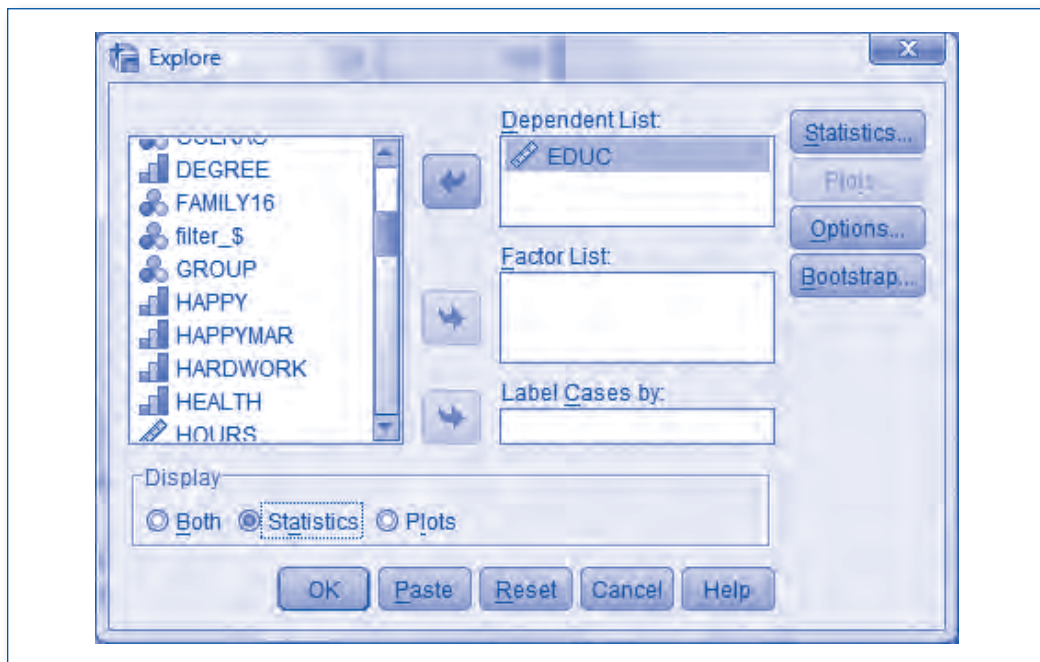


Figure 4.8 Initial Dialog for *Explore* Procedure

Move the variables you want statistical information for into the “Dependent List.” If you only want the statistics and not the plots, set the display in the lower left part of the dialog to “Statistics.” If you want percentiles, click on “Statistics” near the upper right of the dialog. In the dialog (not shown here) that opens, leave “Descriptives” checked but also check “Percentiles.” Click “Continue” to return to the initial dialog. Click OK to generate your output. Figure 4.9 is an example of the statistical output generated by *Explore* when percentiles have not been requested.

Descriptives			Statistic	Std. Error
EDUC	Mean		12.79	.118
	95% Confidence Interval for Mean	Lower Bound	12.56	
		Upper Bound	13.02	
	5% Trimmed Mean		12.79	
	Median		12.00	
	Variance		4.576	
	Std. Deviation		2.139	
	Minimum		7	
	Maximum		19	
	Range		12	
	Interquartile Range		2	
	Skewness		.265	.135
	Kurtosis		.570	.269

Figure 4.9 *Explore* Statistical Output for EDUC for 1980 GSS Young Adults

What Is Available Where?

Table 4.4 shows which statistics discussed in this chapter are available in which procedures. As you have seen, it’s easy to request statistics from SPSS. It is up to you to know the level of measurement of your variables and which statistics are legitimate and which are not.

CONCEPT CHECK

Without looking back, can you answer the following questions:

- What univariate statistics are usually used to describe the distribution of cases on a nominal variable?

(Continued)

(Continued)

- What univariate statistics are usually used to describe the distribution of cases on an ordinal variable?
- What univariate statistics are usually used to describe the distribution of cases on an interval/ratio variable that is not badly skewed?

If not, go back and review before reading on

Table 4.4 Available Univariate Statistics From SPSS Procedures

	Frequencies	Descriptives	Explore
Mode	x		
Median	x		x
Mean	x	x	x
Percent distribution	x		
Minimum and maximum	x	x	x
Range	x	x	x
Percentiles	x		x
Variance	x	x	x
Standard deviation	x	x	x
Skewness	x	x	x

More Univariate Results

Several of the questions asked at the beginning of the chapter about the 1980 GSS young adults can be answered with the *Frequencies* output in Figure 4.10 for the variables POLVIEWS, CLASS, HARDWORK, HAPPY, and HEALTH. All five variables have small numbers of attributes, so their frequency tables are fairly short. They are all ordinal variables, so we want to be sure to note their medians, minimums, and maximums. While we could ask SPSS to report the values for those statistics, it is easy to pick those values out directly from the frequency tables. The minimum is the first valid attribute in the table, the maximum is the last valid attribute, and the median is the first valid attribute to have a cumulative percent of 50.0 or higher.

POLVIEWS					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 liberal	126	38.5	38.7	38.7
	2 moderate	115	35.2	35.3	73.9
	3 conservative	85	26.0	26.1	100.0
	Total	326	99.7	100.0	
Missing	8 don't know	1	.3		
Total		327	100.0		

CLASS					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 lower class	18	5.5	5.5	5.5
	2 working class	187	57.2	57.4	62.9
	3 middle class	114	34.9	35.0	97.9
	4 upper class	7	2.1	2.1	100.0
	Total	326	99.7	100.0	
Missing	8 don't know	1	.3		
Total		327	100.0		

HARDWORK					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 luck/help more important	30	9.2	9.2	9.2
	2 both equally important	82	25.1	25.2	34.4
	3 hard work more important	214	65.4	65.6	100.0
	Total	326	99.7	100.0	
Missing	8 don't know	1	.3		
Total		327	100.0		

HAPPY					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 not too happy	46	14.1	14.1	14.1
	2 pretty happy	185	56.6	56.6	70.6
	3 very happy	96	29.4	29.4	100.0
	Total	327	100.0	100.0	

HEALTH					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 poor	3	.9	.9	.9
	2 fair	40	12.2	12.2	13.1
	3 good	154	47.1	47.1	60.2
	4 excellent	130	39.8	39.8	100.0
	Total	327	100.0	100.0	

Figure 4.10 Frequencies Output for 1980 GSS Young Adults

The GSS asked respondents to politically identify themselves as extremely conservative, conservative, slightly conservative, moderate, slightly liberal, liberal, or extremely liberal. The variable POLVIEWS combines the three conservative answers into a single conservative category and the three liberal categories into a single liberal category. The variable was collapsed into just three categories to simplify some of the analyses in future chapters.

Respondents were asked to identify their social class as lower class, working class, middle class, or upper class. They were also asked, “Some people say that people get ahead by their own hard work, others say that lucky breaks or help from other people are more important. Which do you say is most important?” The responses of 1980 GSS young adults to these two questions are the variables CLASS and HARDWORK.

The variable HAPPY is based on this question in the GSS: “Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?” And the variable HEALTH comes from the following GSS question: “Would you say your own health, in general, is excellent, good, fair, or poor?”

Figure 4.11 shows the output from a *Descriptives* procedure for several other variables relevant to the chapter’s opening questions about 1980 GSS young adults. All of the variables are interval/ratio, so we would usually report for each its mean, standard deviation, minimum, and maximum. The output also includes the skewness coefficient. This is not usually reported when presenting results but is useful for checking if the variable is badly skewed. If it is, then the median should be reported in place of or in addition to the mean.

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
MAEDUC	301	0	20	11.41	2.868	-.888	.140
PAEDUC	256	0	20	11.59	3.665	-.431	.152
SPEDUC	169	5	19	12.72	2.185	.233	.187
AGE	327	20	29	24.69	2.891	-.034	.135
SIBLINGS	327	0	20	3.96	2.756	1.668	.135
CHILDREN	327	0	6	.78	1.043	1.551	.135
INCOME86	244	755	139297	15182.60	14605.817	5.330	.156
TVHOURS	324	0	18	3.12	2.622	2.048	.135
Valid N (listwise)	98						

Figure 4.11 *Descriptives* Output for 1980 GSS Young Adults

Figure 4.11 begins with information about the years of schooling completed by the mothers, fathers, and spouses of 1980 GSS young adults. Only respondents who were currently married were asked about their spouse’s years of schooling. That is why the N for SPEDUC is low.

Next appears information about the age of 1980 GSS young adults. Since this group should consist only of persons in their 20s, any ages younger than 20 or older than 29 would signal that a mistake was made in creating the file.

The GSS asks about SIBLINGS in the following way: “How many brothers and sisters did you have? Please count those born alive, but no longer living, as well as those alive now. Also include stepbrothers and stepsisters, and children adopted by your parents.” The variable CHILDREN comes from the question, “How many children have you ever had? Please count all that were born alive at any time (including any you had from a previous marriage).”

INCOME86 records a respondent’s income the year prior to the survey after converting it into the equivalent amount of 1986 dollars. By converting the incomes of both 1980 and 2010 GSS respondents into 1986 dollars, the effect of inflation is eliminated and comparisons can be made across different survey years. Persons who had no income during the previous year received a missing value code for INCOME86, which largely explains the lower number of valid cases for INCOME86. For the 1980 GSS young adults, INCOME86 was badly positively skewed. A few persons in the data set had incomes much higher than everyone else. These very high scores pulled the mean higher and made it less representative of the data set as a whole. This is a common problem with income data. Since the skewness is severe, the median should be reported. The median income for 1980 GSS young adults who had income in the previous year is \$13,595. (The median can be obtained using the *Frequencies* or *Explore* procedures.)

RESEARCH TIP

The GSS and the Variable INCOME86

Income is always a difficult variable to measure well. A few things need to be noted about the income data reported by the GSS. First, respondents are asked only about their income from work. That means that other income, such as interest from savings, rent from property, or dividends from stock, is not included.

Second, the income data are for the calendar year prior to when the survey was conducted. Since people typically think of total income in terms of calendar years, that makes sense.

Third, because income can be a sensitive topic to some individuals and because exact income amounts are often difficult to recall, the GSS asks respondents to report their income using a set of categories. In their original form, these categories represent an ordinal scale. The GSS then recodes these categories, assigning the midpoint dollar amount of each category to each respondent who chose that category. This converts the variable to an interval/ratio level of measurement, and it also explains why, if you were to look at a frequency table for this variable, certain income amounts seem to be reported so frequently while other amounts are reported not at all.



Finally, TVHOURS is based on the GSS question, “On the average day, about how many hours do you personally watch television?” TVHOURS is also badly skewed but not as severely as INCOME86. The median number of hours of television watched daily by the 1980 GSS young adults is 3 hours.

1980 GSS Young Adults

The chapter began with some questions about 1980 GSS young adults. On the basis of our analyses, what do we now know?

- Most 1980 GSS young adults believed in life after death. Believers outnumbered nonbelievers by a ratio of almost 4 to 1.
- The political party affiliations of the 1980 GSS young adults ranged from strong Democrat to strong Republican. Democrats outnumbered Republicans, but the median reported affiliation was independent.
- At the time of the survey, 1980 GSS young adults had an average of one year of college (mean = 12.79 years). While completed years of schooling ranged from just 7 years to as high as 19 years, most had education levels fairly close to the average (standard deviation = 2.139 years). A lot of 1980 GSS young adults had just a high school education or only a few years of college completed at the time of the survey.
- Among 1980 GSS young adults, liberals outnumbered conservatives, but the median political orientation was moderate.
- Very few 1980 GSS young adults described themselves as upper class. Slightly more said they were lower class, but by far, most described themselves as either working class or middle class. Working class was the most commonly reported class. It was also the median.
- 1980 GSS young adults largely believed that success was primarily due to hard work. Almost two thirds said hard work was more important, while less than a tenth said luck was more important. The remainder believed hard work and luck were equally important.
- The median responses of 1980 GSS young adults were that their life was pretty happy and their health good. More described their life as very happy than as not too happy, and far more described their health as excellent than as poor or fair.
- Both the mothers and the fathers of 1980 GSS young adults averaged less than a complete high school education, which means that 1980 GSS young adults on average had already surpassed by more than a year the average educational levels of their parents. 1980 GSS young adults who were currently married had spouses who averaged almost the same schooling as themselves.
- 1980 GSS young adults ranged in age from 20 to 29 and had an average age of 24.69 years, which is very near the middle of the age range.
- 1980 GSS young adults averaged 3.96 siblings. Although some had no siblings, at least one person reported 20 siblings. If the average seems high and the maximum seems very high, remember that this was the baby boom generation and that stepsiblings and adopted siblings are included in the count.
- At the time of the survey, 1980 GSS young adults averaged less than one child. Of course, for most, their fertility was not yet complete, and for many, it had not yet begun.
- About three fourths of 1980 GSS young adults reported income from a job during the previous calendar year. Of those that did, the median income expressed in 1986 dollars was \$13,595. While this may seem low, remember that this includes persons who may have

worked only part-time or part of the year and that those 1986 dollars purchased more than current dollars.

- 1980 GSS young adults watched quite a bit of television each day. Their median daily television viewing was 3 hours.

Important Concepts in the Chapter

bimodal	minimum
central tendency	mode
<i>Descriptives</i> procedure	multimodal
dispersion	percent distribution
<i>Explore</i> procedure	percentile
<i>Frequencies</i> procedure	range
maximum	skewness coefficient
mean	standard deviation
median	variance

Practice Problems

1. Explain what central tendency means. What are the measures of central tendency presented in this chapter?
2. Explain what dispersion means. What are the measures of dispersion presented in this chapter?
3. What is the lowest level of measurement for which it is permissible to calculate each of the following statistics?
 - a. mean
 - b. standard deviation
 - c. variance
 - d. minimum
 - e. percentiles
4. When reporting on the central tendency and dispersion of an interval/ratio variable, what statistics are usually reported along with the mean and standard deviation because they give a person a quick sense of the dispersion?
5. When reporting on the central tendency of an interval/ratio variable, when is the median reported along with or in place of the mean?

6. Five students have the following number of TV shows they watch regularly: 0, 2, 3, 10, 10. Compute each of the following statistics based on these values:
 - a. mean
 - b. median
 - c. mode
 - d. maximum
 - e. minimum
 - f. range
7. A researcher takes a sample of three students from a large class and asks each of them how many years of formal schooling they have completed thus far and how many sisters they have. Calculate the variance for each variable.
 - a. When asked how many sisters they have, the students' answers were 0, 0, and 3.
 - b. When asked how many years of formal schooling they have completed thus far, the students' answers were 15, 15, and 15.
8. A variable has a standard deviation of 3.00. What is the value of its variance?
9. A variable has a variance of 36. What is the value of the standard deviation?
10. A variable has a variance of 36. What is the approximate average distance of cases from the mean for this variable?
11. You want to know the value of the middle case when the cases are arranged from low to high. Which statistic should you calculate?
12. The range for household incomes in Canon City is \$400,000. The range for household incomes in Cripple Creek is \$250,000.
 - a. Which city has more dispersion in household incomes: Canon City, Cripple Creek, or is there insufficient information to tell?
 - b. Which city has a higher mean household income: Canon City, Cripple Creek, or is there insufficient information to tell?

Problems Requiring SPSS and the *fourGroups.sav* Data Set

13. (*2010 GSS young adults*) This problem looks at the belief in life after death of the twentysomethings who took part in the 2010 GSS. (Variables: POSTLIFE; Select Cases: GROUP = 3)
 - a. What percent believed in life after death?
 - b. Summarize in a few well-written sentences totaling 50 words or less what you found out about the belief in life after death of 2010 GSS young adults.
14. (*2010 GSS middle-age adults*) This problem looks at the belief in life after death of the fiftysomethings who took part in the 2010 GSS. (Variables: POSTLIFE; Select Cases: GROUP = 4)

- a. What percent believed in life after death?
 - b. [Figure 4.1 shows that 78.2% of the 1980 GSS young adults believed in life after death.] Summarize in a few well-written sentences totaling 50 words or less the similarities and/or differences in belief in life after death of these two groups of baby boomers—the 1980 young adults and the 2010 GSS middle-age adults.
15. (2010 GSS middle-age adults) This problem looks at the political party identification of the fiftysomethings who took part in the 2010 GSS. (Variables: PARTYID; Select Cases: GROUP = 4)
- a. For these 2010 GSS middle-age adults
 - i. What was their modal political party identification?
 - ii. What was their median political party identification?
 - iii. What percent identified as Democrat? (Include both “strong Democrat” and “Democrat but not strong” but exclude “independent, nearly Democrat.”)
 - b. Summarize in a few well-written sentences totaling 50 words or less what you found out about the political party identification of 2010 GSS middle-age adults.
16. (2010 GSS young adults) This problem looks at the political party identification of the twentysomethings who took part in the 2010 GSS. (Variables: PARTYID; Select Cases: GROUP = 3)
- a. For these 2010 GSS young adults
 - i. What was their modal political party identification?
 - ii. What was their median political party identification?
 - iii. What percent identified as Democrat? (Include both “strong Democrat” and “Democrat but not strong” but exclude “independent, nearly Democrat.”)
 - b. [Figure 4.2 shows for the 1980 GSS young adults a modal identification of “Democrat but not strong,” a median identification of “independent,” and 34.2% identifying as Democrat.] Summarize in a few well-written sentences totaling 50 words or less the similarities and/or differences in political party identification of the 1980 and 2010 GSS young adults.
17. (2010 GSS young adults) Which, if any, of the following interval/ratio variables are “badly skewed” in the data from the 2010 GSS young adults: EDUC, MAEDUC, PAEDUC, SPEDUC, AGE, SIBLINGS, CHILDREN, INCOME86, and TVHOURS? (See the text or your instructor for a definition of “badly skewed.”) (Select Cases: GROUP = 3)
18. (2010 GSS young adults) This problem looks at the years of schooling completed by the time of the survey by the twentysomethings who took part in the 2010 GSS. (Variables: EDUC; Select Cases: GROUP = 3)
- a. For years of schooling for these 2010 GSS young adults,
 - i. What is the value of the mean?
 - ii. What is the value of the median?
 - iii. What is the value of the skewness coefficient?
 - iv. What is the value of the standard deviation?
 - v. What is the value of the minimum and the maximum?
 - b. [Figures 4.3 and 4.9 both show for years of schooling completed by the 1980 GSS young adults a mean of 12.79, a median of 12, a skewness coefficient of .265, a standard deviation of 2.139, a minimum of 7,

and a maximum of 19.] Summarize in a few well-written sentences totaling 50 words or less the similarities and/or differences in the years of schooling completed by the time of the survey by the 1980 and 2010 GSS young adults. (If either group's data are badly skewed, base your comparison of central tendencies on the medians, otherwise the means.)

19. (1980 GSS middle-age adults and 2010 GSS middle-age adults) This problem compares the years of schooling completed by the time of the survey by the fiftysomethings who took part in the 1980 GSS and the fiftysomethings who took part in the 2010 GSS. Unlike persons in their 20s, most persons in their 50s have obtained all the formal schooling they will ever get.
- For years of schooling for the 1980 GSS middle-age adults (Variables: EDUC; Select Cases: GROUP = 2),
 - What is the value of the mean?
 - What is the value of the median?
 - What is the value of the skewness coefficient?
 - What is the value of the standard deviation?
 - What is the value of the minimum and the maximum?
 - For years of schooling for the 2010 GSS middle-age adults (Variables: EDUC; Select Cases: GROUP = 4),
 - What is the value of the mean?
 - What is the value of the median?
 - What is the value of the skewness coefficient?
 - What is the value of the standard deviation?
 - What is the value of the minimum and the maximum?
 - Summarize in a few well-written sentences totaling 50 words or less the similarities and/or differences in the years of schooling completed by these two groups of fiftysomethings—the 1980 middle-age adults and the 2010 GSS middle-age adults. (If either group's data are badly skewed, base your comparison of central tendencies on the medians, otherwise the means.)
20. (1980 GSS middle-age adults and 2010 GSS middle-age adults) This problem compares the number of children of the fiftysomethings who took part in the 1980 GSS and the fiftysomethings who took part in the 2010 GSS. Since persons in their 50s have generally had all the children they will have, this is a comparison of completed fertility.
- For number of children of the 1980 GSS middle-age adults (Variables: CHILDREN; Select Cases: GROUP = 2),
 - What is the value of the mean?
 - What is the value of the median?
 - What is the value of the skewness coefficient?
 - What is the value of the standard deviation?
 - What is the value of the minimum and the maximum?
 - For number of children of the 2010 GSS middle-age adults (Variables: CHILDREN; Select Cases: GROUP = 4),
 - What is the value of the mean?
 - What is the value of the median?
 - What is the value of the skewness coefficient?

- iv. What is the value of the standard deviation?
 - v. What is the value of the minimum and the maximum?
 - c. Summarize in a few well-written sentences totaling 50 words or less the similarities and/or differences in number of children of these two groups of fiftysomethings—the 1980 middle-age adults and the 2010 GSS middle-age adults. (If either group's data are badly skewed, base your comparison of central tendencies on the medians, otherwise the means.)
- 21. Here are some comparisons you might conduct using other variables introduced in this chapter:
 - a. POLVIEWS
 - i. (1980 GSS young adults and 2010 GSS young adults)
 - ii. (1980 GSS young adults and 2010 GSS middle-age adults)
 - b. CLASS
 - i. (1980 GSS young adults and 2010 GSS young adults)
 - ii. (1980 GSS young adults and 2010 GSS middle-age adults)
 - c. HARDWORK
 - i. (1980 GSS young adults and 2010 GSS young adults)
 - ii. (1980 GSS young adults and 2010 GSS middle-age adults)
 - d. HEALTH
 - i. (1980 GSS young adults and 2010 GSS young adults)
 - ii. (1980 GSS young adults and 2010 GSS middle-age adults)
 - e. TVHOURS
 - i. (1980 GSS young adults and 2010 GSS young adults)
 - ii. (1980 GSS young adults and 2010 GSS middle-age adults)

Questions and Tools for Answering Them (Chapter 5)

