

SPSS DEMONSTRATIONS [GSS18SSDS-A]

Demonstration 1: Producing Measures of Central Tendency With Frequencies

The Frequencies command, which we demonstrated in Chapter 2, also has the ability to produce the three measures of central tendency discussed in this chapter. We will use Frequencies to calculate measures of central tendency for EMAILHR (number of e-mail hours per week) and TVHOURS (hours per day watching television).

Click on *Analyze, Descriptive Statistics, then Frequencies*. Place EMAILHR and TVHOURS in the Variable(s) box. Then click on the *Statistics* button. You will see that the Central Tendency box lists four choices, but we will click on only the first three—(1) Mean, (2) Median, and (3) Mode. Then click on *Continue*, then on *OK* to process this request.

The statistics box for EMAILHR and TVHOURS is displayed here (Figure 3.12). Both variables are interval-ratio measurements. Although we have instructed you to select the Mean, Median, and Mode for your output, the mean is the most appropriate measure of central tendency. SPSS produces exactly the output we asked for, without regard for whether the output is correct for this type of variable. It is up to you to select the proper measure of central tendency.

Suppose we want our results split by sex, allowing us to separate results for males and females. Select *Data, Split File, Organize Output by Groups*. Insert the variable *sex* into the box labeled “Groups Based on.” Click *OK*. Now SPSS will filter our results by SEX. Let’s include AGE in our analysis.

When you want to calculate the mean of interval-ratio variables but you don’t need to view the actual frequency table listing the responses in each category, the Descriptives procedure is often the best choice. Descriptives can be found by clicking on *Analyze, Descriptive Statistics, and then Descriptives*.

The Descriptives dialog box is uncomplicated and requires only that you place the variables of interest (AGE, EMAILHR, and TVHOURS) in the Variable(s) box. By default, Descriptives will calculate the mean, standard deviation (to be discussed in the next chapter), minimum, maximum, and the number of cases with a valid response.

Figure 3.12

Statistics			
		Email hours per week	Hours per day watching tv
N	Valid	923	1014
	Missing	577	486
Mean		7.24	2.97
Median		2.00	2.00
Mode		0	2

You will need to scroll through your SPSS output to locate the descriptive statistics for female respondents. Figure 3.13 displays these descriptive statistics for AGE, EMAILHR, and TVHOURS for female GSS respondents in 2018.

The output from Descriptives automatically lists the variables in the order that we specified in the dialog box. Based on the output, we can determine that on an average, female respondents were 48.50 years old, used their e-mail 7.77 hours per week, and watched 3.17 hours of television per day.

Figure 3.13

SEX = 1 Male

Descriptive Statistics ^a					
	N	Minimum	Maximum	Mean	Std. Deviation
Age of respondent	684	18	89	48.92	18.028
Email hours per week	412	0	90	6.57	10.699
Hours per day watching tv	455	0	24	2.73	2.537
Valid N (listwise)	410				

a. Respondents sex = Male

SEX = 2 Female

Descriptive Statistics ^a					
	N	Minimum	Maximum	Mean	Std. Deviation
Age of respondent	811	18	89	48.50	17.972
Email hours per week	511	0	80	7.77	12.763
Hours per day watching tv	559	0	24	3.17	3.311
Valid N (listwise)	507				

a. Respondents sex = Female

SPSS EXERCISES [GSS18SSDS-A]

- S1. Based on the last SPSS Demonstration on Descriptives, compare the means for each variable (AGE, EMAILHR, and TVHOURS) for men and women.
 - a. On average, which group is older, uses their e-mail more frequently, and watches more television per day?
 - b. Program SPSS to split the file by DEGREE (respondent's educational attainment) and run Descriptives for EMAILHR. Rank the DEGREE groups by the highest to lowest e-mail use per day.

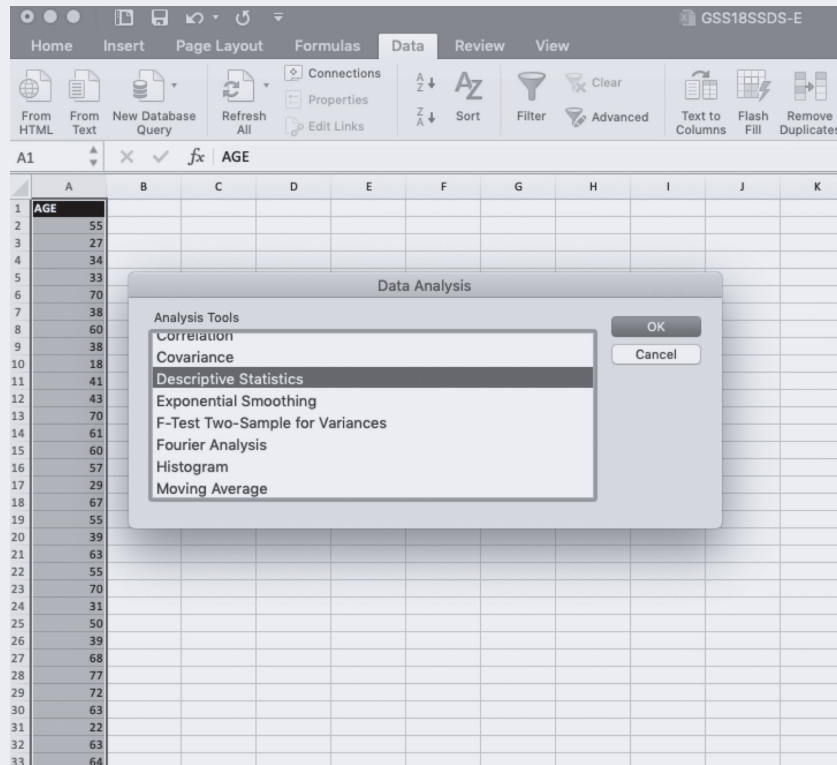
- S2. Picking an appropriate statistic to describe the central tendency of a distribution is a critical skill. Determine the appropriate measure(s) of central tendency for the following variables (remember to reset the Split File option—*Analyze All Cases, Do Not Create Groups*):
- Respondent's marital status [MARITAL]
 - Respondent's general level of happiness [HAPPY]
 - Number of hours a respondent worked last week [HRS1]
 - Which presidential candidate respondent voted for in 2016 [PRES16]
 - Number of brothers and sisters of respondents [SIBS]
 - Beliefs about the Bible [BIBLE]
- S3. Create a frequency distribution, including any appropriate measures of central tendency, for PREMARSX (Does the respondent approve of sex before marriage?).
- Which measure of central tendency is most appropriate to summarize the distribution of PREMARSX? Explain why.
 - Suppose we are interested in whether attitudes about premarital sex vary by beliefs about the Bible. Create a frequency distribution, including any appropriate measures of central tendency, for PREMARSX, this time separating results for the categories in BIBLE. Are there any differences in their measures of central tendency? Explain.
- S4. Create a frequency distribution, including any appropriate measures of central tendency, for EDUC (years of education).
- Which measure of central tendency is most appropriate to summarize the distribution of EDUC? Explain why.
 - Create a frequency distribution, this time separating results for SEX categories. Include in your analysis the appropriate measure of central tendency. Are there any differences in their measures? Explain.
- S5. Some believe that social class influences an individual's decision on the number of children to have. Use SPSS to investigate this question with the GSS data file. The variable CHILDS measures the respondent's number of children. To produce the necessary information, have SPSS split the file by CLASS and then run Frequencies (and Statistics) for CHILDS.
- What is the best measure of central tendency to represent the number of children in a household? Why?
 - Which social class has more children per respondent?
 - Rerun your analysis, this time with the variable CHLDIDEL (ideal number of children). Is there a difference among the social class categories? Explain.

EXCEL DEMONSTRATIONS [GSS18SSDS-E]

Demonstration 1: Producing Measures of Central Tendencies for an Interval-Ratio Variable

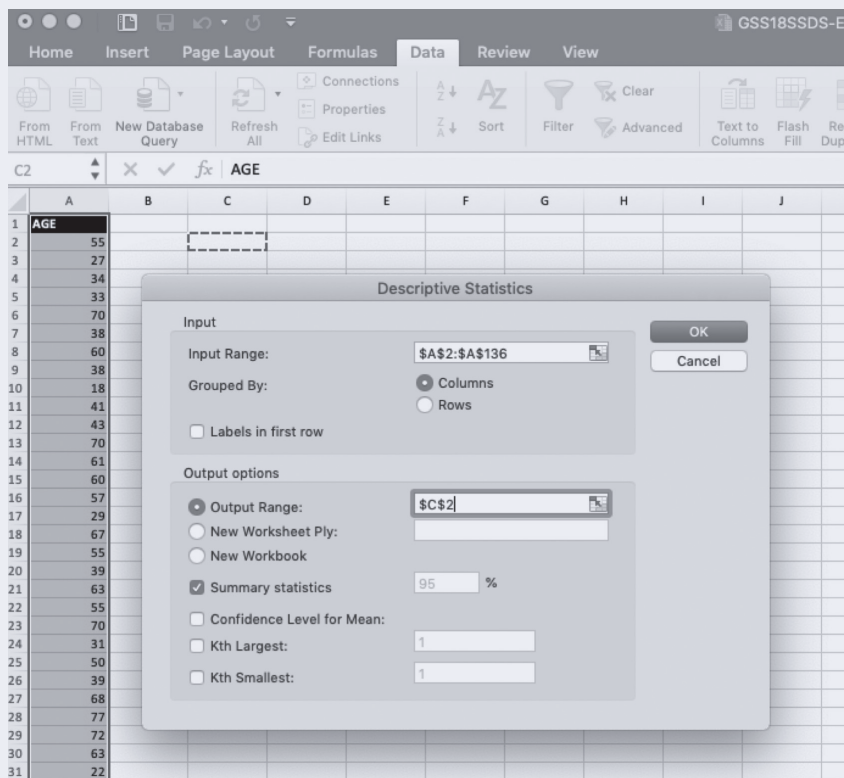
With Excel's Analysis ToolPak installed (see “Introduction to Software, Data Sets, and Variables” at the end of Chapter 1 for step-by-step instructions on how to install this Excel Add-in), you can easily produce descriptive statistics for any interval-ratio variable. To demonstrate, we will work with AGE (respondent's age) from GSS18SSDS-E. Copy the AGE data from the protected Data View sheet and paste it into a new Excel sheet. Navigate to Excel's Data tab and select *Data Analysis*. A window of Analysis Tools will appear (see Figure 3.14).

Figure 3.14



Select *Descriptive Statistics* and then *OK*. A window similar to Figure 3.15 will appear. Click in the empty box next to “Input Range” and highlight the column of AGE data from A2 to A136. Do not select A1 for it contains the variable name, AGE. Under “Output options,” select *Summary statistics*. Click in the empty box next to “Output Range,” and then select any cell in the current sheet you are

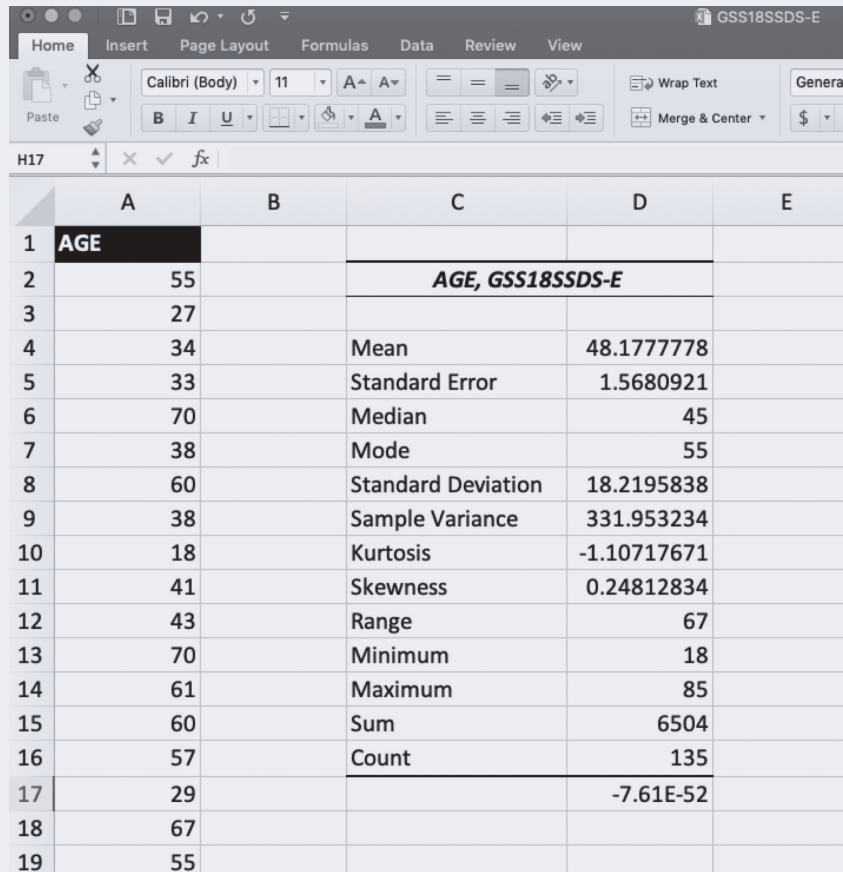
Figure 3.15



working in. This will tell Excel where to place the Descriptive Statistics box it will generate. In our example, we’ve chosen for the output box to begin in cell C2. Click *OK*.

Excel will produce a Descriptive Statistics table that includes measures of central tendency for AGE as well as measures of variability (see Figure 3.16), which we will discuss in Chapter 4. For now, let’s focus on the mode, median, and mean—all three measures of central tendency are useful when working with an interval-ratio variable such as AGE. We can see the mean AGE in our data set is 48.18, and thus the average age of respondents in GSS18SSDS-E is approximately 48 years. The median is 45, indicating half of our respondents are 45 and younger, while the other half of our respondents are 45 and older. Because the mean age is slightly higher than the median age, you can state that age is very slightly positively skewed. The mode is 55, indicating the most common age in our sample is 55 years. Notice how we extended the width of column C. We also replaced the automatically generated table title “Column 1” with “Age, GSS18SSDS-E.”

Figure 3.16

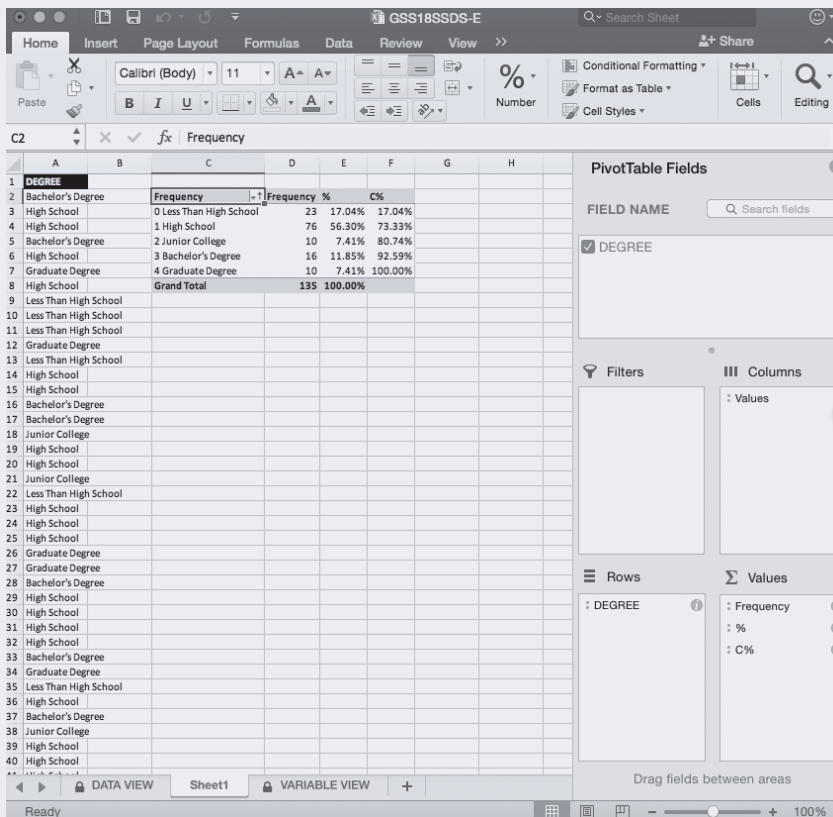


	A	B	C	D	E
1	AGE				
2	55		AGE, GSS18SSDS-E		
3	27				
4	34		Mean	48.177778	
5	33		Standard Error	1.5680921	
6	70		Median	45	
7	38		Mode	55	
8	60		Standard Deviation	18.2195838	
9	38		Sample Variance	331.953234	
10	18		Kurtosis	-1.10717671	
11	41		Skewness	0.24812834	
12	43		Range	67	
13	70		Minimum	18	
14	61		Maximum	85	
15	60		Sum	6504	
16	57		Count	135	
17	29			-7.61E-52	
18	67				
19	55				

Demonstration 2: Producing Measures of Central Tendencies for a Nominal or Ordinal Variable With a PivotTable

The easiest way to identify the measures of central tendency for nominal (mode) and ordinal variables (mode and median) organized in Excel is to create a frequency distribution through the PivotTable command. We first discussed this command in Chapter 2's Excel Demonstration 1. When you create a frequency distribution of a nominal or ordinal variable, you can identify the mode by finding the category with the highest frequency or percentage. When you create a frequency distribution for an ordinal variable, you can examine the cumulative percentage column and determine where the 50th percentile falls. The category associated with the 50th percentile will be your median. Recall that you can report the mode for both nominal and ordinal variables, but the median is only useful for ordinal variables. Figure 3.17 displays a frequency distribution for the variable DEGREE (respondent's highest degree earned). We can see that

Figure 3.17



the mode is “High School.” Seventy-six respondents, or 56.30%, reported a high school diploma was the highest degree they’ve earned. In the cumulative percentage column (C%), we can see that the 50th percentile also falls in the “High School” category. Thus, half of the respondents have a high school diploma or less, whereas the other half have a high school diploma or more. Be careful when interpreting the cumulative percentage column in Excel. You have to make sure the categories of the variable are ordered from low to high. If you would like to review how to order the categories, review Excel Demonstration 1 in Chapter 2.

Demonstration 3: Producing the Mean for an Interval-Ratio Variable by Categories of a Nominal or Ordinal Variable

We will use Excel’s PivotTable command to examine whether the average number of siblings (SIBS) a respondent has varies by sexual orientation (SEXORNT). Begin this process by copying and pasting both the SEXORNT and SIBS column data

into a new Excel sheet. Highlight both columns of data, and then on the *Insert* tab, choose PivotTable. Drag SEXORNT in the Field Name box to the Row box and SIBS to the Values box (see Figure 3.18).

In the frequency table, click on the arrow in the “Row Labels” cell and uncheck the box marked (blank). This will remove the missing data row from the frequency distribution. You can also rename the “Row Labels” cell SEXORNT (the variable name) as we did. Next, double click on “Sum of SIBS.” A Window should appear (see Figure 3.19). In the Field Name box, replace “Sum of SIBS” with “Mean SIBS,” and then in the “Summarize by” box, choose “Average” and click *OK*. Figure 3.19 shows the mean number of siblings by sexual orientation. We can see that bisexual respondents have an average of 4.8 siblings. Heterosexual or straight respondents have an average of 3.46 siblings. And gay, lesbian, or bisexual respondents have, on average, 3.25 siblings.

Excel’s PivotTable command is useful in statistics, and we will be using it throughout our Excel demonstrations. If you are using Excel on a Windows PC (as opposed to a MacBook), you should be able to download Excel’s Power Pivot command by going to *File* → *Options* → *Add-ins* → *Manage* → *COM Add-ins* → *Go* → *Microsoft Power Pivot* → *OK*. Power Pivot offers users more tools to analyze statistics and generate output. At the time this textbook went to press, Power Pivot was not an available add-in for Macintosh users. We thus do not cover Power Pivot in this book.

Figure 3.18

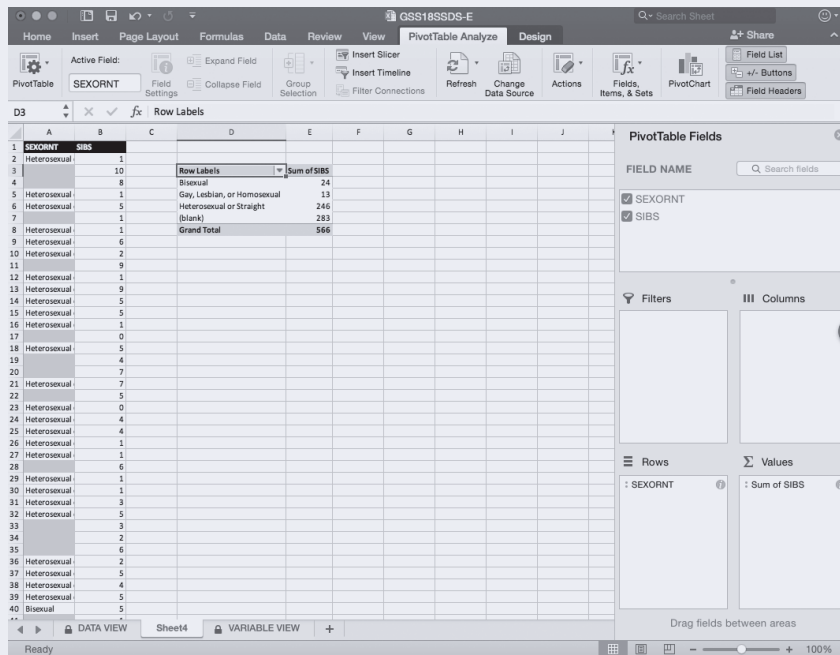
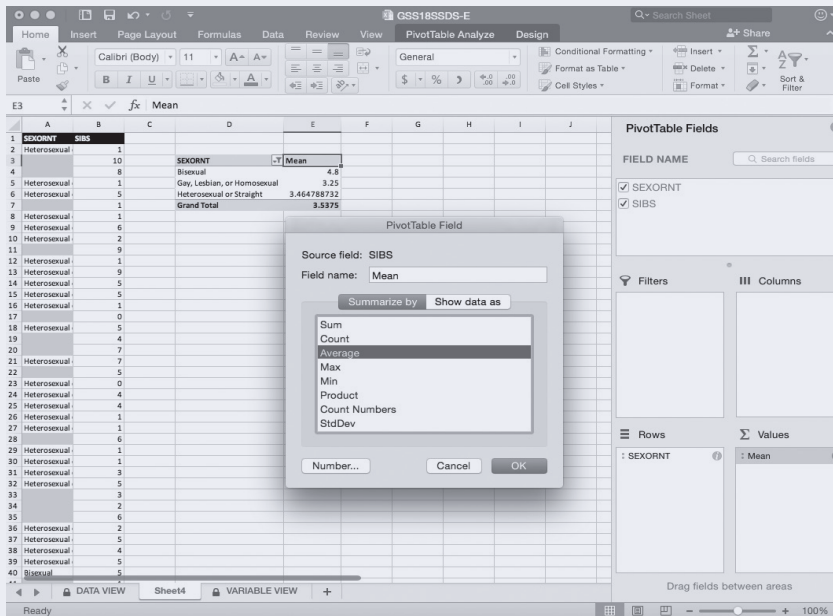


Figure 3.19



EXCEL PROBLEMS [GSS18SSDS-E]

- E1. Examine respondent's highest year of school completed (EDUC).
 - a. Using Excel's Descriptive Statistics command, produce the mode, median, and mean for EDUC.
 - b. Describe the mode, median, and mean for EDUC.
 - c. Compare the mean and median for EDUC. Is this distribution skewed? If so, positively or negatively? How do you know?
- E2. Are married people happy in their marriages?
 - a. Using Excel's PivotTable command, create a frequency distribution of happiness of marriage (HAPMAR).
 - b. What is the level of measurement of HAPMAR?
 - c. What is the mode? The median?
 - d. What variables might help us understand why some folks are very happy in their marriage while others are not too happy? Explain.
- E3. Let's examine a respondent's highest year of school completed (EDUC) by happiness of marriage (HAPMAR).
 - a. Using Excel's PivotTable command, produce the mean number of years of school completed for each category of happiness of marriage.
 - b. Which category of HAPMAR has the highest mean?

- c. Which category of HAPMAR has the lowest mean?
- d. Does EDUC vary by HAPMAR?

SPSS DEMONSTRATIONS [GSS18SSDS-B]

Demonstration 1: Producing Measures of Variability With Frequencies

Except for the IQV, the SPSS Frequencies procedure can produce all the measures of variability we've reviewed in this chapter. (SPSS can be programmed to calculate the IQV, but the programming procedures are beyond the scope of our book.)

We'll begin with Frequencies and calculate various statistics for AGE. If we click on *Analyze, Descriptive Statistics, Frequencies*, then on the *Statistics* button, we can select the appropriate measures of variability.

The measures of variability available are listed in the Dispersion box at the bottom of the dialog box. We've selected the standard deviation, variance, and range, plus the mean and median (in the Central Tendency box) for reference. In the Percentile Values box, we've selected Quartiles to tell SPSS to calculate the values for the 25th, 50th, and 75th percentiles. SPSS also allows us to specify exact percentiles in this section (such as the 34th percentile) by typing a number in the box after "Percentile(s)" and then clicking on the *Add* button.

Earlier, we had seen the frequency table for the variable AGE, so after clicking on *Continue*, we click on *Format* to turn off the display table. This is done by clicking on the button for "Suppress tables with many categories" (see Figure 3.9). There are other formatting options here that you may explore later when using SPSS.

Click on *Continue*, then *OK*, to run the procedure. SPSS produces the mean and the other statistics we requested (Figure 3.10). The range of age is 71 years (from 18 to 89). The standard deviation is 17.993, which indicates that there is a moderate amount of dispersion in the ages. The variance, 323.743, is the square of the standard deviation (17.993).

Figure 3.9 Format Dialog Box

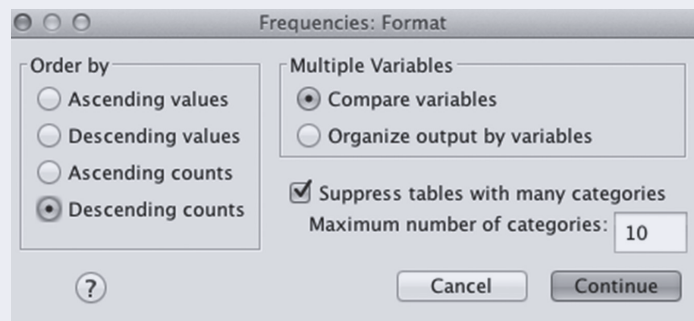


Figure 3.10

Statistics		
Age of respondent		
N	Valid	1495
	Missing	5
Mean		48.69
Median		48.00
Std. Deviation		17.993
Variance		323.743
Range		71
Percentiles	25	34.00
	50	48.00
	75	63.00

The value of the 25th percentile is 34, the value of the 50th percentile (which is also the median) is 48, and the value of the 75th percentile is 63. Although Frequencies does not calculate the IQR, it can easily be calculated by subtracting the value of the 25th percentile from the 75th percentile, which yields a value of 29 years. Compare this value with the standard deviation.

Demonstration 2: Producing Variability Measures and Box Plots With Explore

Another SPSS procedure that can produce the usual measures of variability is Explore, which also produces box plots. The Explore procedure is located in the *Descriptive Statistics* section of the *Analyze* menu. In its main dialog box (Figure 3.11), the variables for which you want statistics are placed in the Dependent List box. You have the option of putting one or more nominal variables in the Factor List box; Explore will display separate statistics for each category of the nominal variable(s) you've selected.

Place the variable EDUC (highest year of school completed) in the Dependent box and SEX in the Factor box to provide separate output for males and females. Click *OK*. By default, Explore will produce statistics and plots, so we don't need to make any other choices. Although our request will not produce percentiles or create a histogram, Explore has options to do both plus several other tasks.

Selected output for males is shown in Figure 3.12. Although not replicated here, you'll notice that the first table is the Case Processing Summary Table. It indicates that 685 males answered this question. The valid sample of females is also reported, 813. Based on the second table, Descriptives, we know that for males, the mean number of years of education is 13.70, and the median is 14.00. The standard

Figure 3.11

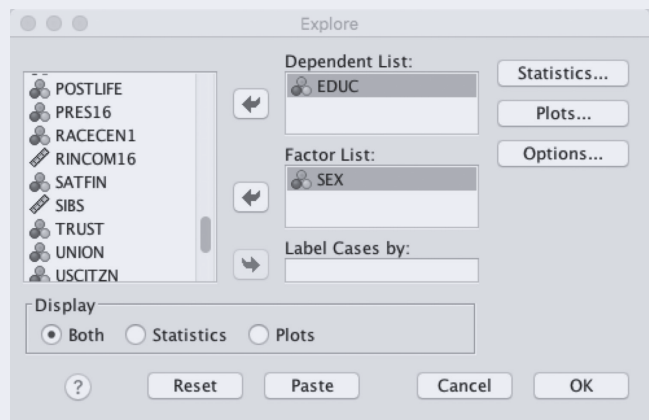


Figure 3.12

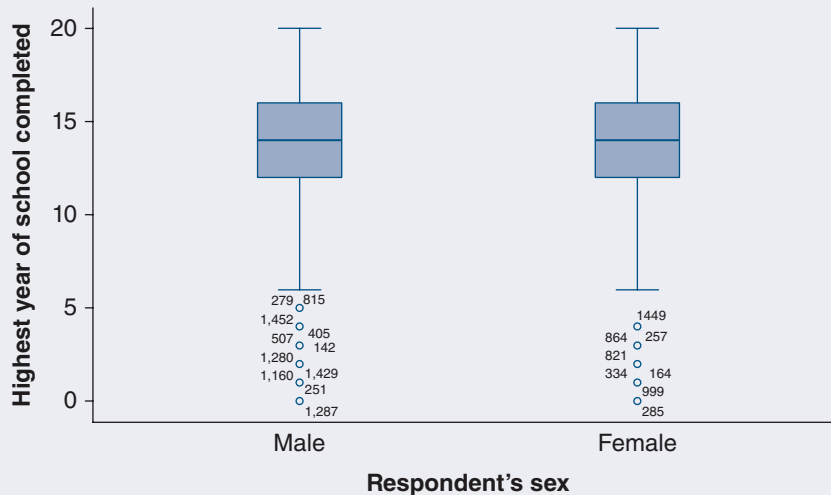
Descriptives					Statistic	Std. Error
Respondents sex						
Highest year of school completed	Male	Mean			13.70	.118
		95% Confidence Interval for Mean	Lower Bound		13.47	
			Upper Bound		13.93	
		5% Trimmed Mean			13.80	
		Median			14.00	
		Variance			9.518	
		Std. Deviation			3.085	
		Minimum			0	
		Maximum			20	
		Range			20	
		Interquartile Range			4	
		Skewness			-.604	.093
		Kurtosis			1.951	.187

deviation is 3.085, the range is 20, and the IQR is 4, which is quite narrow compared with the range. (A stem-and-leaf plot—another way to visually present and review data—is also displayed by default. The option for the stem-and-leaf plot can be changed so that it will not be displayed.)

Although not displayed here, the mean number of years of education for females is 13.73; the median is 14. The standard deviation is 3.002, the IQR is 4, and the range is 20—values similar to those for males.

Explore displays separate box plots for males and females in the same window for easy comparison (see Figure 3.13). Although the SPSS box plot has some differences from those discussed in this chapter, some things are the same. The solid dark line is the value of the median. The width of the shaded box (in color on the screen) is the IQR (4 for males, 4 for females).

Figure 3.13



Note that SPSS only extends whiskers from the box edges to $1\frac{1}{2}$ times the box width (the IQR). If there are additional values beyond $1\frac{1}{2}$ times the IQR, SPSS displays the individual cases. Those that are somewhat extreme ($1\frac{1}{2}$ to 3 box widths from the edge of the box) are marked with an open circle; those considered very extreme (more than 3 box widths from the box edge) are marked with an asterisk.

SPSS PROBLEMS [GSS18SSDS-B]

- S1. Use the Frequencies procedure to investigate the variability of female respondents' current age (AGE). Navigate to *Data* and *Select Cases*. Choose "If condition is satisfied" and then click on the "if" button. Place SEX in the open box followed by equals 2 (SEX = 2). Select *Continue* and then *OK*. Click on *Analyze*, *Descriptive Statistics*, *Frequencies*, and then *Statistics*. Select the appropriate measures of variability. How would you describe the distribution of AGE for women?
- S2. Using the Explore procedure, separate the statistics for HRSRELAX (number of relaxation hours per week) and RINCOM16 (recoded income) for men and women, selecting SEX as a factor variable in the Explore window. Click on *Analyze*, *Descriptive Statistics*, *Explore*, and then insert HRSRELAX and RINCOM16 into the Dependent List and SEX in the Factor List. What differences exist in the hours of relaxation and income between men and women? Assess the differences between men and women based on measures of central tendency and variability. (Reset select Cases to "All cases.")

- S3. Repeat the procedure in Exercise 2, investigating the dispersion in RINCOM16 (recoded income). Select your own factor (nominal) variable to make the comparison (such as CLASS, RACE, or some other factor). Click on *Analyze*, *Descriptive Statistics*, *Explore*, and insert RINCOM16 into the Dependent List and your factor variable of choice in the Factor List. In a paragraph or two, use appropriate measures of variability to summarize the results.
- S4. Use the Explore procedure to study the variability of relaxation hours (HRSRELAX) among social class categories in the GSS sample.
 - a. Is there a difference between the four groups in the variability of relaxation hours?
 - b. Write a short paragraph describing the box plot that SPSS created as if you were writing a report and had included the box plot as a chart to support your conclusions about the difference between social classes in the variability (and central tendency) of hours of relaxation.

EXCEL DEMONSTRATIONS [GSS18SSDS-E]

Demonstration 1: Producing Measures of Variability

We will use the EDUC data (highest year of school completed) to demonstrate how to use Excel to produce measures of variability for an interval-ratio variable. The steps should be familiar to you if you've worked through the Excel Demonstrations we offered in Chapter 3.

Copy the EDUC data from the protected Data View sheet and paste it into a new Excel sheet. Navigate to Excel's Data tab and select *Data Analysis*. A window of Analysis Tools will appear. Select *Descriptive Statistics* and then *OK*.

Click in the empty box next to "Input Range" and highlight the column of EDUC data from A2 to A136. Do not select A1 for it contains the variable name, EDUC. Under "Output options," select *Summary statistics*. Click in the empty box next to "Output Range," and then select any cell in the current sheet you are working in. This will tell Excel where to place the Descriptive Statistics box it will generate. In our example, we've chosen for the output box to begin in cell C2. Click *OK*.

Excel will produce a Descriptive Statistics table that includes measures of central tendency (discussed in Chapter 3) for EDUC as well as measures of variability. We will focus on the standard deviation, sample variation, and range—all appropriate measures of variability for an interval-ratio variable. The range in education is 19 years of schooling (from 0 to 19). The standard deviation is 2.99, which indicates there is a reasonable amount of dispersion in EDUC. The variance (8.96) is the square root of the standard deviation (2.99). Note, for our discussion, we rounded the standard deviation and variance values for EDUC to the second decimal place. Also notice we extended the width of column C, and we replaced the automatically generated table title "Column 1" with "EDUC, GSS18SSDS-E."

Figure 3.14

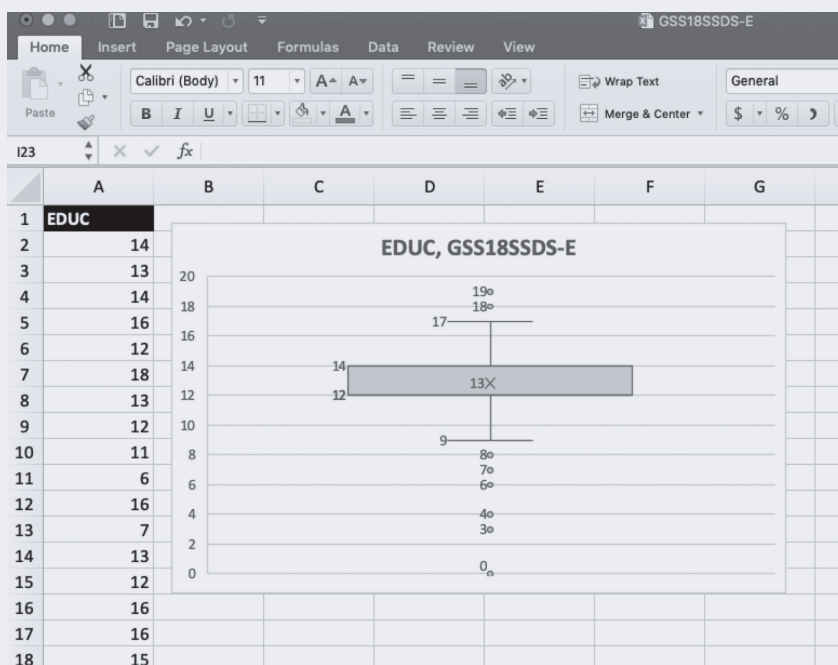
	A	B	C	D	E
1	EDUC				
2	14		EDUC, GSS18SSDS-E		
3	13				
4	14		Mean	12.8296296	
5	16		Standard Error	0.25767184	
6	12		Median	12	
7	18		Mode	12	
8	13		Standard Deviation	2.99387619	
9	12		Sample Variance	8.96329464	
10	11		Kurtosis	2.83894178	
11	6		Skewness	-0.8221172	
12	16		Range	19	
13	7		Minimum	0	
14	13		Maximum	19	
15	12		Sum	1732	
16	16		Count	135	
17	16			-7.605E-52	
18	15				
19	13				

Demonstration 2: Creating a Box-and-Whisker Plot

Let's continue working with the variable EDUC, but this time, we will ask Excel to create a box-and-whisker chart, also called a box plot. If you haven't already done so from Demonstration 1, copy the EDUC data from the protected Data View sheet and paste it into a new Excel sheet. Highlight the column of EDUC data from A2 to A136. Do not select A1 for it contains the variable name, EDUC. In the main Excel toolbar, click *Insert* → *Chart* → *Box and Whisker*. A box plot of EDUC will appear.

In the Chart Design Excel tab, you will be able to edit the box plot you created. We will select *Add Chart Element* → *Data Labels* → *Left*. This will impose the various data points onto the box plot. You can also use the *Add Chart Element* function to alter which data axes are shown, gridlines, titles, and more. We've titled our box plot EDUC, GSS18SSDS-E.

Figure 3.15



The shaded box visually shows us the first quartile ($Q1=12$) as well as the third quartile ($Q3=14$). It also shows us the median (13). This median is different from the median we presented in Demonstration 1 above, for the Box and Whisker Excel function does not consider values it deems outliers when generating this chart. Take, for example, the whiskers, the lines extending either above or below the box. In theory, these whiskers should extend from the highest value of our data (19) to the lowest value in our data (0). But, instead, they extend from 17 to 9. Excel is deeming values of our EDUC data that are beyond 9 to 17 outliers. Thus, it is important to carefully approach any values offered in a box plot, which ought to be easy given that the goal of any visual display of data—be it a box plot, pie chart, bar graph, and so on—is to offer readers a visual, not numeric, representation of data.

EXCEL PROBLEMS [GSS18SSDS-E]

- E1. Examine the measures of central tendency and measures of variability for mother's highest year of school completed (MAEDUC).
 - a. Use Excel to produce a Descriptive Statistics table of MAEDUC.
 - b. Report and describe the measures of central tendency and measures of variability.

- c. Ask Excel to create a box plot for MAEDUC.
 - d. Does there appear to be variability in mother's highest year of school completed?
- E2. Investigate the measures of variability for respondent's age (AGE) and respondent's age when their first child was born (AGEKDBRN).
 - a. Create a Descriptive Statistics table for AGE and AGEKDBRN.
 - b. Report and describe the measures of variability for both variables.
 - c. Was there more variability in respondent's age or respondent's age when their first child was born? Offer an explanation or why one variable had more variability than the other.