SPSS DEMONSTRATION [GSS18SSDS-B]

Selecting a Random Sample With SPSS

In this chapter, we've discussed various types of samples and the definition of the standard error of the mean. Usually, data entered into SPSS have already been sampled from some larger population. However, SPSS does have a sampling procedure that can take random samples of data. Systematic samples and stratified samples can also be drawn with SPSS, but they require the use of the SPSS command language.

When might it be worthwhile to use the SPSS Sample procedure? One instance is when doing preliminary analysis of a very large data set. For example, if you worked for your local hospital and had complete data records for all patients (tens of thousands), there would be no need to use *all* the data during initial analysis. You could select a random sample of individuals and use the subset of data for preliminary analysis. Later, the complete patient data set could be used for completing your final analyses.

To use the Sample procedure, click on *Data* from the main menu, then click on *Select Cases*. The opening dialog box (Figure 5.7) has five choices that will select a subset of cases via various methods. By default, the *All cases* button is checked. We click on the *Random sample of cases* button, then on the *Sample* button to give SPSS our specification.

Figure 5.7

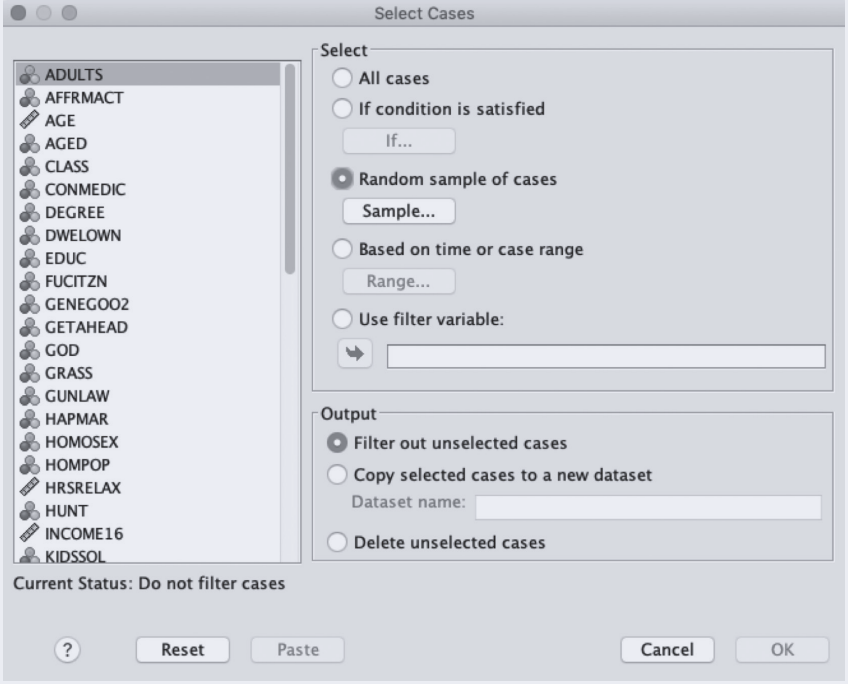
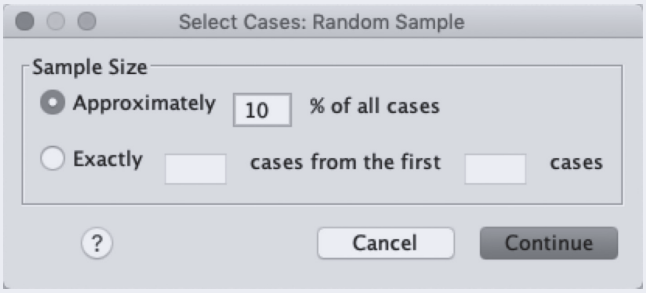


Figure 5.8



The next dialog box (Figure 5.8) provides two options to create a random sample. The most convenient one is the first, where we tell SPSS what percentage of cases to select from the larger file. Alternatively, we can tell SPSS to take an exact number of cases. The second option is available because SPSS will only take approximately the percentage specified in the first option.

We type "10" in the box to ask for 10% of the original sample of 1,500 respondents from the GSS18SSDS-B data set. Then, click on *Continue* and *OK*, as usual, to process the request.

SPSS does not delete the cases from the active data file that aren't selected for the sample. Instead, they are filtered out (you can identify them in the Data View window by the slash across their row number). This means that we can always vreturn to the full data file by going back to the Select Cases dialog box and selecting the *All cases* button.

When SPSS processes our request, it tells us that the data have been filtered by putting the words "Filter On" in the status area at the bottom of the SPSS window (the status area has many helpful messages from SPSS).

To demonstrate the effect of sampling, we ask for univariate statistics for the variable SIBS, the number of brothers and sisters one has. Click on *Analyze*, *Descriptive Statistics*, and then *Descriptives* to open this dialog box. Place SIBS in the variable list. Click on the *Options* button to select the mean, standard deviation, minimum, and maximum values. In addition, we'll add the standard error of the mean by clicking the *S.E. mean* box. Then, click *Continue* and *OK* to put SPSS to work.

The results (Figure 5.9) show that the number of valid cases is 155, or about 10% of the valid cases (1,498 were asked the SIBS question). The mean of SIBS is 3.54, and the standard error of the mean is .236. If we repeat the process, this time asking for a 25% sample, we obtain the results shown in Figure 5.10.

Your results may differ from the results presented here. We are asking SPSS to generate a random selection of cases, and you may not get the same selection of cases as we did.

How closely does the mean for SIBS from these two random samples match that of the full file? The mean for the 1,498 respondents (the other 2 respondents did not have valid responses) is 3.64 siblings. Both samples produced means and standard deviations that are within the range of the population parameters.

Figure 5.9

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Number of brothers and sisters	155	0	12	3.54	.236	2.937
Valid N (listwise)	155					

Figure 5.10

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Number of brothers and sisters	338	0	21	3.74	.169	3.103
Valid N (listwise)	338					

SPSS PROBLEM [GSS18SSDS-A]

S1. Using GSS18SSDS-A, repeat the SPSS demonstration, selecting 25%, 50%, 75%, and 100% samples and requesting descriptives for MAEDUC and PAEDUC. Compare your descriptive statistics with descriptives for the entire sample. What can you say about the accuracy of your random samples?

EXCEL DEMONSTRATION [GSS18SSDS-E]

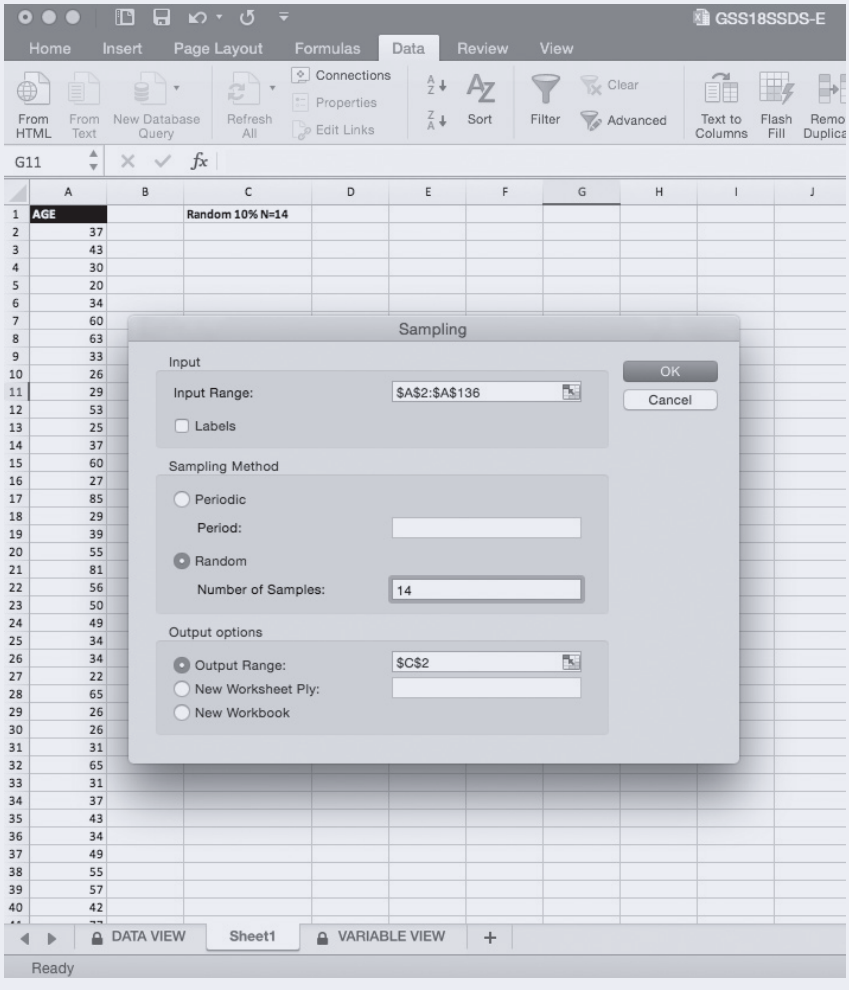
Selecting a Random Sample With Excel

We will use the AGE data (age of respondent) to demonstrate how to use Excel to produce *Z* scores. Copy the AGE data from the protected Data View sheet and paste it into a new Excel sheet. Our copied data were placed in our new sheet from A1 to A136. In cell C1 of this new sheet, we've typed, in bold font, "Random 10% N=14." We also extended the width of column C. Since we are working with 135 respondents, we know that a sample of 10% would be just under 14 cases (135 × .10 = 13.5). We are now ready for Excel to produce a random sample of respondents from our pool of 135 total cases.

Use the *Data Analysis* function to obtain your random sample. Navigate to Excel's Data tab and select *Data Analysis*. A window of Analysis Tools will appear. Select *Sampling* and then *OK*.

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 "Sampling Method," select *Random* and in the box next to *Number of Samples*, insert 14. By entering 14 in this box, we are asking Excel to randomly select 14 respondents from our group of 135 respondents—we are thus creating a new sample 10% of our original sample size ($N = 135$). Under "Output options," select *Output Range*. Click in the box next to *Output Range*. Next select any cell in the current sheet you are working in. This will tell Excel where to place the random sample it will generate. In our example, we've chosen for the random sample to begin in cell C2 directly below our heading "Random 10% N=14." Click *OK*.

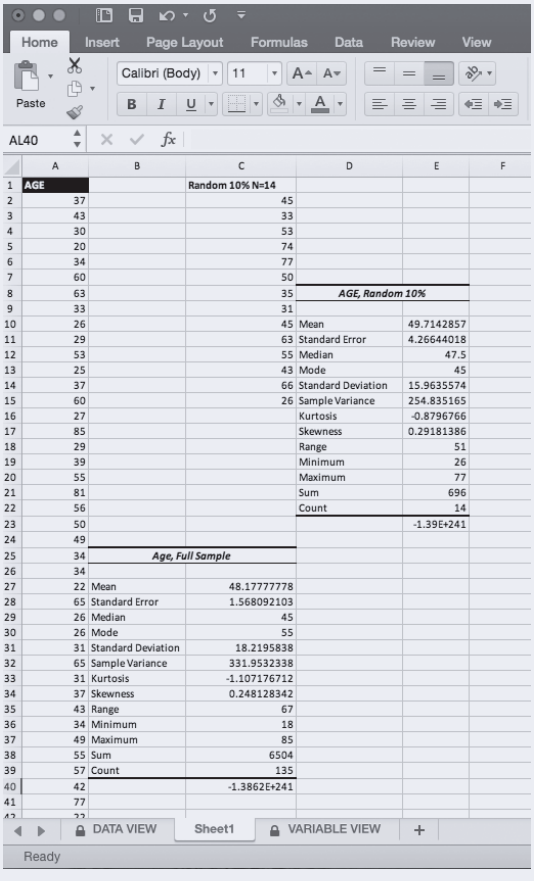
Figure 5.11



You will see a random sample of 14 ages from the original sample of 135 respondents. Because this process is random, your sample will be different from the one we've generated in this demonstration. We will use Excel's *Data Analysis* function to create a table of descriptive statistics for both the original sample as well as our new sample of 14 respondents. For assistance working with *Data Analysis*, see the Excel Demonstrations in Chapters 3 and 4.

In Figure 5.12, you can see both the descriptive statistics table of AGE for the full sample as well as a descriptive statistics table of AGE for our random sample of 14 respondents. The mean for the full sample is 48.18 with a standard deviation of 18.22. Your descriptive statistics for the full sample of AGE should be the same as ours. As for our new sample of 14 respondents, we see a mean of 49.71 with a standard deviation of 15.96. Because we are working with a random sample of 14 participants from our full sample, our descriptive statistics for our new sample will not be the same as the descriptive statistics for your new sample. The 14 respondents

Figure 5.12



	A	B	C	D	E	F
1	AGE		Random 10% N=14			
2	37		45			
3	43		33			
4	30		53			
5	20		74			
6	34		77			
7	60		50			
8	63		35	*AGE, Random 10%*		
9	33		31			
10	26		45	Mean	49.7142857	
11	29		63	Standard Error	4.26644018	
12	53		55	Median	47.5	
13	25		43	Mode	45	
14	37		66	Standard Deviation	15.9635574	
15	60		26	Sample Variance	254.835165	
16	27			Kurtosis	-0.8796766	
17	85			Skewness	0.29181386	
18	29			Range	51	
19	39			Minimum	26	
20	55			Maximum	77	
21	81			Sum	696	
22	56			Count	14	
23	50				-1.39E+241	
24	49					
25	34	*Age, Full Sample*				
26	34					
27	22	Mean	48.17777778			
28	65	Standard Error	1.568092103			
29	26	Median	45			
30	26	Mode	55			
31	31	Standard Deviation	18.2195838			
32	65	Sample Variance	331.9532338			
33	31	Kurtosis	-1.107176712			
34	37	Skewness	0.248128342			
35	43	Range	67			
36	34	Minimum	18			
37	49	Maximum	85			
38	55	Sum	6504			
39	57	Count	135			
40	42		-1.3862E+241			
41	77					

Excel randomly chose for our new sample will not be the same as the 14 respondents Excel randomly chooses to make up your new sample.

You are ready to compare the mean AGE for the full sample with the mean AGE of your new sample. In our example, we can see that our new sample has a mean (49.71) and standard deviation (15.96) of AGE similar to the mean (48.18) and standard deviation (18.22) of AGE for our full sample.

EXCEL PROBLEM [GSS18SSDS-E]

E1. Using GSS18SSDS-E, repeat the Excel demonstration, selecting 10%, 20%, and 30% samples of AGEKDBRN (respondent's age when first child was born).

   a. Using Excel's *Data Analysis* function, create a *Descriptive Statistics* table for each of the three new samples you generated.

   b. What is the mean and standard deviation of AGEKDBRN for each of your new samples?

   c. Create a *Descriptive Statistics* table for the full sample ($N = 135$). What are the mean and standard deviation of the full sample?

   d. How do your new samples compare to the full sample?