Chapter 9

Sampling Distributions

Statistical inference is a process of drawing conclusions about a population value (a parameter) based on a value computed from a random sample (a statistic). Each sample drawn from the population may yield a different statistic, however. The *sampling distribution* is the distribution of all possible values of a particular statistic, each with an associated probability. In this chapter, we use SPSS to simulate drawing random samples from a population, computing a particular statistic, and constructing its sampling distribution. In particular, we shall simulate the sampling distribution of a single observation drawn from a standard normal population distribution, the distribution of the sum of two observations drawn at random, and the distribution of the mean of 100 observations drawn from a population.

9.1 SAMPLING FROM A POPULATION

A random sample of a given size is a group of members of the population selected so that all groups of that size are equally likely to appear in the sample. Randomization ensures an equal chance of selection for every possible subset. A variable assessed for each member of the sample (selected through a random process) is called a random variable.

Random Samples

When we make SPSS toss a coin 10 times (Section 7.2), we are actually taking a random sample from a Bernoulli (two-valued) probability distribution. Now we shall repeat this procedure, but will draw a random sample of 50 observations from a standard normal distribution. Because the procedure is so similar to those described in Chapters 7 and 8, the following instructions are abbreviated.

1. Open a new Data window (see Section 7.1).
2. Enter “999” for the first 50 cases of the first column.
3. Click on **Transform** from the menu bar.
4. Click on **Compute** from the pull-down menu.
5. In the Compute dialog box, name the Target Variable “sample.”
6. Highlight “Random Numbers” from Function Group box and “RV.Normal” from the Functions and Special Variables box and move it into the Numeric Expression box by clicking on the **up arrow button**.
7. Choose a standard normal distribution by replacing the first and second question marks with **0** and **1**, respectively.
8. Click on **OK**.

Now create a histogram of the “sample” variable (see Section 2.2 for details on creating histograms). It should resemble a standard normal distribution, but will differ somewhat because it is a sample. Figure 9.1 shows one possible histogram. This histogram would appear more normal if we had taken a larger number of samples. Because your random sample is different, your histogram will vary somewhat. Note, for instance, that the mean for this sample is –.11, which is close to the mean of 0, and the standard deviation of this random sample (0.90) is close to that of a standard normal distribution, 1.

9.2 SAMPLING DISTRIBUTION OF A SUM AND OF A MEAN

It is possible to use SPSS to simulate the sampling distribution of any statistic computed from a random sample from the population. In this section, we obtain the sampling distribution of the sum of two variables, each having a discrete uniform distribution with minimum of 1 and maximum of 6. This is analogous to constructing the sampling distribution resulting from rolling two dice repeatedly and tabulating the sum of the pips on each roll.

We will direct SPSS to roll dice, one at a time, for a total of 50 pairs of rolls. Next, we will compute the sum of each roll (e.g., the first roll for die one + the first roll for die two, etc.) and then examine the frequency distribution of this new variable.

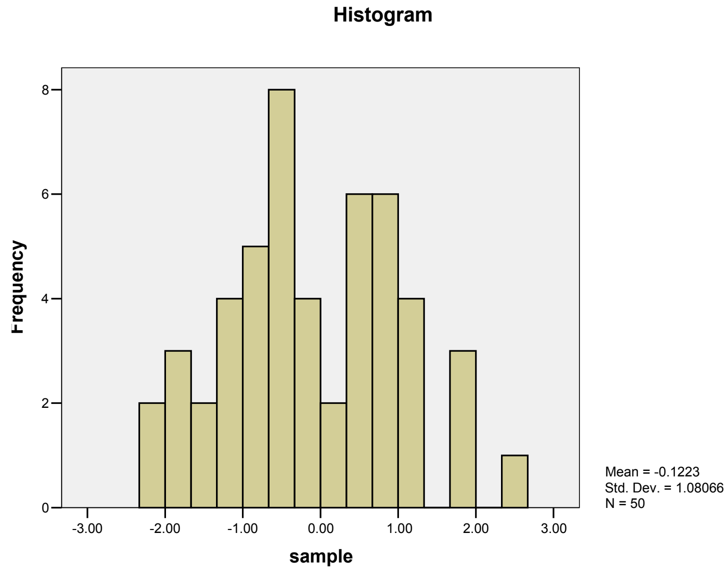


Figure 9.1 Histogram from a Standard Normal Distribution

1. Open a new Data window, and type "999" in the first 50 rows of the first column.
2. Click on **Transform** from the menu bar.
3. Click on **Compute** from the pull-down menu.
4. Name the target variable "die1."
5. In the Numeric Expression box, create the expression "**RND(RV.UNIFORM(1,6))**." (See Section 7.1.)
6. Click on **OK**.
7. Repeat steps 2–6 to compute another variable named "die2."
8. Click on **OK.**
9. Compute a "total" variable, which is the sum of the two sample variables. Click on **Transform** from the menu bar.
10. Click on **Compute** from the pull-down menu.
11. Click on the **Reset button** in the Compute Variable dialog box.
12. Next, name the target variable "**total**."
13. Click on the "die1" variable in the lower left box and move it to the Numeric Expression box with the **right arrow button**, click on the + from the calculator pad, then click on and move the "die2" variable to the Numeric Expression box with the **right arrow button**.
14. Click on **OK**.

15. Obtain a frequency distribution and histogram of the “total” variable (see Section 2.2).

Your output should be similar to that shown in Figure 9.2.

9.3 THE NORMAL DISTRIBUTION OF SAMPLE MEANS

The Central Limit Theorem

Most of the inferential procedures discussed in Part IV of this manual are based on the Central Limit Theorem (CLT) that states that the sampling distribution of the sample mean is approximately a normal distribution. This is true regardless of the parent distribution from which the samples are drawn, as long as the sample size (n) is large. Using the CLT allows us to make probability statements about a sample mean without actually observing the entire population from which it was drawn.

It is possible to use SPSS to illustrate the Central Limit Theorem. The process is not straightforward, but because the CLT is one of the most important principles of inferential statistics, working through this example may help you to understand the concepts more fully.

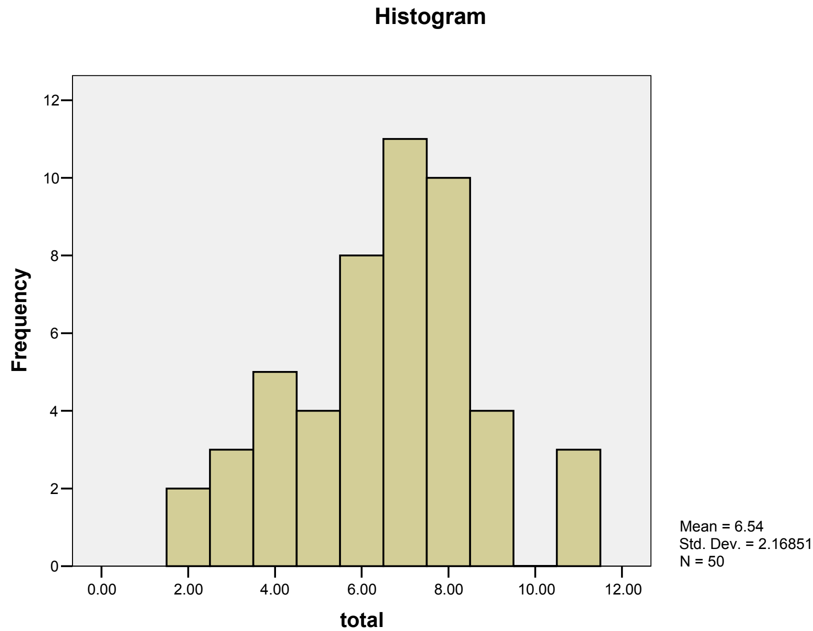
To illustrate this theorem, we will first have to obtain a random sample of size n (e.g., 50) from a specific distribution (e.g., discrete uniform(1,10)), and calculate the mean of the 50 observations. Then we will repeat this process many times (e.g., 99 more times), and inspect the frequency distribution and histogram of the sample means.

The procedures used to obtain a random sample from a specific distribution are given in Chapter 7. The tedious part of this process involves repeating the sampling 100 times. We have completed this step for you, and the results are saved in the data file “clt.sav.” There are 100 variables, u1 to u100, which represent the 100 times that SPSS drew random samples of size 50. Therefore, at present we have a [50 rows × 100 columns] matrix representing [sample size × number of samples].

In order to get a histogram of the means, we first need to calculate the mean of each of the 100 samples (of the columns). We could direct SPSS to compute the mean of each of the “u” variables separately, but we would then have to manually input each of these means into another column. A less time consuming method is to transform the matrix in such a way as to make SPSS keep track of the means. To do so, we have to transpose the matrix — interchange the rows and the columns — and then compute the means. Open the “clt.sav” data file, and then:

total

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	2	4.0	4.0	4.0
	3.00	3	6.0	6.0	10.0
	4.00	5	10.0	10.0	20.0
	5.00	4	8.0	8.0	28.0
	6.00	8	16.0	16.0	44.0
	7.00	11	22.0	22.0	66.0
	8.00	10	20.0	20.0	86.0
	9.00	4	8.0	8.0	94.0
	11.00	3	6.0	6.0	100.0
	Total	50	100.0	100.0	



o **Figure 9.2** Frequency Distribution and Histogram of Total

1. Click on **Data** from the menu bar.
2. Click on **Transpose** from the pull-down menu.

3. Highlight all of the variable names (click on **u1**, hold the mouse button down, and drag down to the name of the last variable in the list).
4. Move the variable names to the Variable(s) box by clicking on the **upper right arrow button**.
5. Click on **OK**.

You should now have a transposed data file with a [100 × 50] matrix (excluding the first column). The rows now represent the 100 samples drawn, and the columns represent the 50 draws in each sample.

We can now compute the mean of each of the rows. To do this:

1. Click on **Transform** from the main menu bar.
2. Click on **Compute** from the pull-down menu.
3. Enter "mean" in the Target Variable box.

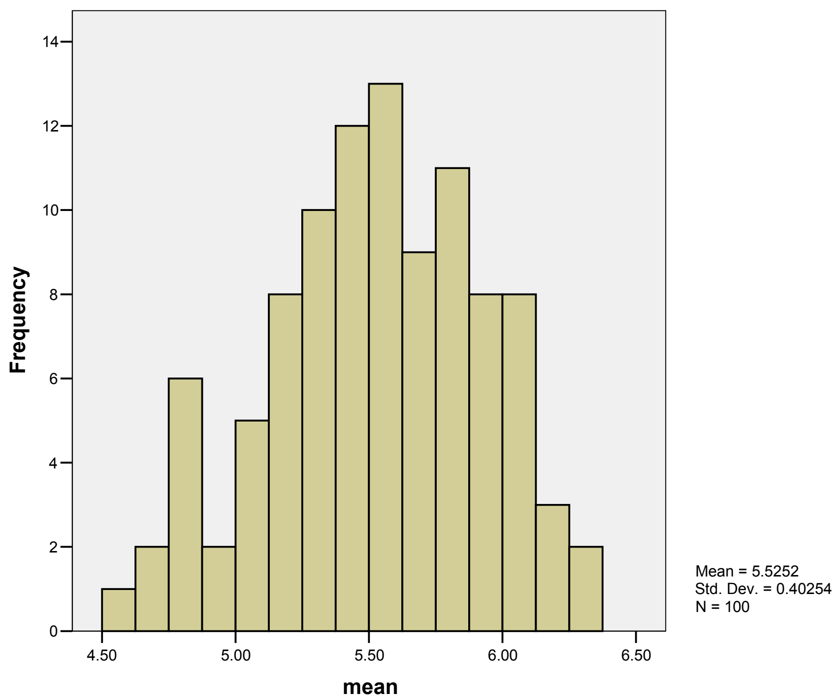


Figure 9.3 Histogram of Means

4. In the Numeric Expression box, create the expression: "**mean(var001 to var050)**." (The easiest way to do this is to type it in the box.)
5. Click on **OK**.

This should create a new variable, "mean," which contains the means of the 100 samples. Create a histogram of this new variable. Figure 9.3 displays this graph. Although the distribution of the means is not exactly normal, it is close to normal. Normality would be improved if the parent distribution were more normal-like or if we had drawn more than 100 samples.

Chapter Exercises

9.1 Use SPSS to:

- **a.** Obtain 10 random samples of size 8 from a standard normal distribution, and then compute the mean of each of these samples. Are all or any of the sample means equal to 0, the mean of the population? Sketch a histogram of these means and summarize the distribution.
- **b.** Repeat part (a), using 10 random samples of size 40.
- **c.** Compare your results in part (a) and part (b). What principle do these results illustrate?

9.2 Use SPSS to:

- **a.** Obtain three random samples of size 10 from a Bernoulli distribution with $p = .5$.
- **b.** Compute a new variable, which represents the sum of these three original variables.
- **c.** Based on probabilities, sketch the histogram you would expect to obtain for this composite variable.
- **d.** Create the histogram with SPSS and compare the actual results to those in part (c). How would you explain the differences?

Part IV

Inferential Statistics