CHAPTER 13 Appendix

Sampling Distributions

Your text discusses creating estimated sampling distributions using simulation. All of these technologies can create simulated distributions using random number generators, but not all can sample from a specified set. Also, some of the technologies require you to create all of the parts for the sampling distribution, while others include built-in modules that greatly simplify the process of creating the sampling distribution.

Recall that you will want to repeatedly obtain samples of size n. The number of samples (repetitions) is your decision; let's call the number of repetitions k. The value of k can be 100, 1000, or something else, but don't confuse k with n. The value n is the number of individuals in each sample.

The next step is to use these k samples of size n to find the statistic of interest. Here we are interested in the sample mean, $\bar{x}$. You will need to find the sample mean for each of the k samples. These k sample means are now your new data set that you want to explore because you have simulated the sampling distribution of the sample mean of samples of size n.

Calculating probabilities for the sample mean and the sample proportion employ the Normal distribution functions as detailed in Chapter 11. Be sure to check that the necessary conditions hold:

- The sample mean will **always** have a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ **if the population is Normal**. The **Normal distribution will be approximate** if the population has any distribution and the sample size is at least 30 (by the central limit theorem).
- The sample proportion will have an approximately Normal distribution with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ if **both** $np \geq 10$ and $n(1-p) \geq 10$.

Excel

1. Enter the column of values and the probability of selecting each ($1/n$ in decimal form) in two columns. The values must be in the left-hand column.
2. **Data ➔ Data Analysis**
3. Select **Random Number Generator** from the menu box.
4. Enter k, the number of columns (samples) to generate, in the box labeled **Number of Variables**.
5. Enter n, the number of observations per sample in the box labeled **Number of Random Numbers**.
6. Specify the range of cells containing the values to select from and their probabilities in the box labeled **Value and Probability Input Range**.
7. Specify an output region; if generating more than one sample, it is a good idea to use the **New Worksheet Ply** default option.

Excel should take you to the new worksheet automatically. You still need to create the statistic of interest, such as the sample mean. To calculate the mean of each sample:

1. Click in an empty row below the first simulated sample (this should be in column A). Enter the command

   `=AVERAGE(a1:an)`

   where "n" is the number of observations in the sample.

2. Use the cursor to grab the square at the lower right of the cell just created and fill that row across all the *k* samples.
3. The simulated sampling distribution of the statistic (here, the mean) is this row. It will be easier to explore these data using knowledge from Chapters 1 and 2 if the row is transposed into a column. To transpose a row to a column, copy the row, click into the cell that you wish to copy the cells below (usually in a new worksheet), **HOME ➔ Paste Special ➔** click in the radio button next to Value and click in the box next to **Transpose ➔ OK**.

At this point, you can graph the distribution or find the mean of the means using the column (or the row, if you didn't perform the transpose step) just created as the input range.

To simulate the distribution of a sample proportion, enter **0** (denoting a "failure") and **1** (denoting a "success") in a column, with their specified probabilities in a second column. Use the previously given procedure to then draw samples from that population. The statistic of interest is the sample proportion, $\hat{p}$. You will need to sum the columns to find the number of successes in each column. Then divide the number of successes by the sample size, *n*, to find the sample proportion. You should have *k* sample proportions from samples of size *n*.

A newer version of the JMP Teaching Demos allows you to do this simulation for any arbitrary sample size with any population size. If your version wants to set the sample size to 1 with small "populations," go to jmp.com/tools for the latest version. These teaching demonstrations function like applets, so readers are encouraged to make changes and explore the different options to understand the sampling distribution of the sample mean or the sampling distribution of the sample proportion. Be sure to **Reset Demo** when you change the population or some other value.

1. Enter the column of values to sample from.
2. **Help ➔ Teaching Demos ➔ Sampling Distribution of Sample Mean**
3. In the dialog box, select **My Data** in the **Population Shape** pull-down.
4. Make certain the correct variable name appears in the **Name of Variable** box.
5. Enter the **Sample Size**, *n*, and **Number of Samples**, *k*.

To simulate the distribution of a sample proportion:

1. **Help ➔ Teaching Demos ➔ Sampling Distribution of Sample Proportion**
2. Enter the desired **Population Proportion.**
3. Enter the desired **Sample Size**, *n*, and **Number of Samples**, *k*, in their respective boxes.
4. Click **Draw Additional Samples**. JMP will create a histogram of the sample values and display the mean and standard deviation of the sample proportions.

Minitab

For this chapter, Minitab users may find it useful to Enable Commands in the Sessions window. See Chapter 1.

To simulate the sampling distribution of the sample mean from a specified population:

1. Enter the values (population) in a column.
2. **Calc ➔ Random Data ➔ Sample From Columns**
3. Enter the number of rows to sample (the desired sample size, *n*).
4. Select the original data and enter that into the **From columns** box.
5. Enter a **New columns** as Cx, where "x" is a number.
6. Click **OK**.

Note: Minitab will create only one sample at a time. To automate this process, use a text editor such as Notepad or Wordpad to create a command file with a .MTB extension. The file should contain the following three commands (assuming the original data are in C1):

```
Sample k2 c1 c3
Let c4(k1)=mean(c3)
Let k1=k1+1
```

Be sure you have enabled commands in the Session window. Then, at the MTB> prompt, enter the following commands:

```
MTB> Let k1=1
MTB> Let k2=n (replace the n here with your desired
sample size)
```

Then, click **Tools ➔ Run an Exec**. Enter the number of times to execute the command set (*k*), and use **Select File** to find the .MTB file you just created. Click **Open** to run the exec file. When the calculation is finished, c4 (or your destination) column will contain all of the sample means, which can then be graphed or summarized.

Note: If you use this exec file more than once, reenter the "Let k1 = 1" command before you execute the file each time! For a different sample size, you will need to reissue the "Let k2 = n" command again as well, replacing the "n" with the desired size.

To simulate the sampling distribution of the sample proportion $\hat{p}$ from a specified population, we will create two columns, Cx and Cy. The "x" and "y" in Cx and Cy are numbers. Each row in Cx will contain the number of successes in n trials with p = probability of success. Thus, each row represents a sample of size n. We will find the sample proportion by dividing the values in Cx by n, the sample size. Column Cy will contain the sampling distribution of the sample proportion. The value for k, the number of repeated samples, is the number of rows in the columns.

1. **Calc ➔ Random Data ➔ Binomial**
2. Enter k, the number of "samples" in the box labeled **Number of rows of data to generate.**
3. Enter a column number to store the results in the **Store in column(s)** box as Cx.
4. Enter n, the sample size for each sample (**Number of trials**), and the probability of "success" (**Event probability**).
5. Click **OK**.
6. Convert the observed number of successes into sample proportions with the following command (be sure commands have been enabled):

   ```
   MTB> Let Cy = Cx/n
   ```

   where "n" is the number of trials for each "sample."

Cy is the column of sample proportions that can be graphed or summarized.

SPSS

SPSS cannot create repeated random samples from a specified list (its "random sampling" scheme is intended to work only as a means of specifying smaller samples from one very large one for further analysis). You can, however, simulate the sampling distribution of a sample proportion:

1. Page down and enter some value in the first column of an empty worksheet to correspond with the number of random numbers (samples) you wish to generate (this should be in row *k*). You may have to enter interim values to be able to go down as far as you want.
2. **Transform ➔ Compute Variable**
3. Enter a destination column name.
4. In the Function Group box, select **Random Numbers**.
5. Select **RV.Binom.**
6. Enter *n* and *p* separated by a comma to replace the question marks.
7. Click **OK**.

Turn the observed number of successes per sample into sample proportions (replace "SimCol" below with your name and "n" with your numeric sample size):

8. **Transform ➔ Compute Variable**
9. Enter a destination column name.
10. Enter the function as SimCol/n in the **Numeric Expression** box.
11. Click **OK**.

CRUNCH

CrunchIt! cannot create random samples from a specified list, so it cannot be used to create a sampling distribution of the sample mean. It could be used to simulate the sampling distribution of a sample proportion.

1. **Insert ➔ Random Numbers ➔ Binomial**
2. Enter the size for each sample, *n*, and the probability of success, *p*.
3. Enter the number of samples, *k*.
4. Click **Sample**.

You will have a column that contains the number of "successes" in each trial. Convert those to sample proportions that can be summarized or graphed using the following steps:

5. **Insert ➔ Evaluate Formula**
6. Enter a formula such as

```
[Var5]/50
```

7. **Evaluate**.

Note: CrunchIt! will default to placing the first set of random numbers after the last column in the current worksheet. With a blank worksheet, that would be Var5. The square brackets around the variable name in the formula are important! The new variable Var6 is the sampling distribution of the sample proportion, which can be graphed or summarized.

TI-83/-84

TI calculators can simulate random integers (equally likely), or observations from Normal and binomial distributions. They cannot sample from lists of specific values, but you can use the binomial distribution to simulate sample proportions. However, because their processors are limited, this is not recommended if you have other technologies available.

1. Press [MATH] and arrow to PRB.
2. Press [7] for option 7:randbin(.
3. Enter the number of samples (k), the sample size (n), and the probability of a success (p) followed by a right parenthesis. Press [STO▶] and [2nd] [1] to store the samples into L1. This should look like

   randbin(100,10,.3) ➔ L1

   to simulate 100 samples of a binomial with $n = 10$ and $p = 0.3$.
4. Convert the observed numbers of successes to sample proportions stored in L2, say, by pressing [2nd] [1] /"n" [STO▶] [2nd] [2], where "n" is replaced with the sample size. Using the above example, this will look like

   L1/10 ➔ L2

Using R to simulate the distribution of the sample mean is not covered here.

It is easy to create a simulated sampling distribution for a sample proportion:

```
> x <- rbinom(k, n, p)
```

where "k" is the number of simulated samples, "n" is the number of trials for each sample, and "p" is the probability of success for each observation. Convert these to sample proportions:

```
> phats <-x/n
```

where "x" is the previous result and "n" is again the number of trials per sample. The variable phats can then be graphed or summarized, since phats represents the sampling distribution of the sample proportion.