

4

CENTRAL LIMIT THEOREM

Objectives

- Set up a spreadsheet model to examine the properties of the central limit theorem.
- Develop frequency distributions and sampling distributions, and differentiate between the two.
- Develop a bootstrap analysis of the mean for various sample sizes.
- Evaluate the relationship between standard error and sample size, and standard deviation and sample size.

Suggested Preliminary Exercise: Statistical Distribution

INTRODUCTION

You have probably come across the term “population” in your studies of biology. In the biological sense, the term “population” refers to a group of organisms that occupy a defined space and that can potentially interact with one another. The Hardy-Weinberg equilibrium principle is an example of a population-level study. In statistics the term population has a slightly different meaning. A statistical population is *the totality of individual observations about which inferences are made, existing anywhere in the world, or at least within a specified sampling area limited in space and time* (Sokal and Rohlf 1995).

Suppose you want to make a statement about the average height of humans on earth. Your statistical population would include all the individuals that currently occupy the planet earth. Usually, statistical populations are smaller than that, and the researcher determines the size of the statistical population. For example, if you want to make a statement about the length of dandelion stems in your hometown, your statistical population consists of all of the dandelions currently occurring within the boundaries of your hometown. Other examples of statistical populations include a population of all the record cards kept in a filing system, of trees in a county park, or motor vehicles in the state of Vermont.

In practice, it would be very difficult to measure the heights of *all* the individuals on earth, or even to measure *all* the dandelions in your hometown. So we take a sample from the population. A **sample** is a subset of the population that we can deal with and measure. The goal of sampling is to make scientific statements about the greater population based on the information we obtain in the sample. Quantities gathered from samples are called **statistics**.

“How many samples should I take?” and “How should I choose my samples?” are very important questions that any investigator should ask before starting a scientific study. In this exercise, we’ll consider **simple random sampling**. If you sample 10 dandelions in your hometown with the intent of making scientific statements about all of the dandelions that occupy your town, then each and every individual in the population must have the same chance of being selected as part of the sample. In other words, a simple random sample is a sample selected by a process that gives every possible sample (of that size from that population) the same chance of being selected.

Let’s imagine that you use a simple random sampling scheme to sample the stem lengths of 10 dandelions in your hometown. And let’s further imagine that the *actual* average stem length of the dandelion population in your hometown is $\mu = 10$ mm; you are trying to estimate this parameter through sampling. You carefully measure the stem length of each of the 10 sampled dandelions, and then calculate and record the mean of the sample on your computer spreadsheet. The mean you have calculated is called an **estimator**, usually designated as \bar{x} , which estimates the true population mean, μ (which in this case is 10 mm). If you plot your raw data on a graph, your graph is called a **frequency distribution**. This is a pictorial description of how frequent or common different values (in this case stem lengths) appear in the population. A frequency distribution reveals many things about the nature of your samples, including the sample size, the mean, the shape of the distribution (normal, skewed, etc.), the range of values, and modality of the data (Figure 1).

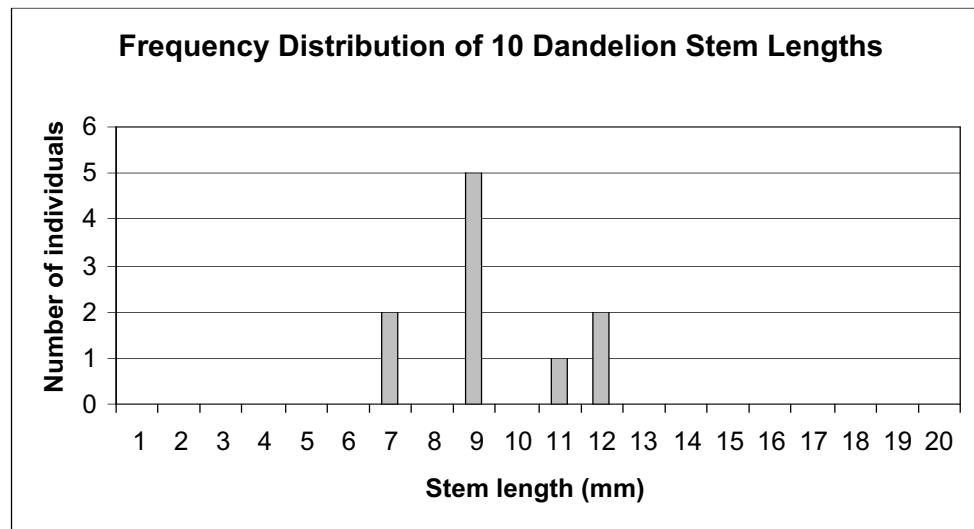


Figure 1

In the example in Figure 1, our sample of 10 dandelions had a mean value of 9.4 mm. How do you know how close your estimator is to the true mean, μ , if you can’t actually measure μ ? The central paradox of sampling is that it is impossible to know, based on a single sample, how well the sample represents μ . If you obtain another sample of 10 dandelions, and calculate a mean, you will now have two estimates of the population mean, μ . What if they are different? How will you know which is the “best” estimator?

Here is where the central limit theorem comes into play. If you repeat this sampling process and obtain a set of estimators (say, for example, 10 estimators in total, each based on a sample size of 10 dandelions), you now have a **sampling distribution** of the sample average (note the difference between the sampling distribution and the frequency distribution). The sampling distribution shows the possible values that the estimator can take and the frequency with which they occur. The standard deviation of a sampling distribution is called the standard error.

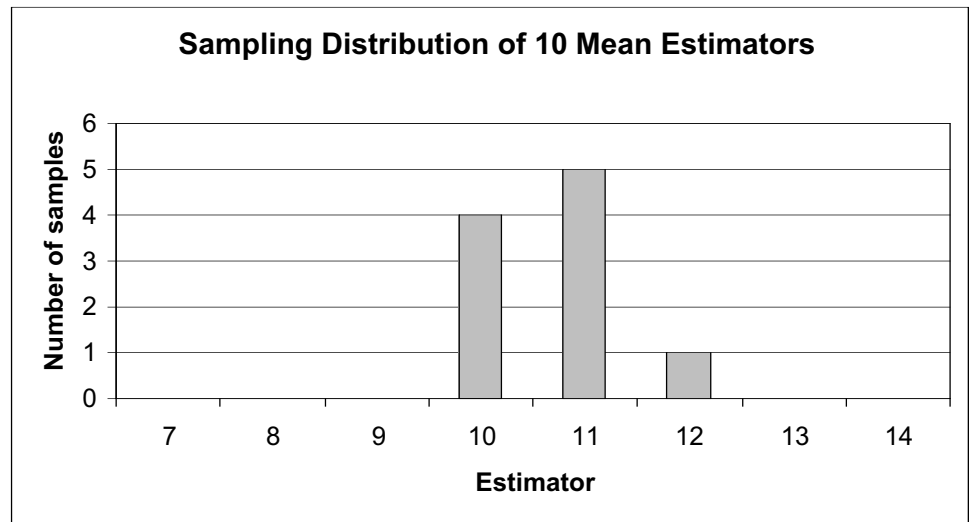


Figure 2

The central limit theorem, one of the most important statistical concepts you will encounter, states that in a finite population with a mean μ and variance σ^2 , the *sampling distribution of the means* approaches a *normal distribution* with a sampling mean μ and a sampling variance σ^2/N as N (N = number of individuals in the sample) increases. In Figure 2, 4 of our 10 samples had a mean of 10 mm, 5 samples had a mean of 11 mm, and 1 sample had a mean of 12 mm. The central limit theorem says that this sampling distribution will become more and more “normal” (a bell-shaped curve on a graph) as the sample size increases. It also says that the mean of the sampling distribution is an unbiased estimator of μ , and that the variance of the estimators is σ^2/N .

In this exercise, you will set up two populations that have the same mean, μ , of 50 mm. You will try to estimate this parameter through sampling. Both populations contain 500 individuals. The mean stem lengths of Population 1 follow a normal distribution. Population 2 has a somewhat funky, **bimodal distribution** in which individuals have stem lengths of either 0 or 100. We will obtain samples from each population, from which we will estimate the mean of each population.

The method by which we will sample is called the **bootstrap method**, a very common sampling method in statistics (Efron 1982). The bootstrap involves *repeated reestimation of a parameter (such as a mean) using random samples with replacement from the original data*. Because the sampling is with replacement, some items in the data set are selected two or more times and other are not selected at all. We will do a bootstrap analysis of the mean when sample sizes of 5, 10, 15, and 20 are drawn (with replacement) from each population. When the procedure is repeated a hundred or a thousand times, we get “pseudosamples” that behave similarly to the underlying distribution of the data. In turn, you can evaluate how biased your estimator is (whether your estimator gives a good estimate of μ or not), the confidence intervals of the estimator, and the bootstrap standard error of your estimator. All of this will become more clear as you work through the exercise.

As always, save your work frequently to disk.

INSTRUCTIONS

A. Set up the spread-sheet.

1. Open a new spread-sheet and set up column headings as shown in Figure 3.

2. In cells A7–A506, assign a number to each individual in the populations, starting with 1 in cell A7 and ending with 500 in cell A506.

3. Enter a population mean of 50 in cell C3.

4. Enter the standard deviation for Population 1 in cell C4.

ANNOTATION

	A	B	C
1	Central Limit Theorem Exercise		
2			
3	Population Mean => μ		50
4	Population Std => σ		10
5			
6	Individual	Pop 1	Pop 2

Figure 3

Enter 1 in cell A7.
Enter =A7+1 in cell A8.
Copy this formula down to cell A506 to designate the 500 individuals.

We will compare two populations of dandelions (actual statistical populations), each consisting of 500 individuals. Both populations, Population 1 and Population 2, have an actual mean stem length (μ) of 50 mm, which is designated in cell C3.

Population 1 will consist of 500 individuals that have a mean, μ , of 50 mm and a standard deviation of 10 mm. We'll assume that Population 1 is normally distributed. Thus, the raw data are distributed in a bell-shaped curve that is completely symmetrical and has tails that approach but never touch the x -axis. The shape and position of the normal curve is determined by μ and σ : μ sets the position of the curve while σ determines the spread of the curve. Figure 4 shows two normal curves. They have different means (μ) but have the same σ , thus they are similar in shape but are positioned in different locations along the x -axis.

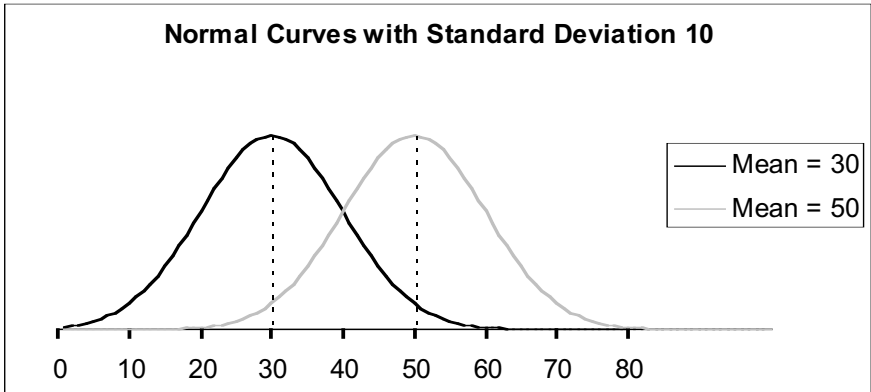


Figure 4

5. In cell B7, use the **NORMINV** function to obtain a stem length for Individual 1 in Population 1, whose mean and standard deviation are given in cell C3 and C4. Copy this formula down to obtain stem lengths for the remaining 499 individuals in Population 1.

6. Copy cells B7–B506 and paste their values in place of the formulae.

7. Enter 0 in cell C7, and fill this value down to cell C256. In cell C257, enter 100 and fill this value down to cell C506.

8. Label cell A507 as “Mean” and cell A509 as “Std” as shown in Figure 5.

A property of normal curves is that the total area under the curve is equal to 1. (This is true of all probability models or models of frequency distributions). Another property is that the most of the data fall in the middle of the curve around the mean. For normal distributions, approximately 68% of the observations will fall between the mean and ± 1 standard deviation. In our dandelion population, this means that 68% of the individuals in the population will have a stem length between 40 mm and 60 mm (which is the mean, 50 mm, ± 10 , which is 1 standard deviation). About 95% of the observations will fall between the mean and ± 2 standard deviations. Since our dandelion Population 1 is normally distributed, approximately 95% of the individuals will have stem lengths between 30 mm and 70 mm (2 standard deviations, or 20 mm, from the mean in either direction).

We used the formula **=NORMINV(RAND(),\$C\$3,\$C\$4)**. This formula allows us to draw a random probability from a normal distribution whose mean is 50 and standard deviation is 10, and convert it to a data point from the same distribution. In this way we can assign stem lengths to each individual in Population 1 and end up with a population that has (approximately) the desired mean and standard deviation.

Let’s look at the formula carefully. The **NORMINV** function consists of three parts, each separated by a comma. It has the form **NORMINV(probability, mean, standard_dev)**, where probability corresponds to the cumulative probability from the normal distribution, mean is the arithmetic mean of the distribution, and standard_dev is the standard deviation of the distribution. For example, the formula **=NORMINV(RAND(),\$C\$3,\$C\$4)** tells Excel to draw a random cumulative probability between 0 and 1 (the **RAND()** portion of the formula) from a normal distribution that has a mean given in cell C3 and a standard deviation given in cell C4. The formula returns the inverse of this probability; it changes the cumulative probability into an actual number from the distribution. Excel will return a value, which is the stem length of the individual.

Now we need to “fix” the stem lengths for Population 1 in cells B7–B506. (Otherwise, Excel will generate new stem lengths for Population 1 every time the spreadsheet recalculates its formulae).

Copy cells B7–B506.

Select cell B7.

Go to Edit | Paste Special | Paste Values. The **NORMINV** formula will be overwritten and the values will occupy the cells.

Population 2 also has a mean stem length, μ , of 50 mm. Stem lengths in this population are highly variable, where individuals either have a very long stem of 100 mm or no stem at all (0 mm).

	A	B	C
507	Mean =		
508	Std =		

Figure 5

9. Calculate the mean stem lengths and standard deviation for the two populations in cells B507–C508.

10. Save your work.

B. Construct a frequency distribution of the raw data.

1. Set up new column headings as shown in Figure 6. Enter values in cells F7–G16.

2. Use the **FREQUENCY** formula to generate the frequencies of the various stem lengths in Population 1. For example, in cell H7, count the number of individuals in Population 1 whose stem lengths are <10 mm. In cell H8, count the number of individuals whose stem lengths are within 10 and 19 mm, and so on.

We used the following formulae:

- Cell B507 =**AVERAGE(B7:B506)**
- Cell B508 =**STDEV(B7:B506)**
- Cell C507 =**AVERAGE(C7:C506)**
- Cell C508 =**STDEV(C7:C506)**


Note that both populations have approximately the same mean, but are very different in terms of how stem lengths are distributed in the population.

	F	G	H	I
3	Frequencies of Values in Populations			
4				
5				
6	"Bin"	Stem lengths	Pop 1	Pop 2
7	9	<10	0	250
8	19	<20	0	0
9	29	<30	6	0
10	39	<40	64	0
11	49	<50	153	0
12	59	<60	186	0
13	69	<70	79	0
14	79	<80	12	0
15	89	<90	0	0
16		<100	0	250

Figure 6

The most common way to depict a population's values is through a **frequency distribution**. A frequency distribution is a plot of the raw data, which we can generate using Excel's **FREQUENCY** function. This is an array formula (see pp xxx) and is a bit tricky, so proceed carefully.

The **FREQUENCY** function calculates how often values occur within a range of values, and then returns an array (or series) of numbers. For example, you will use it to count the number of stems that fall within 0 and 9 mm, 10 and 19 mm, and all of the other potential categories listed in Figure 6. Because **FREQUENCY** returns an array, it must be entered as an array formula. The function has the syntax **FREQUENCY(data_array, bins_array)**, where **data_array** is a set of values for which you want to count frequencies, and **bins_array** is a reference to intervals into which you want to group the values. You can think of a "bin" as a bucket in which specific numbers go. The bins may be very small (hold only a few numbers) or very large (hold a large set of numbers). In our example, we used bins that hold 10 numbers each. For example, a bin labeled 9 holds numbers 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The bin labeled 19 holds numbers 10, 11, 12, 13, 14, 15, 16, 17, 18, and 19. The bin labeled 89 holds numbers 80, 81, 82, 83, 84, 85, 86, 87, 88, and 89. Any data points greater than 89 go into a final "default" bin, which is not technically listed as a bin.

The **FREQUENCY** function works best when you use the f_x button and follow the cues for entering a formula. Since you will be entering this formula for an array of cells, the mechanics of entering this formula is different than the typical formula entry. Instead of selecting a single cell to enter a formula, you need to select a series of cells, then enter a formula, and then press <Control>+<Shift>+<Enter> (Windows) to simultaneously enter the formula for all of the cells you have selected. (Press the <Control>, <Shift>, and <Enter> keys in that order, making sure to hold the <Control> and <Shift> keys—or the  key if you use a Mac—down until the <Enter> key is pressed.

OK, let's try it. Select cells H7–H16 (where we are building the frequency distribution for Population 1) with your mouse, then press the f_x button and select the **FREQUENCY** function. Click on the button just to the right of the Data_array box (the button with the little arrow pointing up and left; see Figure 9 on p. 11); this will allow you to indicate the cells with the appropriate data by selecting them with your mouse. Select all of the individuals in Population 1 (i.e., cells B7–B506 of your data array) and click again on the button just to the right of the box again to return to the Frequency dialogue box. Then use the button next to the Bins_array box to select cells F7–F15 for your bins. Instead of clicking OK, press <Control>+<Shift>+<Enter>, and Excel will return your frequencies for Population 1. After you've obtained your results, examine the formulas in cells H7–H16. Your formula should look like this: **{=FREQUENCY(B7:B506,F7:F15)}**. This formula will be identical in all of the cells. The {} symbols indicate that the formula is an array formula.

3. Obtain the frequencies for Population 2 in I7–I16.

4. Construct a frequency histogram of the two populations. Select the data in G6–I16.

The data in the G column will form the x -axis, and the data in the H and I columns will make up the frequencies. Make sure you label your axes fully.

5. Save your work.

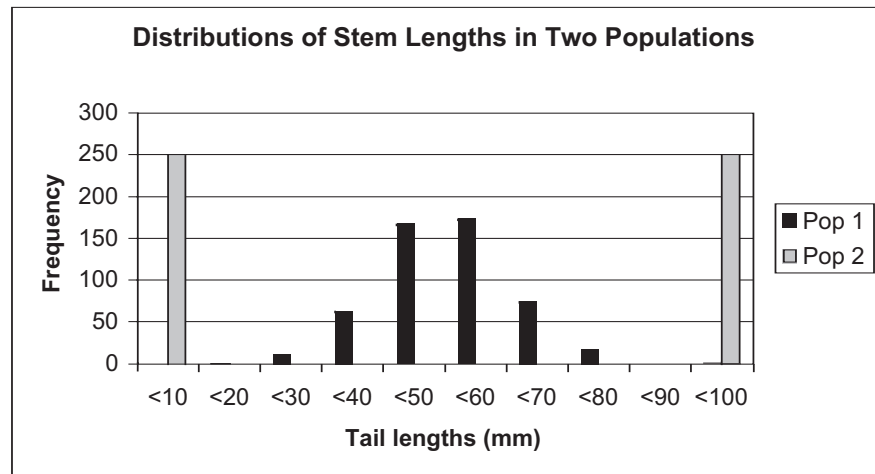


Figure 7

Based on Figure 7, it's easy to see that both populations have a mean around 50 mm, although their variances are quite different.

C. Obtain random samples from each population.

1. Set up new column headings as shown in Figure 8, but extend the series in row F to cell F40.

	F	G	H	I	J	K	L	M	N
19		$n = 5$		$n = 10$		$n = 15$		$n = 20$	
20	Individual #	Pop 1	Pop 2	Pop 1	Pop 2	Pop 1	Pop 2	Pop 1	Pop 2
21	1								
22	2								
23	3								
24	4								
25	5								

Figure 8

2. In cell F18, generate a random number between 1 and 500.

3. Enter a formula in cell G21 to return the stem length of a random individual in Population 1.

4. Copy cell G21 into cells I21, K21, and M21.

5. Copy the formula in G21 down to G25. Copy the formula in I21 down to I30. Copy the formula in K21 down to K35. Copy the formula in M21 down to M40.

6. Obtain samples from Population 2 and output stem lengths in the appropriate cells.

Remember that our goal is not to measure all 500 individuals in each population, but to sample from each population and estimate μ with a statistic. We will now randomly sample individuals from the population (with replacement), and estimate μ . We will do this for sample sizes of 5, 10, 15, and 20 individuals.

The random number will select which individuals from the population will be part of a sample. For example, if the random number is 324, then individual number 324 will be selected as part of the sample. Two formulae can be used to generate a random number between 1 and 500: **=RANDBETWEEN(1,500)** and **=ROUNDUP(RAND()*500,0)**

Press F9, the calculate key, several times to obtain new random numbers in cell F18.

Now we will draw a random sample from Population 1, and output the individual's stem length in cell G1. We'll use the **VLOOKUP** formula, combined with the **RANDBETWEEN** (or **ROUNDUP(RAND())** formula) above, to accomplish this task. The **VLOOKUP** formula searches for a value in the leftmost column of a table you specify (in this case, the table consists of cells A7–B506; the leftmost column is column A, which gives the individual's number). The function finds the individual's number, then returns a value associated with that individual from a different column in the table (in this case, the stem length associated with the randomly drawn individual).

Enter one of the following formulae (depending on whether or not you have the **RANDBETWEEN** function) in cell G21: **=VLOOKUP(RANDBETWEEN(1,500),A\$7:\$B\$506,2)** or **=VLOOKUP((ROUNDUP(RAND()*500,0),A\$7:\$B\$506,2)**. This formula tells Excel to generate a random number between 1 and 500 (the **RANDBETWEEN** or **ROUNDUP(RAND())** portion of the formula), find that number in the left-hand column in the table, and then return the value listed in the second column of the table.

At this point, for Population 1, you have drawn a random sample of 5 individuals (in cells G21–G25), a random sample of 10 individuals (in cells I21–I30), a random sample of 15 individuals (in cells K21–K35), and a random sample of 20 individuals (in cells M21–M40).

We used the formula **=VLOOKUP(RANDBETWEEN(1,500),A\$7:\$C\$506,3)**. Note that our VLOOKUP table now includes columns A through C, and returns the value associated with the third column of data (stem lengths from Population 2).

Your spreadsheet should now look like Figure 9 (the values in the cells will be different).

	F	G	H	I	J	K	L	M	N
19		n = 5		n = 10		n = 15		n = 20	
20	Individual #	Pop 1	Pop 2	Pop 1	Pop 2	Pop 1	Pop 2	Pop 1	Pop 2
21	1	49	0	37	0	51	100	34	100
22	2	46	0	54	0	62	0	69	100
23	3	43	0	32	0	69	100	58	100
24	4	52	0	51	0	49	0	62	100
25	5	46	100	52	100	62	100	28	0
26	6	4		8	0	48	100	45	0
27	7	2		6	100	58	0	54	100
28	8	4		9	0	33	100	56	0
29	9	5		8	0	31	0	32	100
30	10			62	100	54	0	39	100
31	11					46	100	42	100
32	12					63	0	62	0
33	13					46	100	44	0
34	14					54	0	58	100
35	15					44	0	41	100
36	16							45	100
37	17							59	0
38	18							45	0
39	19							69	100
40	20							50	0

Figure 9

7. Calculate the mean for each population and each sample size in cells G41–N41.

8. Save your work.

D. Set up the bootstrap.

1. Set up new column headings as shown, but extend the trials to 25 in cell F69.

Enter **=AVERAGE(G21:G40)** in cell G41. Copy this formula over to cell N41. Now you have an estimator of the mean for each population when various sample sizes (N) are taken.

The central limit theorem says that if we repeat this process many times and construct a graph of the frequency distribution of our *sampling means*—or estimates—the average of that sampling distribution will in fact be close to μ , the actual mean stem length of the population. So far, you’ve run one “trial.” To make a sampling distribution of the means, you’ll want to run several trials with a bootstrap analysis. We’ll do 25 trials in this exercise, which should be just enough to show you the general principles of the central limit theorem. (You can do more trials if you’d like.)

	F	G	H	I	J	K	L	M	N
43		n = 5		n = 10		n = 15		n = 20	
44		Pop 1	Pop 2	Pop 1	Pop 2	Pop 1	Pop 2	Pop 1	Pop 2
45	Trial 1								
46	Trial 2								
47	Trial 3								
48	Trial 4								
49	Trial 5								

Figure 10

2. Develop a bootstrap macro.

The following steps will create a bootstrap macro:

- Open Tools | Options | Calculation and set the calculation key to manual.
- Open Tools | Macro | Record New Macro. A dialog box will appear. Type in a name (bootstrap) and a shortcut key (<Control>+b).
- Press F9, the calculate key, to generate a new set of random samples from both populations.
- Select cells G41–N41, the estimators of μ for various sample sizes.
- Open Edit | Copy. Select cell G44.
- Open Edit | Find. A dialog box will appear. Leave the Find What box completely blank. Search by columns and look in values, then select Find Next and then Close. Your cursor should move down to cell G45 (the next blank cell in that column).

3. Save your work.

E. Construct a Sampling Distribution of the Means.

1. Set up column headings as shown in Figure 11.

2. Use the **FREQUENCY** function to count the frequency in which certain values (estimators) were obtained for Population 1 for various sample sizes.

3. Construct a sampling distribution of the *means* (Figure 12) by plotting the results from the previous step.

- Open Edit | Paste Special | Paste Values. Select OK.
- Open Macros | Stop Recording (or, if the Stop Recording menu is visible, press the Stop Recording button).
- Open Tools | Options | Calculation and return your calculation to automatic.

Your bootstrap macro is finished. When you press <Control>+b 24 more times, you will have resampled your population and computed new means for 25 different trials. This is the bootstrap analysis.

	E	F	G	H	I
72		Frequency of Estimated Mean			
73	"Bin"	<i>n</i> = 5	<i>n</i> = 10	<i>n</i> = 15	<i>n</i> = 20
74	40				
75	41				
76	42				
77	43				
78	44				
79	45				
80	46				
81	47				
82	48				
83	49				
84	50				
85	51				
86	52				
87	53				
88	54				
89	55				
90	56				
91	57				
92	58				
93	59				
94	60				
95	>60				

Figure 11

We used the following formulae:

- F74–F95 {=FREQUENCY(G45:G69,E74:E94)}
- G74–G95 {=FREQUENCY(I45:I69,E74:E94)}
- H74–H95 {=FREQUENCY(K45:K69,E74:E94)}
- I74–I95 {=FREQUENCY(M45:M69,E74:E94)}

For clarity, we have graphed only the cases $N = 5$ and $N = 20$. Your own graph will look different.

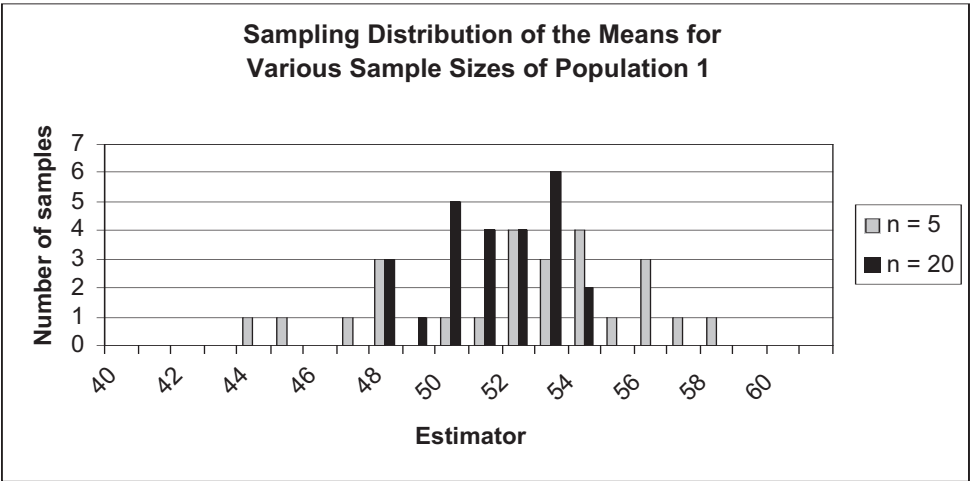


Figure 12

QUESTIONS

1. Examine your graph from Part E, Step 3. How does N , the sample size, affect the sampling distribution’s mean and variance?
2. Repeat Part E for Population 2. Set up column headings and bins as shown in Figure 13. Explain why different bins are necessary for this population. Population 2 has a very strong bimodal distribution. Does the sampling distribution at $N = 20$ also have a bimodal shape? How does the shape of the sampling distribution change as sample size changes?

	J	K	L	M	N
71			Population 2		
72		Frequency of Estimated Mean			
73	"Bin"	$n = 5$	$n = 10$	$n = 15$	$n = 20$
74	0				
75	10				
76	20				
77	30				
78	40				
79	50				
80	60				
81	70				
82	80				
83	90				
84	100				

Figure 13

3. Review the definition of the central limit theorem (given at the top of Page 67). How close was the average of your bootstrap analyses to μ ? How did sample size affect this? Did the two populations show similar results? Why or why not?

4. What is the relationship between the standard error of the sample means and the sample size? What is the relationship between the standard deviation of the raw data and the sample size? Calculate the standard deviation of samples in row 42, and calculate the standard deviation of your 25 trials in row 70. Plot your results for Population 1. Does the variance in the sampling distribution tell you anything about the variance in the raw data? If your sample size is 1 and you repeatedly estimate the mean, what will the variance of your sampling distribution be?

LITERATURE CITED

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Sokal, R. R., and F. J. Rohlf. 1995. *Biometry*. 3rd Edition. W. H. Freeman & Co., New York.