Microsoft

Sampling, Sampling Distributions, Parameter Estimation and Confidence Intervals

All work for this handout can be found in file “Module4Demos.xlsx”.

Populations and Samples

A **population** is a collection of all objects of interest. Some populations include:

- All voters registered for a US Presidential election.
- All Americans who have a CPA.
- All cows in India.
- All customers shopping at a department store on a chosen day.
- All computer chips produced this month at a semiconductor plant.
- All families in Houston, Texas.

Often, we are interested in estimating a numerical characteristic of a population, or a population **parameter**. Some examples of population parameters include:

- Proportion of voters preferring the Democratic candidate
- Average age of all CPA’s
- Average weight of all cows in India
- The standard deviation of the amount spent by a department store customer
- Fraction of all computer chips that are defective
- Median income of families in Houston, Texas

A complete enumeration of the population is a census. A **sample** is a part of the population that we observe to glean insights about the population. Samples are used for several reasons:

- If the population is large, it is impractical to take a census. For example, it would be impractical to sample every registered voter to learn how they are voting.
- Sampling involves examining a smaller set than a census, reducing measurement error.
- Sampling may involve destroying elements of the population. For example, testing a chip to see if the chip is defective may involve destroying the chip.

Estimates from sample data are called **statistics** and are used to estimate a population parameter. For example, we could estimate the:

- Median income of Houston families by taking a sample of 100 Houston families and using the median income in the sample to estimate the population’s median income.
- Proportion of defective chips by testing 100 chips and using that proportion to estimate the fraction of defective chips in the population. For example, if 5 of the 100 tested chips are defective we would estimate that 5% of the chips produced are defective.

- Average weight of all cows in India by weighing 100 cows and using the average weight of the cows in the sample to estimate the average weight of all Indian cows

Simple Random Sample (SRS)

Suppose a population has N individuals and we want to take a sample of size n. The sample is a **simple random sample** if each set of n individuals has the same chance of being chosen. For example, consider a random sample without replacement of two items from a population of size 5 (n = 2, N = 5.) Each of the possible ten samples shown below has the same chance of being chosen.

(1,2) (1,3) (1,4) (1,5) (2,3) (2,4) (2,5) (3,4) (3,5) (4,5).

Note that each population member has as 40% chance (2/5) of being chosen; in general, the chance of any population member being chosen is n/N.

It is easy to use Excel to generate a simple random sample. Using the players listed in the "SRS" worksheet in the **Sample.xlsx** spreadsheet, let's select a simple random sample of 10 from a list of NBA players. Simply enter the formula =RAND() next to each player's name in column F, and then "Copy Paste Special Values" in column G. Then, sort the columns in descending order based on the values in column G. Your simple random sample is the first 10 players listed. The values in column D show the list of 10 that was selected when we originally ran this analysis (see Figure 4-1). Note the differences between your sample and ours.

	D	E
6	Rand	**Player**
7	0.998152	**Chris_Douglas-Robe**
8	0.994134	**Nick_Calathes**
9	0.987062	**Paul_Pierce**
10	0.986242	**Patrick_Beverley**
11	0.985116	**Roger_Mason_Jr.**
12	0.973976	**Carlos_Delfino**
13	0.973319	**Kelly_Olynyk**
14	0.970552	**David_Lee**
15	0.967327	**Tony_Mitchell**
16	0.961087	**Blake_Griffin**

Figure 4-1. SRS of 10 NBA players

Other Types of Sampling

We could also do some more sophisticated types of sampling although some of these strategies will lead to some inherent biases. Let's look at some other options, starting with stratified random sampling.

With stratified random sampling, the population is divided into groups, called strata, based on some characteristic. Then, within each group, a sample, usually random, is selected. How many are selected from each strata depends on the purpose for creating the strata initially. As an example, suppose we want to make sure that the percentage of genders in our study is equal to that of the population. We plan to sample 100 people from the population and the percent of women is 45%. In this case, we'd randomly select 45 women and 55 men to participate in our study. In most cases, stratification is done

to ensure that sample percentages match population percentages on some key characteristic. However, imagine that you are studying something that tends to be more of an issue for some part of your population than for another, more so for women rather than men, such as the glass ceiling effect, for example. In this case, you might use stratified random sampling to over select women. You might select 60, rather than 45, women and 40 men.

Cluster sampling involves dividing up the population into clusters and then selecting clusters to be part of the sample. Every cluster should represent the population on a small scale and be as heterogenous as possible. Every population element must belong to one and only one cluster. If the researcher decides to include all individuals from a cluster in the sample, this called one-stage clustering. If the researcher randomly selects from the clusters, this is called multi-stage clustering. How is this different from stratified sampling? In stratified, some members from each group or strata are selected, but in cluster, all or part of some clusters are used but not all clusters are included. If the clusters are heterogeneous and representative, this sampling strategy tends to work well.

In systematic random sampling, the researcher first randomly picks the first item or subject from the population. Then, the researcher will select each n'th subject from the list. The results are usually representative of the population unless certain characteristics of the population are repeated for every n'th individual, which is highly unlikely.

Convenience sampling is sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher. For example, professors who have their students participate in their research projects are using convenience sampling. Clearly, there are a lot of problems with this type of sampling because it is very unlikely to represent the population… unless, of course, your population is those that are close to you.

Problems with Sampling

A sample that is not random can cause serious errors. Suppose every 10^{th} chip produced is defective. Then the population has 10% defectives. If we sample every 10^{th} item, however, we would estimate that 100% of the chips would be defective. As the following examples show, many problems can arise when conducting a sampling study.

Bias is a systematic error that is introduced into your study that can prejudice your results in some way. You should be aware of the most common biases as you consider your sample. In most cases, bias is unintentional and happens because the sampling procedure wasn't well thought out. A classic example of this is a company who wants information about one of their products in a particular area and decides to do a phone survey. Even if they do a random sample, there's still a problem… they are missing input from customers who don't have a phone. There may be something systematically different about people who use their product and own a phone and those who use it but do not have a phone.

This is an example of selection bias. **Selection bias** occurs when each element in the population does not have same chance of being chosen in sample. Other examples include the 1972 Presidential election; although Nixon won the election in a landslide, liberal movie critic, Pauline Kael, said she did not know anyone who voted for Nixon, a very conservative candidate. A few other examples…Bernie Sanders outperformed polls in the 2016 Democratic Michigan primary because pollsters assumed the electorate would be like that of previous primaries where 50% of the electorate is 50 years or older. In reality, there were many more young voters than expected. Similarly, in the 2016 election, many pollsters

assumed the electorate would look like 2012 electorate, but in 2016, more rural voters and fewer African Americans voted compared to 2012. Complicating this, today, only 10% of all people contacted by pollsters respond, making it difficult to get a random sample. We consider poll results with some skepticism.

Nonresponse bias occurs when respondents differ in meaningful ways from non-respondents. Similarly, voluntary response bias occurs when sample members are self-selected volunteers; the resulting sample tends to over-represent individuals who have strong opinions.
Publication bias occurs because studies with positive results are more likely to be published than negative or null results. This is a big problem in many fields because the failure to publish these types of studies leads us to draw conclusions without all the information. For example, in studies of antidepressants 94% of studies with positive results were published but only 14% of studies with negative results were published.

Survivorship Bias occurs when a meaningful part of a population is not considered in your sample. For example, in World War 2, engineers noticed that after a mission, bullet holes tended to be clustered in the wing, rear gunner, and body of returned planes. They started reinforcing those areas without giving thought to where the bullets were doing the most damage to the planes that did NOT return. Today, we this in finance when mutual fund companies drop poorly performing mutual funds, which overinflates their past returns, or when you look at successful people, identify an interesting characteristic, such as dropping out of school, and assuming that all successful people have dropped out of school. Essentially, you are ignoring failures—all those people who dropped out of school and went nowhere—when interpreting the results of your study.

Response bias refers to the bias that results from problems in the measurement process, for example, when you ask leading questions, when people tell you want they think you want to hear, or because they want to present themselves in a favorable way (this is known as social desirability bias). This happens when a small fraction of those sampled respond, and the respondents may not be representative of the population. For example, the TV news show Nightline asked people to call in about whether the US should leave the UN. 67% of those calling wanted the US to leave the UN. A correctly designed sample study estimated only 28% of people wanted the US to leave the UN.

Point Estimates of Population Parameters and Sampling Distributions

Suppose we want to estimate the unknown mean of a random variable. Suppose this unknown mean = μ. We use the sample statistic $\bar{x}$, the sample mean as an estimate of μ. If we take a sample of n independent observations x1, x2, …, x_n from a population, then

$$\bar{x} = \frac{x1+x2+\cdots xn}{n}.$$

The sample mean is an **unbiased** estimate of μ. This means that if we take many samples and average our $\bar{x}$ values, we should get μ. To show this, we use the rule that the expected value of a sum of random variables is the sum of expected values shows:

$E(\bar{x}) = E(x_{n1}/n) + E(x2/n) + \ldots E(x_n/n) = n*(\mu/n) = \mu.$

It can be shown that among all unbiased estimators of the population mean, the sample mean has the smallest variance. For that reason, we use the sample mean as a point estimate for the population mean.

Because the sample mean $\bar{x}$ is a random variable, it has a variance. Suppose the population from which we are sampling has a variance of σ^2. To find the variance of $\bar{x}$, we use the logic of expected values like we did for the mean. As a result,

$\text{var}(\bar{x}) = \text{var}(x1/n^2) + \text{var}(x2/n^2) + \ldots \text{var}(xn/n^2) = n\sigma^2/n^2 = \sigma^2/n$.

The standard deviation of $\bar{x}$ is $(\sigma/\sqrt{n})$ and is referred to as the standard error of $\bar{x}$.

As an illustration, suppose our population is the roll of a die, and we take a sample of size 2. All possible outcomes, each having probability 1/6, are shown in Figure 4-2 and "Two Dice" worksheet of **Sample.xlsx**. Recall that when one die is tossed, the expected mean is 3.5 and variance is 2.92. From Figure 4-2, we find $E(\bar{x}) = \mu = 3.5$ and $\text{var}(\bar{x}) = 2.92/2$ (2 being the number of die that are tossed) = 1.46.

	B	C	D	E	F	G	H
3							Meanxbar
4	Probability	Die 1	Die 2	Xbar	Squared dev		3.5
5	0.027777778	1	1	1	6.25		Variance xbar
6	0.027777778	1	2	1.5	4		1.458333333
7	0.027777778	1	3	2	2.25		
8	0.027777778	1	4	2.5	1		
9	0.027777778	1	5	3	0.25		
10	0.027777778	1	6	3.5	0		
11	0.027777778	2	1	1.5	4		
12	0.027777778	2	2	2	2.25		
13	0.027777778	2	3	2.5	1		
14	0.027777778	2	4	3	0.25		
15	0.027777778	2	5	3.5	0		
16	0.027777778	2	6	4	0.25		
17	0.027777778	3	1	2	2.25		
18	0.027777778	3	2	2.5	1		
19	0.027777778	3	3	3	0.25		
20	0.027777778	3	4	3.5	0		
21	0.027777778	3	5	4	0.25		
22	0.027777778	3	6	4.5	1		
23	0.027777778	4	1	2.5	1		
24	0.027777778	4	2	3	0.25		
25	0.027777778	4	3	3.5	0		
26	0.027777778	4	4	4	0.25		
27	0.027777778	4	5	4.5	1		
28	0.027777778	4	6	5	2.25		
29	0.027777778	5	1	3	0.25		
30	0.027777778	5	2	3.5	0		
31	0.027777778	5	3	4	0.25		
32	0.027777778	5	4	4.5	1		
33	0.027777778	5	5	5	2.25		
34	0.027777778	5	6	5.5	4		
35	0.027777778	6	1	3.5	0		
36	0.027777778	6	2	4	0.25		
37	0.027777778	6	3	4.5	1		
38	0.027777778	6	4	5	2.25		
39	0.027777778	6	5	5.5	4		
40	0.027777778	6	6	6	6.25		

Figure 4-2. Sample Means When Tossing Two Dice

Suppose we want to estimate an unknown population proportion, *p*. For example, let *p* be the fraction of registered voters in Seattle, Washington, who are Independents. To estimate *p*, we might ask n randomly chosen Seattle voters if they are independents and estimate p by

$\hat{p} = \frac{number\ voters\ in\ sample\ who\ are\ independent}{n}$.

In short, we estimate the population parameter, *p,* by the fraction of "successes" in the sample. For example, if 100 of 400 samples voters say they are independents, we would estimate p by $\hat{p}$ = 100/400 = 0.25. Of course, $\hat{p}$ is a random variable.

We see that $\hat{p}$ is an unbiased estimate of the population proportion, *p*. It can be shown that among all unbiased estimates of the population mean, $\hat{p}$ has the smallest variance. Therefore, we use $\hat{p}$ as a point estimate for *p*.

It can easily be shown that E($\hat{p}$) = p and standard deviation $\hat{p} = \sqrt{\frac{p(1-p)}{n}}$. Because we often do not know *p*, we usually assume that the standard deviation ($\hat{p}$) = $\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$. The standard deviation of $\hat{p}$ is called the standard error of $\hat{p}$.

Our estimates of μ by $\bar{x}$ and *p* by $\hat{p}$ are **point estimates** of population of parameters. Of course, if we take a different sample, our point estimates will change. Therefore, it is important to measure the precision or accuracy of our point estimates of population parameters. We next turn our attention to the study of **confidence intervals** that measure the precision, or accuracy, of our point estimates of population parameters.

The Standard Normal

A normal random variable with mean 0 and standard deviation 1 is called a standard normal and is often referred to as **Z** (remember Z scores have a mean of 0 and standard deviation of 1). Confidence intervals require finding the percentiles of the standard normal.

In the "Standard Normal" worksheet, find the 2.5%ile (often referred to as $z_{.025}$) and 97.5%ile (often referred to as $z_{.975}$) using the formulas in Figure 4-3. The NORM.S.INV function returns percentiles for a standard normal. From Figure 4-3, we find that there is a 95% of the values in a standard normal are between -1.96 and 1.96.

	E	F	G	H	I
2					
3					
4	**2.5 %ile**	**-1.95996**	**-1.95996**	**=NORM.INV(0.025,0,1)**	**=NORM.S.INV(0.025)**
5	**97.5%ile**	**1.959964**	**1.959964**	**=NORM.INV(0.975,0,1)**	**=NORM.S.INV(0.975)**

Figure 4-3. Finding Percentiles for Standard Normal

95% Confidence Interval for Population Mean

Suppose we have a random variable, **X**, with an unknown mean, μ and a known standard deviation, σ. To estimate μ, we take a random sample X1, X2, …, X_n from **X**. If we assume n >= 0, then the Central Limit Theorem tells us that the sample mean $\bar{x}$ will be approximately normal. After standardizing $\bar{x}$, we find that:

Prob($z_{.025}$ <= $(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}})$ $\leq z_{.975}$) = .95.

Rearranging this a bit, we find that:

Prob($\bar{x}$ + $z_{.025}\sigma/\sqrt{n}$ <= μ <= $\bar{x}$ + $z_{.975}\sigma/\sqrt{n}$) = 0.95.

This equation tells us that if we take 100 samples of size n from a random variable **X** and construct for each sample the interval, the

lower limit = $\bar{x}$ - $z_{.025}\sigma/\sqrt{n}$ or $\bar{x}$-1.96$\sigma/\sqrt{n}$ and the

upper limit = $\bar{x}$ + $z_{.975}\sigma/\sqrt{n}$ or $\bar{x}$+1.96$\sigma/\sqrt{n}$.

Confidence intervals provide the range where approximately 95% of the 100 intervals will contain the true value of μ.

Example of 95% Confidence Interval

You are told the standard deviation of invoice values is $500. A sample of 100 invoices taken from a large sample of invoices has a sample mean value of $4500. You are 95% sure the mean size of an invoice is within what range?

As shown in the "CI for Mu" worksheet and Figure 4-4, we find that our 95% confidence interval for μ is [$4402, $4598].

The half width of a 95% confidence interval is the **margin of error** in the estimate, or $z_{.025}\sigma/\sqrt{n}$. Thus, our margin of error in the estimate of the population mean is $98.

	C	D	E
1			
2		**samplemean**	**4500**
3		**popsigma**	**500**
4		**samplesize**	**100**
5		**z.025**	**-1.95996**
6		**z.975**	**1.959964**
7			
8		**Lower Limit**	**4402.002**
9		**Upper Limit**	**4597.998**

Figure 4-4. 95% Confidence Interval for Mean Invoice Size

Demonstration of Meaning of 95% Confidence Interval

In the "IQ CI" worksheet, we take 100 samples of 36 IQs and compute 100 95% confidence intervals. IQ is a normal random variable with mean 100 and standard deviation 15. If you hit F9, the sample values and the confidence intervals change. You will see that invariably around 95 of the 100 intervals contain the true value (100) of the population mean.

95% Confidence Interval for Population Proportion

Consider a large population in which an unknown fraction or proportion, *p*, of the population has a given attribute. We now calculate a 95% confidence interval formula for a population proportion, *p*. If $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$, then the central limit applies, and $\hat{p}$ will follow a normal random variable. Standardizing $\hat{p}$ we find that:

$$\text{Prob}(-1.96 <= \frac{\hat{p}-p}{\sqrt{\frac{(\hat{p}*(1-\hat{p}))}{n}}} <= 1.96) = 0.95.$$

Rearranging this a bit, we see that:

$$\text{Prob}(\hat{p} - 1.96\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}} <= p <= \hat{p} + 1.96\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}) = 0.95.$$

This equation tells us that if we take 100 samples of size n from a large population in which an unknown fraction, *p*, of the population has an attribute, and construct for each sample the interval the

$$\text{lower limit} = \hat{p} - 1.96\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}} \text{ and the}$$

$$\text{upper limit} = \hat{p} + 1.96\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}.$$

Approximately 95 of these intervals will contain the true proportion, *p*.

Example

Assume that every US voter will either vote for the Democratic or Republican candidate in a Presidential election. Suppose in a random sample of 1500 registered voters, 800 prefer the Democratic candidate. Construct a 95% confidence interval for the proportion of registered voters that prefer the Democratic candidate. In the "Voters" worksheet and Figure 4-5, we see that the confidence interval is [.5081, .5586], meaning that the actual percentage of registered voters preferring the Democratic candidate is between 50.8% and 55.9%. The half width of the confidence interval is 2.5%, or the margin of error. This is consistent with the usual margin of error of 3% reported in Presidential election polls. It is truly remarkable that a sample of 1500 voters (out of 219 million registered voters) can, in theory, deliver this level of accuracy. Columbia statistician, Andrew Gelman, recently examined a large number of political polls and found that, for complex reasons, the actual margin of error turned out to be double the theoretical 3%.

	D	E	F	G	H	I
2						
3	**n**	**1500**				
4	**phat**	**0.533333**				
5	**Std Error phat**	**0.012881**	**=SQRT((phat)*(1-phat)/n)**			
6	**Lower Limit**	**0.508086**	**=SQRT((phat)*(1-phat)/n)**			
7	**Upper Limit**	**0.558581**	**=SQRT((phat)*(1-phat)/n)**			
8	**Margin of erro**	**0.025247**	**=SQRT((phat)*(1-phat)/n)**			
9						

Figure 4-5. 95% Confidence Interval for Population Proportion

Blyth's Method for Confidence Intervals on a Proportion

If all trials result in success or failure, our proportion confidence interval will result in a confidence with 0 width. This is unreasonable, so Blyth developed a confidence interval formula to be used when all trials result in success or failure. The worksheet Blyth in the file Module4demos.xlsx shows how to use Blyth's Method. Suppose my son has driver to work 500 times without an accident. Let's find a 95% confidence interval for the chance he will have, or will not have an accident. Simply enter the number of trials in cell C3 and alpha of .05 for a 95% confidence interval (alpha of .01 for a 99% confidence interval) in C4. In cells C8 and D8 we find that we are 95% sure the chance of an accident is between 0 and 0.005974 and from cells C11 and D11 we find that we are 95% sure the chance of no accident is between 0,99492645 and 1.

	A	B	C	D	E	F
1	**Blyth Confidence Interval**					
2						
3		**n**	**500**			
4		**alpha**	**0.05**			
5						
6						
7		**Successes**	**Lower**	**Upper**		
8		**0**	**0**	**0.005974**	**0**	**=1-alpha^(1/n)**
9		**1**	**5.0634E-05**	**0.007351**	**=1-(1-0.5*alpha)^(1/n)**	**=1-(0.5*alpha)^(1/n)**
10		**499**	**0.99264939**	**0.999949**	**=(0.5*alpha)^(1/n)**	**=(1-0.5*alpha)^(1/n)**
11		**500**	**0.99402645**	**1**	**=(alpha)^(1/n)**	**1**

Sample Size Determination

Suppose we want to be 95% sure that our estimate $\bar{x}$ of the population mean μ is accurate within and error amount E. How large a sample size n is needed? Simply set the half-width of the 95% confidence interval for μ equal to E: $E = 1.96\sigma/\sqrt{n}$. Solving for n we find $n = (1.96\sigma/E)^2$. Note that if the sample size n exceeds 10% of the population size, then a smaller sample size is needed.

Example

Suppose we know the standard deviation of a large population of uncashed checks is $100. If we want to be 95% sure we can estimate the average size of an uncashed check is within $20, how large of a sample is needed? From the "Sample Size" worksheet and Figure 4-6, we find a sample size of 96 is needed.

	C	D	E	F
2		**ESTIMATING POPULATION MEAN**		
3				
4		**SIGMA**	**$100.00**	
5		**ERROR**	**$20.00**	
6		**SAMPLE SIZE**	**96.04**	**=(1.96*SIGMA/ERROR)^2**
7				

Figure 4-6. Sample Size Determination for Estimating Population Mean

Suppose we want to estimate a population proportion, *p*, and be 95% sure our estimate of p ($\hat{p}$) is accurate within a given amount E. How large does the sample size n need to be? From what we saw above, we know that sample size is based on the margin of error: E = $1.96\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$.

The problem is that in advance of taking the sample, we do not know $\hat{p}$. It is easy to show, however, that the maximum value of $\hat{p}$ *(1- $\hat{p}$) is 0.25 and occurs when $\hat{p}$ = 0.5. Therefore, a conservative bound on the needed sample size can be obtained by setting $\hat{p}$ = 0.5. Solving for n we obtain a sample size

$n = 1.96^2/4E^2$.

To illustrate this, suppose we want to estimate the fraction of registered voters preferring the Republican candidate in a Texas election. We would like our estimate to have a 95% chance of being accurate within 3%. How large of a sample is needed? As shown in the "Sample Size" worksheet and Figure 4.7, a sample size of 1067 is needed.

	D	E	F
7			
8	**Estimating Population Proportion**		
9	**Error**	**0.03**	
10	**Sample Size**	**1067.111**	**=(1.96*SIGMA/ERROR)^2**
11			

Figure 4-7. Sample Size Determination for Estimating a Proportion

What if Sample Size is large Fraction of Population?

If sample size n is a large percentage, such as greater than 10%, of population size, N, we have more confidence in our estimate of the population mean, and we can bee 95% sure that μ is between

$\bar{x}$ - $\frac{1.96*\sigma*FC}{\sqrt{n}}$ and $\bar{x}$ + $\frac{1.96*\sigma*FC}{\sqrt{n}}$.

Here FC is the **Finite Correction Factor** = $\sqrt{\frac{N-n}{N-1}}$. This formula is known as the **Finite Correction Confidence Interval for the Population Mean.**

Note that if the sample size n is equal to the total population size, N, then FC and the width of the confidence interval are 0 because we know population mean exactly. Similarly, if the sample size is more than 10% of the population, then we are 95% sure the population proportion, *p*, is between a

lower limit = $\hat{p}$ - 1.96∗ FC ∗ $\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$ and an

upper limit = $\hat{p}$ + 1.96∗ FC ∗ $\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$.

A More General Sample Size Formula

The finite correction factor also impacts sample size determination. Suppose N_0 is the original sample size as determined above and N is the population size. If we sample without replacement, then a more accurate sample size formula would be:

n = $\frac{N_0*N}{N_0+N-1}$.

Examples

Suppose we want to estimate the average salary of Fortune 500 CEOs. Assume the standard deviation of these salaries is known to be $5 million. If we sample 100 CEOs and find an average salary of $40 million, then as shown in the "Finite Correction" worksheet and Figure 4-8, we are 95% sure that the actual mean salary of Fortune 500 CEOs is between $39.12 and $40.88. Note that without the Finite Correction Factor the 95% confidence interval would be $39.02 to $40.88, which is slightly wider than the actual 95% CI.

	F	G	H
1			
2	**samplesize**	**100**	
3	**popsize**	**500**	
4	**sigma**	**5**	
5	**xbar**	**40**	
6			
7			
8	**FC**	**0.895323**	**=SQRT((popsize-samplesize)/(popsize-1))**
9			
10	**lowerlimit**	**$39.12**	**=xbar-1.96*FC*sigma/SQRT(samplesize)**
11	**upperlimit**	**$40.88**	**=xbar+1.96*FC*sigma/SQRT(samplesize)**
12			
13	**WITHOUT FC FACTOR**		
14	**lower**	**39.02**	**=xbar-1.96*sigma/SQRT(samplesize)**
15	**upper**	**40.98**	**=xbar+1.96*sigma/SQRT(samplesize)**

Figure 4-8. Finite Correction Factor Confidence Interval for Population Mean

Now, suppose we want to estimate the mean salary of Fortune 500 CEOs and be 95% sure our estimate is accurate within $1 million. How large of a sample is needed? In the "FC Sample Size" worksheet and Figure 4-9, we find that a sample size of 81 is needed. Without incorporating the Finite Correction Factor, we find that a larger sample size (96) would be needed.

	F	G	H	I
1	**Error**	**1**		
2	**N**	**500**		
3	**sigma**	**5**		
4	**samplesizenoFC**	**96.04**	**=(1.96*sigma/Error)^2**	
5	**samplesizeFC**	**80.70046**	**=samplesizenoFC*N/(samplesizenoFC+N-1)**	

Figure 4-9. Sample Size with Finite Correction Factor