CHAPTER 17

INTRODUCTION TO CONFIDENCE INTERVALS AND HYPOTHESIS TESTING

Chapter 16 focused on the reality that repeated sampling will produce "statistics" (such as the sample mean) that will deviate around the true population parameter. This creates a problem in that, typically, we would like to assume that the sample statistic was either the same value as the parameter or was very close to it. Thus, adopting the strategy of using the statistic as a good estimate of the parameter is an easy course of action to take. However, we saw - depending on sample size and the amount of variability in the population - that the sample mean could be "off" considerably from the parameter. Therefore, always adopting the strategy that the sample statistic would be a good estimate of the parameter would not always lead to a "good" inference as to what the parameter value is most likely to be. Therefore, whenever an inference is being made about what the parameter is, sampling error must be factored into that process. The basic rule is: if sampling error appears to be small, then the statistic is likely to be close to the parameter value but, if sampling error appears to be relatively large, then we must be very cautious when stating that the statistic is a good estimate of the population parameter. When sampling error is large, our assumption that the statistic is a good estimate of the parameter could be quite inaccurate.

In inference, there are usually two different (but highly interdependent) concerns as to how we may use sample information to "generalize" to the larger population. First, one may simply be interested in the answer to the question:"what is the parameter"? Perhaps you have developed a new testing instrument and, since you do not have any data yet, would like some approximate idea of what the mean might be if this new instrument were administered to all the people in some specified population. The sample data, using the sample mean, can give you some idea assuming of course that the sample is representative of the population.

Second, another possibility would be to formulate some hypothesis that says: "I think the parameter is ____". For example, you might hypothesize that the population mean for this new test (based on some logical analysis you make) should be about 30 out of 40 questions or about 75%. By taking some sample data, administering the test, and looking at the percentage scores in the sample, you can get some "feel" for whether your hypothesis of 75% seems to be reasonable or not. If, for example, the sample mean was close to 75%, then you probably would feel that your hypothesis was supported and therefore would want to **RETAIN** your hypothesis. But, if the sample mean was quite different than 75%, then perhaps you would feel that your hypothesis was not supported and therefore you would be more tempted to **REJECT** your hypothesis. Whether you are asking a question about the parameter or testing some hypothesis about the population parameter, sampling error must be taken into account. The present Chapter continues our use of sample means and the population mean to outline the process of asking what the

parameter is or testing some hypothesis about the population mean. Again, the reason why the "mean" is used is simply because the procedures for it tend to be more straightforward and therefore the "logic" behind these types of inferences should be easier to understand.

Confidence Intervals for the Population Mean

Assume for a moment that we are interested in how many "watts" of light output you get from "60 watt" light bulbs in the local light bulb manufacturing plant. For purposes of our illustration, further assume that the "true mean" is 60 watts and the standard deviation of the wattage output is about 2 watts. Regardless of how good the manufacturing process is, each and every light bulb will not produce exactly 60 watts; some may give off a little lower amount and some may give off a little higher amount. To show the type of sample mean variation there could be in such a situation, I did a Minitab simulation generating 200 samples of 25 bulbs each, each sample taken at random. One of those random samples is shown below.

DATA FROM ONE RANDOM SAMPLE

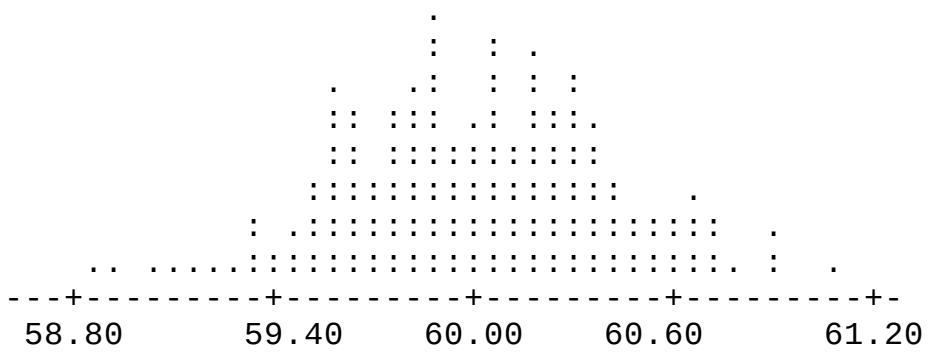
61	61	61	61	58	58	57	64	62	63	62	57	62
61	61	61	58	60	63	68	60	59	55	60	60	

A frequency distribution of that one set of sample data looks as follows.

Watt	Count
55	1
57	2
58	3
59	1
60	4
61	7
62	3
63	2
64	1
68	1
n =	25

Note that one value, 68, is a rather rare event if the population mean is about 60 and the standard deviation is 2. In a normal distribution, values are not very likely to be much more than 3 units of 2 around the mean or about 6 watts above and below. But, such is life; things like a 68 can happen.

Now, in this simulation, there were actually 200 different samples where n = 25 in each sample. Thus, I took the means from all 200 samples and made a dotplot.



Means from 200 Samples of Wattages

As usual, this sampling distribution of wattage means is similar to the normal distribution where the center point is near 60, or the population parameter. From the descriptive statistics shown below, notice that while the average of the means is close to 60, some of the means were as low as approximately 59 and some of the means were about as high as 61. Thus, while the "mean of the means" is near the parameter, sampling error will produce some variations around the parameter due to random sampling fluctuation. However, note that the means in this case do not vary too far from the value of 60. Also note the standard deviation of the distribution of means, which is .403, represents the extent of the sampling error. Recall that the standard deviation of the sampling distribution of means is called the standard error of the mean, and the larger this value is, the more sampling error is present in a typical sample drawn from the population. In this case, since the sampling distribution is approximately normal, the sample means would vary approximately 3 standard deviations of about .4 (about 1.2 watts in either direction of the parameter) around the center.

	# SAMPS	MEAN	STDEV	MIN	MAX
MeanWatt	200	60.007	0.403	58.852	59.733

The standard deviation of the means, .403, is our error factor or standard error of the mean. Thus, our estimate of the amount of sampling error in the means, if we took many different samples at random of n = 25, is a value of about .403 which is our standard error of the mean. The simulation above was simply another attempt to illustrate the idea that random samples of a certain size will provide you information about the sampling error that would be present if many, many random samples had in fact been selected.

Now we will use this standard error of the mean, or the error factor, to build what is called a confidence interval. **A CONFIDENCE INTERVAL IS AN INTERVAL AROUND THE STATISTIC WHERE THERE IS A CERTAIN CHANCE OR PROBABILITY OF INCLUDING OR CAPTURING THE PARAMETER**. In this case, the confidence interval

will be built around the sample mean in hopes of capturing the population mean. Recall that our first inferential application was trying to answer the question: what is the parameter? Since confidence intervals represent an interval where the population parameter (mean in this case) is likely to fall (within certain probability limits), then the confidence interval should help tell us what the value of the parameter is likely to be (within limits of chance of course).

What I have first done below is to build what we call 68% confidence intervals where about 68% of the intervals should capture the true population mean. Remember in a normal distribution, which is like our sampling distributions, about 68% of the area will be from the middle out one standard deviation on either side. So, to each sample mean, I will add and subtract one unit of error, the standard error of the mean. For each sample therefore, there will be an interval around the sample mean. Then we can see how many of these intervals actually include the population mean. For each of the 200 samples, there will be a sample mean. Thus, when adding and subtracting one unit of error from each of those sample means, we will have 200 different confidence intervals. How many will include the value of 60? For illustration purposes, I have listed out the first 15. The Yes1No0 listing merely indicates whether or not the interval contained 60.

Samp#	LowptCI	Popmean	UpptCI	Yes1No0
1	59.5121	60	60.3190	1
2	59.4945	60	60.3014	1
3	59.2899	60	60.0968	1
4	59.8644	60	60.6712	1
5	59.1323	60	59.9391	0
6	59.6297	60	60.4366	1
7	60.4970	60	61.3039	0
8	59.2955	60	60.1024	1
9	58.9564	60	59.7633	0
10	59.5351	60	60.3419	1
11	59.4669	60	60.2737	1
12	59.8362	60	60.6431	1
13	59.2403	60	60.0471	1
14	59.8124	60	60.6192	1
15	59.4768	60	60.2837	1

To find out how many of the intervals captured the parameter value of 60, we could count the number of 1's in the last column and, in this simulation, that value turns out to be 140. Out of 200 intervals, 140/200 = 70%. Thus, about 70 percent of the confidence intervals did in fact include the parameter value of 60.

What if you wanted to be more confident? Going one error unit on either side of the mean gives you about 68% accuracy (70% in this specific simulation). In a normal

distribution, you need to go out about 2 standard deviations to include about 95% of the area. So, if you wanted to have 95% confidence, we should find that going 2 units of error on either side of the mean should capture the parameter about 95% of the time. Will it? In the example above, I added .403 (or 1 unit of error) and subtracted .403 to each of the 200 sample means. To be approximately 95% confident, I will add and subtract 2 units of error or about .806 to each of the 200 sample means. Again, to see if the interval captures the population mean of 60, I print out the first 15 below. Note that if the interval contained 60 before, it would still have to capture 60 since the interval will be twice as wide. The question is: what if the interval before did not contain 60? Will being twice as wide now allow it to capture 60?

Samp#	LowptCI	Popmean	UpptCI	Yes1No0
1	59.1087	60	60.7224	1
2	59.0911	60	60.7048	1
3	58.8864	60	60.5002	1
4	59.4609	60	61.0747	1
5	58.7288	60	60.3426	1
6	59.2263	60	60.8400	1
7	60.0936	60	61.7073	0
8	58.8921	60	60.5058	1
9	58.5530	60	60.1667	1
10	59.1316	60	60.7454	1
11	59.0634	60	60.6772	1
12	59.4328	60	61.0466	1
13	58.8368	60	60.4506	1
14	59.4089	60	61.0227	1
15	59.0734	60	60.6871	1

In this case, the intervals are twice as wide as the 68% ones above and therefore should capture the value of 60 more often. In fact, they did. If we see how many 1's there are, which is how many of the 95% confidence intervals captured the parameter value of 60, we find that 191 intervals capture 60 and that is 191/200 = 96%. The process of going about 2 units of error around the sample mean will capture the parameter about 95% of the time. Since about 95% of the CI's captured the true mean, that indicates that we are about 95% sure that any one particular confidence interval would in fact capture the parameter. Thus, when you build a 95% confidence interval based on your sample data, assuming of course that the sample was selected at random, then the chances are very good that the true mean is somewhere within the boundaries of the confidence interval. Since the confidence interval has a very high chance of including the true mean, then the confidence interval gives us a very good idea of a range in which the true means is likely to fall. The bottom line is: confidence intervals help us answer the question, "... what is the parameter?".

At this point, it might be helpful to present an example of a confidence interval problem according to the "typical" way it is done. In most instances, a researcher will select one sample and, based on the data from that one sample, try to make some inference about what the population parameter might be. So, what if we stay with the same example where the question is: what is the mean wattage for the population of 60 watt light bulbs produced by one manufacturer? A random sample is selected from the production line and the bulbs are tested for their wattage output. The data for the 25 bulbs are below.

DATA FOR ONE RANDOM SAMPLE WHERE n = 25 BULBS ARE TESTED

61	60	62	60	61	60	56	65	61	57	59
57	57	61	60	56	59	58	60	61	60	61
61	57	61								

DESCRIPTIVE STATISTICS ON THE SAMPLE

	N	MEAN	STDEV	MIN	MAX
Wattage	25	59.640	2.119	56	65

In the sample, the wattages ranged from 56 up to 65 with a mean of 59.64 and the estimate of the population standard deviation was 2.119. We can estimate the amount of sampling error by calculating the standard error of the mean by the following process.

Standard Error of Mean = S / Sqrt n = 2.119/5 = **.424**

It is very common practice to build 95% confidence intervals when working with inferential statistics. To do so, you therefore need to go approximately 2 units of error (two standard error of the mean units) on either side of the sample mean, as follows.

Lowpt95%CI = Mean - 2 * .424 = 59.64 - .848 = 58.792

Uppt95%CI = Mean + 2 * .424 = 59.64 + .848 = 60.488

Thus, based on this single random sample of n = 25 cases, the 95% confidence interval for the population mean goes from about 58.79 watts up to about 60.49 watts. We feel reasonably confident that the true mean, whatever it is, falls within the boundaries of our confidence interval. Important note: In this last example, I made no assumption about what the true mean was so do not necessarily think that it is 60. The fact is, when you take a sample, you do not know what the parameter is; in fact, that is what you are taking the sample for, to give you some idea of the value of the parameter. In the simulation examples above, we assumed the parameter to be a certain value simply to illustrate that confidence intervals will capture the parameter at that approximate "confidence percentage" rate. In real situations however, the parameter is unknown. If we knew the value of the parameter, we would not waste our time worrying about all this sampling error stuff!

Above, we have concentrated on the basic idea of confidence intervals in the context of using the mean of sample data to allow you to estimate the mean of the population. However, estimating the population mean is only one type of parameter estimation that may be of interest. There are many other parameters out there! For example, what if you developed a new measure of "attitudes about statistics" and wanted to correlate this measure with "performance in a statistics course". You could select a random sample from the appropriate population (students taking a statistics course) and then administer the attitude scale. Later, at the end of the course, you could obtain their statistics course grades. Assuming that higher values mean more positive attitudes, perhaps a correlation of .45 was obtained in your sample which suggests that more positive attitudes are associated with better course performance. The real question may be however: what is the correlation between these two variables (X=attitudes and Y=course performance) out there in the larger population? The concept of a confidence interval can assist us in providing an answer to that question. If we had a way to estimate sampling error of correlation coefficients (and there is), we could do the following to build 95% confidence intervals for the true population correlation.

Statistic +/- 2 units of error

r_{XY} +/- 2 units of S_r

The r value is the sample correlation and the S sub r represents the standard deviation of the sampling distribution of r. Remember, if many many samples were taken and r values computed each time, one would find that sample r values will vary (around the parameter) just like sample means vary around the population mean. Sometimes the r value will be too high and other times the sample r value will be too low. So, sampling error is involved in inferences about the population correlation just as it is involved in inferences about the population mean.

One last example before moving on to the concept of hypothesis testing. What if you were interested in what the variance of test scores is for a 50 item test given to all members of the population (perhaps all freshpersons at one university). To estimate this, you take a random sample of students, administer the test, and find the variance of the sample (S squared). But, we know that if we took other samples at random, the sample sets of data would show differing variance values. Some of the variance values would be too high compared to the population variance and some would be too low. Thus, sampling error of variances is also a reality. What we could do is to build a confidence interval around the sample variance where we have a good chance of capturing the true population variance. If we can estimate sampling error in this case, we can build the 95% CI like the following.

Statistic +/- 2 units of error

S^2 +/- 2 units of S_{S^2}

The S squared on the left is our sample variance estimate and the (right side value of) S sub S squared is the standard error of the variance since repeated samplings would show that the variance values (S squared) would vary from one sample to another. The standard deviation of the distribution of sample variances is the relevant standard error in the present case. Thus, whether you are interested in population means, correlation coefficients or variances, one could conceptually estimate sampling error and then build a confidence interval around the statistic to give you a better idea of the parameter value. Confidence intervals are very important in inferential statistical work and we will "bump" into them on several other occasions in future Chapters.

Testing a Hypothesis about a Population Mean

Finally, I want to introduce the concept of hypothesis testing which is the other inferential concern raised at the beginning of this Chapter. Rather than asking "what is the parameter?", we might have formulated a hypothesis about what the parameter is and then would let the sample data help to confirm or disconfirm that hypothesis. For the moment, and in this introduction to hypothesis testing, I will only cover one procedure for testing a hypothesis but, in later Chapters, we will cover other techniques.

To continue our "quest" of the true wattage for 60 watt light bulbs, what if we give the benefit of doubt to the manufacturer and formulate the hypothesis that the true mean wattage for all bulbs coming down that production line is really 60. In the literature, such a hypothesis is usually called H sub zero [H(0)] or the null hypothesis. Why it is called the "null" will be clarified later. So, for the moment, we state the hypothesis as follows.

$$H(0): F_X = 60 \text{ Watts}$$

Recall that the symbol F stands for the population mean. **A HYPOTHESIS IS ALWAYS A STATEMENT ABOUT THE PARAMETER OF INTEREST**! So, how can we test this hypothesis? First we must, as usual, take a nice random sample. Assume for a moment that we had selected the same sample (used above) where the data looked like the following.

Mean of sample	= 59.640 watts
Standard deviation of sample	= 2.119 watts
Standard error of the mean	= .424 watts

The simplest way to test the null hypothesis that the mean wattage is 60 would be to build a 95% confidence interval. Recall that the 95% confidence interval was as follows.

$$59.64 \pm (2 * .424)$$

$$58.79 <))))))))))))))) > 60.49$$

Since the 95% CI gives us a very good idea of what the true populations mean is, if the hypothesized value of 60 falls within the confidence interval, then we would conclude that a value of 60 could be the value for the population mean. That is, the available evidence would lead us to **RETAIN** the H(0) or null hypothesis and conclude that the true mean could be 60. And, for our example, that is what happens. The hypothesized value of 60 does indeed fall inside the confidence interval. So, for our sample and the confidence interval, our decision is to retain the null hypothesis.

On the other hand, if the H(0) or null value had not fallen within the confidence interval, we would have **REJECTED** the H(0) or null hypothesis. After all, values inside the confidence interval are the ones most likely to be the population mean whereas values outside of the confidence interval are the unlikely ones. Therefore, if the null value is outside the confidence interval, we consider it not to be a very good candidate for being the population mean.

What happens, in the unusual situation, if the H(0) or null hypothesis value falls exactly at the edge of the confidence interval? As was indicated previously, the rule will be to consider it outside if it falls at the exact edge of the confidence interval. Therefore, the H(0) value will be rejected if it falls at the boundary of the 95% confidence interval.

For the moment, the simplest way to test some null or H(0) hypothesis would be to first build a 95% confidence interval and then check to see if the null value falls inside or outside of the confidence interval. If H(0) falls inside, retain the null hypothesis. However, if the H(0) value falls at the exact edge or outside, then reject the null hypothesis. In later sections, we will explore further the concept of using confidence intervals for testing hypotheses further along with discussing what, in the literature, has become the "most common way" of testing hypotheses for a variety of parameters. Remember, one can test hypotheses about parameters other than the population mean. For example, previously I noted that one could build confidence intervals for the population correlation coefficient or the population variance. We can also test hypotheses about the population correlation or the population variance. For example, we could formulate a null hypothesis that states that the correlation between statistics attitudes and statistics course performance is 0. This may certainly not be what you believe but acts as a point of reference to compare to sample data. In this case, it is likely that you will want to reject the null hypothesis since your view is that there should be a positive correlation between these two variables. We could build a confidence interval around the sample correlation coefficient and then see if the null value of 0 was in or out of the interval. Finally, in regards to variances, we could also test the null hypothesis that the population variance of IQ scores is 225. It should be if the population standard deviation is about 15 since 15 squared is 225. Again, we could take sample data, build a confidence interval for the true population variance and then check to see if 225 is inside or outside of that interval. More on this later!

In brief summary, the two primary inferential concerns in statistical work are: to ask the question "what is the parameter?" or to formulate and test some hypothesis about the population parameter. The first concern is handled by building confidence intervals, usually

95% ones. To test a hypothesis, we can also use a confidence interval (in many situations). But, as we will see later, the more common procedures for testing a hypothesis are a little different than using the confidence interval approach described above although they produce the same identical decision with respect to the null hypothesis. As we discuss inferences about other parameters later, we will encounter these other methods for testing H(0) hypotheses.

Practice Problems

1. Note: This exercise applies if you have access to a statistical package that will allow you to randomly generate columns of data. To examine how confidence intervals capture the parameter, do the following. First, randomly put 200 values into 16 columns based on a normal distribution with mean = 100 and standard deviation = 15. Then, using the row as the sample, find the means of each sample and put them in another column. Make a distribution of that new column and find the descriptive statistics on that sampling distribution. Using the standard deviation of that column (which is the standard error of the mean), build 95% confidence intervals around the 200 sample means by going 2 units of error below and above each mean in each sample. Finally, since 100 is the population mean (from which the samples were drawn), see how many of the confidence intervals capture the value of 100.

2. You did a study where you selected a random sample of 16 students and gave them several statistics problems to solve using a calculator. You recorded the time, in minutes, for them to complete their work. These data are below.

   94, 65, 55, 63, 55, 75, 107, 77,
   97, 45, 84, 60, 54, 44, 80, 84

   Using the describe command and the mean and semean, build a 95% confidence interval for the population mean. If you had formulated a H(0) hypothesis that the true mean time should be 60 minutes, would you retain or reject that hypothesis?