12 – Confidence Intervals, Part I

A confidence interval is an interval within which a population statistic is believed to lie, with a quantified level of certainty. In order to make this understandable, consider how a typical confidence interval problem would end.

> Example 1: The 95% confidence interval for the mean age of college students is (21.34, 24.16). This translates to: *we are 95% confident that the average age of US college students is somewhere between 21.34 years and 24.16 years.*

When data is collected, the sample mean can be used to generate a confidence interval for population mean. Essentially, we will compute an "error" value to add and subtract from the sample mean for our interval.

To compute of a confidence interval for population mean with "large" data set of 30 or more values, we use the formula $\bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$. The parameters in this formula are:

$\bar{x}$ is the sample mean

z is a z-score from the standard normal distribution (determined by level of certainty)

σ is the population standard deviation

n is the number of data points in the sample
(must be 30 or more in order to use the standard normal distribution)

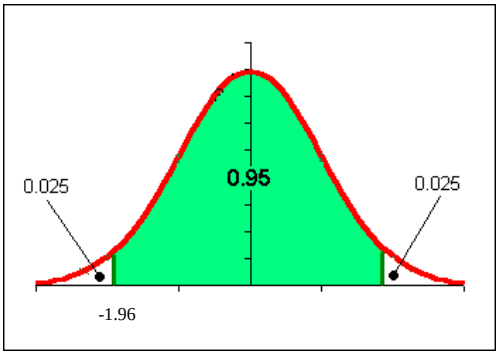
Month	Average Price of Gas	Standard Deviation
Jan 2009	1.86	0.29
Jul 2009	2.44	0.24
Jan 2010	2.65	0.29
Jul 2010	2.71	0.25
Jan 2011	3.08	0.20
Jul 2011	3.68	0.28
Jan 2012	3.37	0.27
Jul 2012	3.52	0.24
Jan 2013	3.29	0.30

Notice that σ (the population standard deviation) is one of the values needed for computation. At first glance, this may seem absurd, because if we knew the population standard deviation then we certainly must know the population mean. The reason it is not absurd is because standard deviations generally remain the same from scenario to scenario. For example, while the average price of gasoline in the US fluctuates, the standard deviation of the same data does not vary by much (see the table of means and standard deviations of gas prices on the right).

So, once a variable's standard deviation can be assumed to be a specific value, that value may then be used as the population standard deviation for future data.

> Example 2: A standardized test is used to assess students' knowledge of world events (max score is 75). The national reported standard deviation is 5. A sample of 30 Point Park students are tested and the mean score on the test is 49. Compute a 95 percent confidence interval based on this sample's data.

Solution: The values of $\bar{x}$, σ, and n are given directly in the problem: $\bar{x}$ = 49, σ = 5, and n = 30. We will use the formula $\bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$. The value of z needed will correspond to the 95% level of certainty prescribed by the problem. The value we seek is the z-score that corresponds to an area of 5% being *shared* by the two tails in the standard normal distribution. That is, we must look up the z-score that gives an area of 2.5% = .0250 to its left (see image). That value is z = -1.96.



The formula thus becomes $49 \pm (-1.96)\left(\frac{5}{\sqrt{30}}\right)$. That gives two values of 47.21 and 50.79. These numbers represent the lower and upper boundaries of the confidence interval and our conclusion is that we are 95% confident that the average score of all Point Park students would be between 47.21 and 50.79.

When the population standard deviation of a variable is not known and cannot be accurately estimated, of if our sample size is not sufficiently large (less than 30 data points), then we cannot use the standard normal distribution. Rather, we use the t-distribution (table F, p. 648, or in the foldout reference) and (unfortunately) looking up values on its respective table is different than looking up z-scores. However, the table in our textbook makes it simple for

confidence intervals: just use the column that corresponds to the prescribed confidence level. There is one other parameter necessary in looking up values of the *t*-distribution: *degrees of freedom (df)*. For now, *df = n – 1*.

The formula for confidence intervals that require the *t*-distribution is nearly the same as before: $\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$. Note that the only difference is that it uses *t* instead of *z* and the sample standard deviation (*s*) instead of the population standard deviation (σ).

> Example 3: Ten randomly selected high school students were asked how many hours they sleep at night. The mean was 6.8 hours with a standard deviation of 1.1 hours. Find a 90% confidence interval for the amount of time that high school students in general sleep at night.
>
> *Solution*: Since the population standard deviation is not known, the *t*-distribution must be used. The value in table F for a 90% confidence interval with nine degrees of freedom is 1.833 (note that *df = n* – 1.) So, the formula becomes $6.8 \pm (1.833)\left(\frac{1.1}{\sqrt{10}}\right)$, which gives values of 6.16 and 7.44 hours. So, we can be 90% confident that the average of the number of hours high school students sleep at night is between 6.16 and 7.44 hours.

When studies are first designed it often is necessary to determine the appropriate sample size for a specific confidence interval size. For example, suppose a researcher wishes to collect sample data in order to compute a 95% confidence interval for the average amount of debt carried by college graduates so that the interval is no more than $10,000 wide (Error is ± $5000). In this case, the formula for confidence interval can be used to find the number of college graduates to collect data from to achieve that precision. Specifically, if we solve the formula for *n*, we get $n = \left(\frac{z\sigma}{E}\right)^2$, where *E* is the "error," or desired precision.

> Example 4: A local bank wishes to know the average amount that its customers spend each month on cable television. Nationally, the standard deviation for cable service is $19.23. How many customers should the bank survey in order to estimate, with 90% confidence, the mean amount spent on cable to within $5?
>
> *Solution:* The values given to us are σ = 19.23 and *E* = 5. To find the appropriate value for *z*, we must find the nearest number to 0.05 in the standard normal distribution table. That value is -1.645. (note that 0.0495 and 0.0505 are equally spaced above and below the desired number and so we can use a value for *z* midway between their corresponding *z*-scores).
>
> The formula gives us $n = \left(\frac{-1.645 \times 19.23}{5}\right)^2 \approx 40.02$. Since this represents a number of people to be surveyed, we round up to 41. The bank should survey at least 41 of its customers.

When computing **confidence intervals for a population mean**, we must decide which distribution (*t* or *z*) to use.

Here is an easy way to decide: