Handout 03: Confidence Intervals

<u>Definition and Derivation</u>

Let θ be a quantity of interest that we are trying to estimate from a random sample drawn from a distribution $\mathcal{G}$. A **confidence interval** with **confidence level** $(1 - \alpha)$ is a pair of sample statistics L and U such that:

$$\mathbb{P}\left[L \leq \theta \leq U\right] \geq 1 - \alpha.$$

The idea is that we want to have a high probability that the quantity of interest falls between the lower bound L and upper bound U.

A standard approach to deriving a confidence interval is to start with a random random variable called a **pivot**. A pivot is defined as a function of the random sample and parameters defining the population $\mathcal{G}$ whose distribution does not depend on the unknown parameters. Let's walk through an example where $\mathcal{G}$ is equal to $N(\mu_X, 1)$ with an unknown mean μ_X. The following value is a pivot because, as we have written, it will have a standard normal distribution regardless of the value of μ_X:

$$Z = \frac{\mu_X - \bar{X}}{\sqrt{1/n}} \sim N(0,1).$$

Since we know the distribution of Z, we can write something that looks like a confidence interval for a given confidence level. For example, with $\alpha = 0.01$, we have:[1]

$$\mathbb{P}\left[-2.58 \leq Z \leq 2.58\right] \approx 0.99 = 1 - 0.01$$

$$\mathbb{P}\left[-2.58 \leq \frac{\mu_X - \bar{X}}{\sqrt{1/n}} \leq 2.58\right] \approx 0.99$$

To get the actual confidence interval, we manipulate the part inside the probability so that the parameter μ_X is alone in the middle and the lower and upper bounds depend only on the random sample:

$$\mathbb{P}\left[\bar{X} - 2.58 \cdot \sqrt{1/n} \leq \mu_X \leq \bar{X} + 2.58 \cdot \sqrt{1/n}\right] \approx 0.99$$

And that's really it! We now have a confidence intervale for the unknown mean parameter μ_X. One slightly more compact way to write this is as the following:

$$\bar{X} \pm 2.58 \cdot \sqrt{1/n}.$$

We could also write it down as an interval with the lower and upper bound, which looks nice when we are working with specific numbers.

[1] The first line is something that we would need to lookup on a table or use software such as R to compute. We will see that next class.

Notation for Tail Bounds

A handy notation for defining formulae for confidence intervals would be to replace the constant 2.58 with a more general term that we can fill in based on the confidence level. To do this, let z_α be the following quantity:[2]

$$\mathbb{P}[z_\alpha \geq Z] = \alpha, \quad Z \sim N(0,1).$$

This gives the tail probability that a standard normal is greater than the value z_α. Since the normal is symmetric around the origin, we can use $\pm z_{\alpha/2}$ as the parameter for a confidence interval. That is, we have the problem above, we would have the a confidence interval with confidence level $(1 - \alpha)$ given by the following:

$$\bar{X} \pm z_{\alpha/2} \cdot \sqrt{1/n}.$$

We will also define analogous quantities $\chi^2_\alpha(k)$, $f_\alpha(d_1, d_2)$, and $t_\alpha(k)$ for the chi-squared, F-, and t-distributions.

T-Statistic

Assume that we have two independent random variables: $Z \sim N(0,1)$ and $C \sim \chi^2(k)$. Then, define the following ratio between the two random variables:

$$T = \frac{Z}{\sqrt{C/k}}.$$

It should be clear that this random variable has a well defined distribution that depends only on the degrees of freedom k of the chi-squared distribution. The distribution is called **Student's t-distribution** with k degrees of freedom, which is denoted by $t(k)$. This will be important for our work today to derive the standard confidence interval for the mean of an unknown distribution.

This will be useful combined with the facts we showed last week. Namely, that if $\mathcal{G}$ is a normal distribution, then we have the following:

$$\frac{\mu_X - \bar{X}}{\sqrt{\sigma^2_X/n}} \sim N(0,1)$$

$$\frac{(n-1)S^2_X}{\sigma^2_X} \sim \chi^2(n-1).$$

We will create a pivot statistic based on these two quantities and then derive a confidence interval for the mean.