Handout 03: Confidence Intervals

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 $N(\theta,1)$ with an unknown mean θ. The following value is a pivot because, as we have written, it will have a standard normal distribution regardless of the value of θ:

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

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:

$$\mathbb{P}\left[-2.58 \leq Z \leq 2.58\right] \approx 0.99 = 1 - 0.01$$
$$\mathbb{P}\left[-2.58 \leq \frac{\theta - \bar{X}}{\sqrt{1/n}} \leq 2.58\right] \approx 0.99$$

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

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

We see that we can get a confidence interval by picking something centered on the sample mean with length $2 \cdot 2.58 \cdot \sqrt{1/n}$.

A handy notation for defining formulae for confidence intervals is to define z_α to be the following quantity:

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

We will also define analogous quantities $t_\alpha(k)$ and $\chi^2_\alpha(k)$ for the t-distribution and chi-squared distributions. Replacing this with the ± 2.58 above, we have the more general formula:

$$\mathbb{P}\left[\bar{X} + z_{\alpha/2} \cdot \sqrt{1/n} \leq \frac{\theta - \bar{X}}{\sqrt{1/n}} \leq \bar{X} + z_{1-\alpha/2} \cdot \sqrt{1/n}\right] \approx 1 - \alpha$$

The confidence interval above is valid for any confidence level α. Because the normal distribution is symmetric around the origin, we have that $z_{\alpha/2} = -z_{1-\alpha/2}$. This means that you could rewrite the confidence interval as:

$$\mathbb{P}\left[\bar{X} - z_{1-\alpha/2} \cdot \sqrt{1/n} \leq \frac{\theta - \bar{X}}{\sqrt{1/n}} \leq \bar{X} + z_{1-\alpha/2} \cdot \sqrt{1/n}\right] \approx 1 - \alpha$$

Or even:

$$\bar{X} \pm z_{1-\alpha/2} \times \sqrt{\frac{1}{n}}.$$

The latter is a common way of writting a confidence interval for a mean.

The example above is quite artificial because it assumes that we already known the variance of $\mathcal{G}$. On today's worksheet, we will see that the following is a pivot statistic in the general case where the distribution is $N(\mu, \sigma^2)$:

$$T = \frac{\mu - \bar{X}}{\sqrt{S_X^2/n}} \sim t(n-1).$$

Importantly, it can also be used as an approximation for any $\mathcal{G}$ with finite mean and variance for large n due to the central limit theorem. For reference, here is the confidence interval that we will be deriving:

$$\bar{X} \pm t_{1-\alpha/2} \times \sqrt{\frac{S_X^2}{n}}.$$

In addition to being a helpful formula to have as a computational tool, this quantity also helps conceptualize how our ability to estimate a mean scales with the desired confidence (α), the variation in the data (S_X^2), and the sample size (n).