STAT 231 — LECTURE 20

Bartosz Antczak *Instructor: Michael Wallace* *October 27, 2017*

20.1 More on Interval Estimation

Last Time

We learned about *interval estimation*. Particularly, a $100p\%$ *confidence interval* for a parameter is an interval estimate $[L(y), U(y)]$ for which

$$P(\theta \in [L(Y), U(Y)]) = P[L(Y) \leq \theta \leq U(Y)] = p$$

This definition means that out of all of the independent random samples we collect, 95% of them will have the true value of θ within the range $[L(y), U(y)]$.
We define p as the **confidence coefficient**. Also, note that $P(\theta \in [L(y), U(y)])$ is *either* 1 or 0.

Example 20.1.1. *Constructing a confidence interval.*

Suppose we have a normally distributed population with unknown mean μ and known standard deviation 1. We have n random samples $Y_1, Y_2, \ldots, Y_n \sim G(\mu, 1)$. From this, we also know that $E[Y_i] = \mu$ and $sd(Y_i) = 1$. Recall that the standard deviation of a normal distribution is defined as $\frac{\sigma}{\sqrt{n}}$. Our confidence interval for the true value of μ is then defined as:

$$\left(\overline{Y} - \frac{1.96}{\sqrt{n}}, \overline{Y} + \frac{1.96}{\sqrt{n}}\right)$$

and so we have

$$P\left[\mu \in \left(\overline{Y} - \frac{1.96}{\sqrt{n}}, \overline{Y} + \frac{1.96}{\sqrt{n}}\right)\right] = 0.95$$

Recall that for a normal distribution, approximately 95% of data is contained within 1.96 standard deviations of the mean, and so $\overline{Y} \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
And so, to conclude, our **95% confidence interval** for the unknown mean μ is

$$\left(\overline{y} - \frac{1.96}{\sqrt{n}}, \overline{y} + \frac{1.96}{\sqrt{n}}\right)$$

Note: that the size of the interval is $2\left(\frac{1.96}{\sqrt{n}}\right)$, which means that as the number of samples increases, the size of our interval decreases. This should make intuitive sense because we're able to construct a more accurate estimate given our sample data.

Applying some numbers

Say for a sample size of $n = 16$ and an observed sample mean of $\overline{y} = 10.4$, our 95% confidence interval is

$$\left(10.4 - \frac{1.96}{\sqrt{n}}, 10.4 + \frac{1.96}{\sqrt{n}}\right)$$

Recall that we **cannot** say that $P(\mu \in (10.4 - \frac{1.96}{\sqrt{n}}, 10.4 + \frac{1.96}{\sqrt{n}})) = 0.95$. The probability of the true value of μ being in a particular range is either 1 or 0 (i.e., it's in the range or it isn't). We can only say that we are 95% confident that μ is in that particular range.

Definition

A **pivotal quantity** is a function of the data Y and the unknown parameter θ such that the distribution of Q is completely known. Pivotal quantities can be used to construct confidence intervals.

Example 20.1.2. *Determining the pivotal quantity of a normal distribution*

Suppose we have the data $Y_1, Y_2, \ldots, Y_n \sim G(\mu, \sigma)$, where μ is unknown and σ is known. A *point estimator* for μ is $\widetilde{\mu} = \overline{Y}$. Its *sampling distribution* is

$$\widetilde{\mu} = \overline{Y} \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

What could be a possible pivotal quantity for this data? Well, since

$$\frac{\overline{Y} - \mu}{\sigma \ \sqrt{n}} \sim G(0, 1)$$

is a completely known distribution, we can define $Q(Y, \mu) = \frac{\overline{Y} - \mu}{\sigma \ \sqrt{n}}$.

Constructing a Confidence Interval using Pivotal Quantities

We follow a series of steps (*these are very general*):

1. Determine numbers a, b such that
$$P(a \leq Q(Y, \theta) \leq b) = p$$
2. Re-express the inequality as such:
$$p = P(a \leq Q(Y, \theta) \leq b) = P(L(Y) \leq \theta \leq U(Y))$$
3. For the observed data y, the interval $[L(Y), U(Y)]$ is a $100p\%$ confidence interval for θ.

So let's apply an example to understand these steps. As stated before, our pivotal quantity is $Q(Y, \mu) = \frac{\overline{Y} - \mu}{\sigma \ \sqrt{n}}$. Let's find a 95% confidence interval.

Step 1

Let's find values a, b. Since $Q \sim G(0, 1)$, we know that $P(a \leq Z \leq b) = 0.95 \implies -a = b = 1.96$.
Note that there are an infinite number of values for a, b, but we want to find such values that are *symmetric* (i.e., $-a = b$). And so we have

$$P\left(-1.96 \leq \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \leq 1.96\right)$$

Step 2

Isolate μ using basic algebra:

$$\left(\overline{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Step 3

A 95% confidence interval for μ based on the observed data $y_1, y_2, \ldots, y_n$ is:

$$\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{n}}\right) \quad \square$$