8.1: Statistical Intervals Based on a Single Sample

Taylor

University of Virginia

Motivation

Say you want to estimate θ with $\hat{\theta}$. It's never going to be the case (with probability 1) that $\hat{\theta} = \theta$.

We need some routine where we can pick a confidence level (0-100%) and be able to get an interval that we know will contain θ that percent of the time.

Example

Recall that when $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, we have $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. So we can say stuff like

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

where z_c is the $1 - c$th standard normal percentile.

In this example we'll assume μ is unknown, but σ^2 is known. This is unrealistic, but it's our first example.

Motivation

Remember how we can do algebra inside the probability parentheses. Then we have

$$P\left(-\frac{\sigma}{\sqrt{n}}z_{\alpha/2} \leq \bar{X} - \mu \leq \frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right) = 1 - \alpha$$

or

$$P\left(\frac{\sigma}{\sqrt{n}}z_{\alpha/2} \geq \mu - \bar{X} \geq -\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right) = 1 - \alpha$$

or

$$P\left(\bar{X} + \frac{\sigma}{\sqrt{n}}z_{\alpha/2} \geq \mu \geq \bar{X} - \frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right) = 1 - \alpha$$

So before we even observe any random data, if our data are a random sample from a normal distribution, and we know σ, we know that $\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$ covers/contains μ exactly $(1 - \alpha)$ percent of the time.

It's important to mention, however, that after we observe data, and we compute this interval, we can no longer use the word "probability" since nothing is random after the data are observed. If you are asked to interpret a confidence interval, you're not allowed to do this. Rather, you should say that if we were allowed to repeat this data collection procedure over and over again, and we computed the same interval each time, then in the long run $(1 - \alpha)$ percent of these would cover the true μ.

Motivation

Let's take a closer look at our first CI formula: $\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$. The width of this interval is

$$w = (\bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}) - (\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}) = 2 \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

1. if $n \uparrow$, then width $\downarrow$
2. if $\sigma^2 \uparrow$, then width $\uparrow$
3. if $(1 - \alpha) \uparrow$, then $z_{\alpha/2} \uparrow$, which means width $\uparrow$

Motivation

Sometimes you are given a width, and a desired confidence level, and then asked how much data are needed to achieve this desired width. This new formula

$$n = \left(2z_{\alpha/2}\frac{\sigma}{w}\right)^2$$

is just a rearrangement of the old formula

$$w = 2\frac{\sigma}{\sqrt{n}}z_{\alpha/2}.$$

Motivation

How can we generalize our process to give us a confidence interval for more parameters, taken one at a time?

We need two things. First, we need a random variable that depends on *both* the data *and* the parameter of interest. We'll call this $h(X_1, \ldots, X_n; \theta)$. Second, we need the probability distribution of this $h(X_1, \ldots, X_n; \theta)$ to not have any θ involved in it.

If you want to think in terms of the last example.
$h(X_1, \ldots, X_n; \theta) = Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. Notice that the representation/label of Z depends on μ, but its probability distribution doesn't have a μ in it.

Some textbooks call $h(X_1, \ldots, X_n; \theta)$ a pivotal quantity, but this textbook doesn't give this thing a name.

Example 8.5 on page 389

$X_1, \ldots, X_n \overset{iid}{\sim}$ Exponential(λ). Find a 95% confidence interval for λ.

Example 8.5 on page 389

$X_1, \ldots, X_n \overset{iid}{\sim}$ Exponential(λ). Find a 95% confidence interval for λ.

Let $h(X_1, \ldots, X_n; \lambda) = 2\lambda \sum_i X_i$. Using the arguments we've talked about for gamma random variables

1. $X_1, \ldots, X_n \overset{iid}{\sim}$ Exponential(λ) so
2. $X_1, \ldots, X_n \overset{iid}{\sim}$ Gamma($1, \frac{1}{\lambda}$) so
3. $\sum_i X_i \sim$ Gamma($n, \frac{1}{\lambda}$) so
4. $2\lambda \sum_i X_i \sim$ Gamma($n, 2$) so
5. $2\lambda \sum_i X_i \sim \chi^2_{2n}$

This distribution depends functionally on the unknown parameter of interest, but it has a known distribution...

Example 8.5 on page 389

Let's use specific numbers now. We have $n = 10$ data points. So we get the .025 and the .975 quantiles from that $\chi^2_{2 \cdot 10}$ distribution (use excel/tables/calculator or some software you have). They turn out to be 9.590777 and 34.16961, respectively.

$$P(9.590777 \leq 2\lambda \sum_i X_i \leq 34.16961) = .95$$

which is true if and only if

$$P\left(\frac{9.590777}{2\sum_i X_i} \leq \lambda \leq \frac{34.16961}{2\sum_i X_i}\right) = .95$$

This means our confidence interval is $\left[\frac{9.590777}{2\sum_i X_i}, \frac{34.16961}{2\sum_i X_i}\right]$