8.3: Intervals Based on a Normal Population Distribution

Taylor

University of Virginia

Motivation

This section talks about the difference between three types of intervals when dealing with normal data. Unlike last section, the assumption of normal data is required here.

Proposition

First let's recall some facts about a t_ν distribution:

1. it has one parameter: ν
2. it doesn't have an MGF
3. it only has a mean if $\nu > 1$
4. it only has a variance if $\nu > 2$
5. it's basically a standard normal distribution with fatter tails
6. as $\nu \to \infty$, $t_\nu \overset{D}{\to} \mathcal{N}(0, 1)$

Motivation

In this chapter, we only use it for its percentiles:

$$t_{\alpha,\nu}$$

will denote the $(1-\alpha)$ 100th percentile.

The book calls these **critical values**.

Definition

Recall in section 6.4 We proved that $T = \frac{(\bar{X}-\mu)}{s/\sqrt{n}} \sim t_{n-1}$ when $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

so

$$P\left(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}\right) = 1 - \alpha$$

and with a little algebra we can show that

$$P\left(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

...so our **confidence interval** *for* μ is $[\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}]$

Motivation

This is an interval for μ. That means that in the long run, CIs constructed this way will cover μ $(1-\alpha)100$ percent of the time.

What if our real goal is to predict the next data point? We want something like this:

$$P(lower \leq X_{n+1} \leq upper) = 1 - \alpha$$

Prediction Error

First, $X_{n+1} - \bar{X}$ is going to be the random prediction error. X_{n+1} is still coming from a normal distribution with center μ, so we're going to use $\bar{X}$ again as a point estimate.

$$E(\text{prediction error}) = E[X_{n+1} - \bar{X}] = EX_{n+1} - E\bar{X} = 0$$

and

$$V[\text{prediction error}] = V[X_{n+1} - \bar{X}] = V[X_{n+1}] + V[\bar{X}] = \sigma^2 + \frac{\sigma^2}{n}$$

We're using independence in the variance calculation.

Prediction Error

Since a linear combination of normals is normally distributed, we have

$$\frac{(X_{n+1} - \bar{X}) - 0}{\sqrt{\sigma^2 + \frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$$

and by similar reasoning as in section 6.4 (this will be a quiz problem), we can show that

$$\frac{(X_{n+1} - \bar{X}) - 0}{\sqrt{s^2 + \frac{s^2}{n}}} \sim t_{n-1}$$

Prediction Error

Last steps:

$$P\left(-t_{\alpha/2,n-1} \leq \frac{(X_{n+1} - \bar{X}) - 0}{\sqrt{S^2 + \frac{S^2}{n}}} \leq t_{\alpha/2,n-1}\right) = 1 - \alpha$$

$$P\left(-t_{\alpha/2,n-1}\sqrt{S^2 + \frac{S^2}{n}} \leq X_{n+1} - \bar{X} \leq t_{\alpha/2,n-1}\sqrt{S^2 + \frac{S^2}{n}}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{\alpha/2,n-1}S\sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{X} + t_{\alpha/2,n-1}S\sqrt{1 + \frac{1}{n}}\right) = 1 - \alpha$$

so $[\bar{x} \pm t_{\alpha/2,n-1}s\sqrt{1 + \frac{1}{n}}]$ is our $(1 - \alpha)100$th **prediction interval**

Motivation

Key points so far

1. CIs are for inferring about *parameters*
2. PIs are for inferring about future *obeservations*
3. Prediction variance is non-vanishing, no matter how much data you throw at it