STAT 231 — Lecture 24

Bartosz Antczak *Instructor: Michael Wallace* *November 6, 2017*

Last Time

We learned about the $t-$distribution. We use the following result: suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, \sigma)$ distribution. Then

$$A = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where S is a point estimator defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

So the distribution A is a function of the data and unknown parameter μ. This can be used to construct confidence intervals for μ!

24.1 Gaussian Data with Unknown Parameters

Say we want to determine the confidence interval of some random sample of data $Y_1, Y_2, \ldots, Y_n \sim G(\mu, \sigma)$ where $E[Y_i] = \mu$ and $sd(Y_i) = \sigma$ are **both** unknown. We will show how to do this with an example.

Example 24.1.1. *Bad vs. good movies*

Some summary statistics for movie lengths are shown. Movies are classified as either *bad* or *good*.

	Bad	**Good**
n	42	58
Mean	86.190	97.707
SD	16.251	18.229

Bad Movies

It seems reasonable to assume the model

$$Y_i \sim G(\mu, \sigma) \qquad i = 1, 2, \ldots, 42$$

where Y_i is the length of the ith bad movie, in minutes.
Since we don't know σ, we must refer to our point estimate of $s = 16.251$ (defined above) and $\mu = 86.190$. Our confidence interval will follow the structure

$$\overline{y} \pm a\frac{s}{\sqrt{n}}$$

where $P(T \leq a) = (1+p)/2$ and $T \sim t_{41}$. To calculate a for a 95% CI, we refer to our distribution table for $T \sim t_{41}$ and look up $P(T \leq a) = 0.975$, which results in $a = 2.019541$.

Therefore a 95% CI for μ of bad movies is

$$86.190 \pm 5.064168 \implies [81.126, 91.254]$$

We can interpret this result as:

- We are 95% confident that $\mu \in [81.126, 91.254]$
- 86.190 is a reasonable guess for μ
- A plausible range of values for μ is $[81.126, 91.254]$

Note: if we assumed σ was known, our CI would be *narrower*. This is because we're referring to the $G(0,1)$ distribution rather than the $t-$distribution. This reflects the natural result that we're more certain in our outcome.

Good Movies

Applying the same logic as for bad movies, we get the 95% confidence interval $[92.914, 102.500]$

24.1.1 CI Characteristics

Given a 95% CI, what would happen to its width if we applied the following changes (recall the width of a CI is $\frac{2as}{\sqrt{n}}$:

- **CI increases to 99%:** width increases (to accommodate the greater certainty)
- **Sample Size increases** width decreases (since we're more confident with a larger sample size)
- **Sample std. deviation decreases:** width decreases (since the distribution of elements is closer together)
- **Sample mean decreases:** width remains the same (since the width is not dependent on sample mean)

Say we want to reduce the CI to have a width of 2. What should our sample size be? We define the width of a CI to be $2d$, that way we have the nice range of

$$\overline{y} \pm d$$

And so, to answer this question, we should choose a sample size n such that

$$a\frac{\sigma}{\sqrt{n}} \approx d \implies n \approx \left(\frac{a\sigma}{d}\right)^2$$

As an **example**, say we want to have a 95% CI for μ of width 2 for our good movies. To determine how many more films (on top of the already 58 sampled) we'd have to create an estimate for σ (which is $s = 18.229$) and refer to the appropriate value that satisfies $P(Z \leq a) = p$. Also we set $d = 1$ since the width we want is $2 = 2d$. From there, we simply calculate

$$n \approx \left(\frac{a\sigma}{d}\right)^2 = 1276.55$$

24.1.2 CI for Unknown Sigma

So we went over how to construct a CI for μ, now let's see how to build one for σ.
First, recall the *point estimator* for σ^2 is:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

From this, we'll use the following theorem:

Theorem

Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, \sigma)$ distribution. Then

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

This means that the random variable $\frac{(n-1)S^2}{\sigma^2}$ is a function of the data $Y_1, Y_2, \ldots, Y_n$ with unknown parameter σ. This means it is a pivotal quantity and we can use it to construct a CI for σ! Oh frick yeah pal.

Building our CI

The first thing we do is determine our confidence coefficient. Notice however, that the chi-squared distribution is not symmetric, so we cannot simply choose one value a like we have been doing.
For a $100p\%$ confidence interval, we must determine $P(a \leq W \leq b) = p$, with $W \sim \chi^2(n-1)$. We choose our coefficients by determining:

- $P(W \leq a) = (1-p)/2$
- $P(W \geq b) = (1-p)/2$ or $P(W \leq b) = (1+p)/2$

Once we have those values, we simply calculate

$$P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = p \tag{24.1}$$

$$P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) = p \qquad \text{(Rearrange)} \tag{24.2}$$

And so a $100p\%$ CI for σ^2 is

$$\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right)$$

and trivially a CI for σ is

$$\left(\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right)$$