Week 5 Confidence interval

3 Student t model

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

In general, a confidence interval is of the form:

$$\text{Point Estimate} \pm \text{Critical value} \times \text{SE}$$

How do you compute SE?

Proportion

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Mean

$$SE = \frac{\sigma}{\sqrt{n}}$$

The central limit theorem says ...

- For a sample proportion

$$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

where p is the population proportion, and n is the sample size

- For a sample mean

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

where μ is the population mean, σ is the population standard deviation and n is the sample size

$$\text{CI : point estimate} \pm \text{critical value} \times \text{SE}$$

If the parameter of interest is the population mean, and the point estimate is the sample mean,

$$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

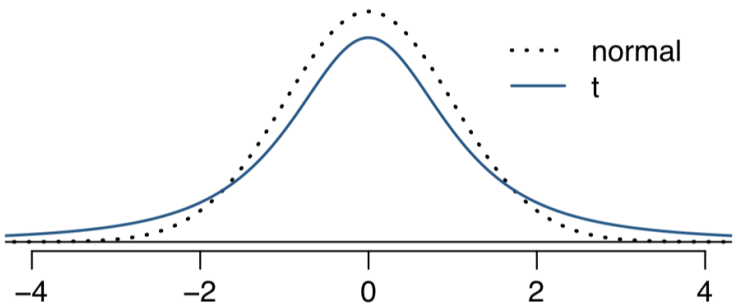
- z^* is critical value, which comes from a standard normal model $\mathcal{N}(0, 1)$
- $\bar{x}$ is the sample mean
- s is the sample standard deviation

How reliable is the estimate s?

Plugging in an estimate introduces additional uncertainty. We make up for this by using a more "conservative" distribution than the normal distribution.

t-distribution also has a bell shape, but its tails are *thicker* than the normal model's

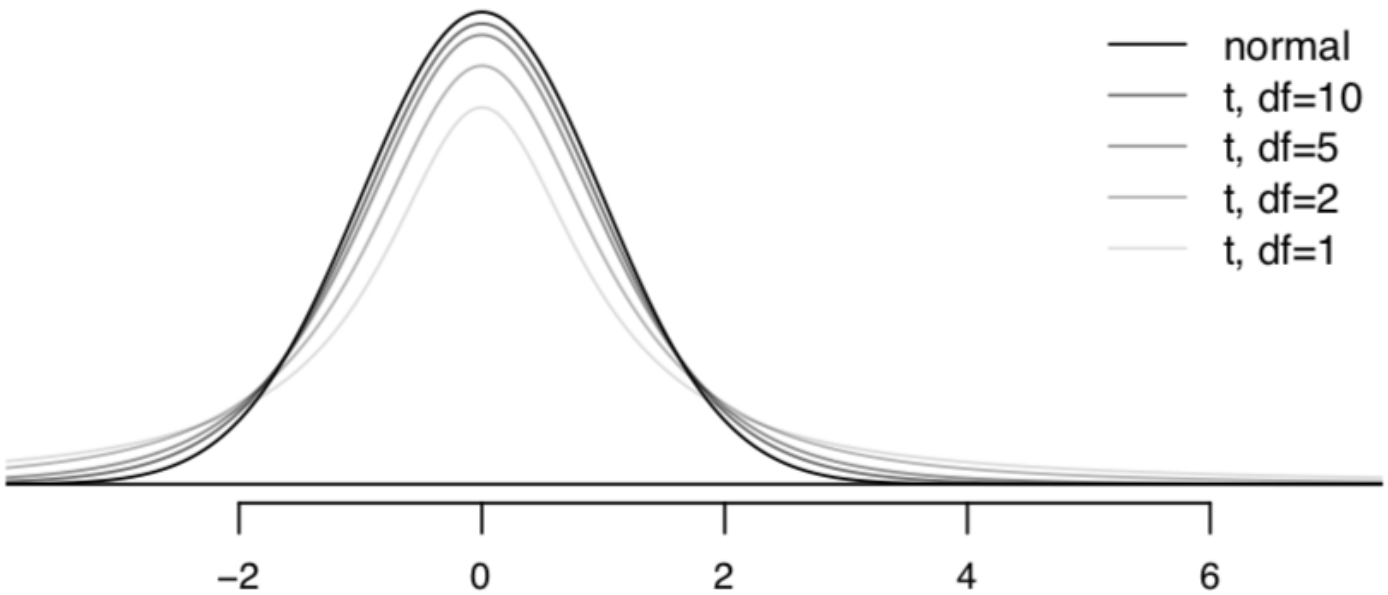
- Observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- Extra thick tails help mitigate the effect of a less reliable estimate for the standard error of the sampling distribution.



William Sealy Gosset

- Always centered at zero, like the standard normal distribution
- Has a single parameter, degrees of freedom (df), that is tied to sample size.

$$df = n - 1$$



How to make confidence intervals for means with t-distribution?

$$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

The critical value t^*_{n-1} depends on

- confidence level
- degrees of freedom, n - 1, which we get from the sample size.

In 2004, a team of researchers published a study of contaminants in farmed salmon. One of those contaminants was the insecticide mirex, which has been shown to be carcinogenic and is suspected to be toxic to the liver, kidneys, and endocrine system.

Summaries for the mirex concentrations (in parts per million ppm) inthe farmed salmon are

$$n = 150 \qquad \bar{x} = 0.0913\text{ppm} \qquad s = 0.0495\text{ppm}$$

What does a 95% confidence interval say about mirex?

The summaries are given as

$$n = 150 \qquad \bar{x} = 0.0913\text{ppm} \qquad s = 0.0495\text{ppm}$$

We can compute

- $df = n - 1 = 149$, $t^*_{149} = 1.977$
- $\text{SE} = \frac{s}{\sqrt{n}} = \frac{0.0495}{\sqrt{150}} = 0.0040$

So the 95% confidence interval for μ is

$$\begin{aligned} \bar{x} \pm t^*_{n-1} \times \text{SE} &= 0.0913 \pm 1.977 \times 0.0495 \\ &= 0.0913 \pm 0.0079 = (0.0834, 0.0992) \end{aligned}$$

I'm 95% confident that the mean level of mirex concentration in farm-raised salmon is between 0.0834 and 0.0992 parts per million

These are two conditions we need to check before using the Student's t-models.

- Randomization: This condition is satisfied if the data arise from a random sample or suitably randomized experiment
- Nearly Normal Condition: The data come from a distribution that is unimodal and symmetric.

You should check the Nearly Normal Condition by making a histogram or Normal probability plot.

- $n < 15$ or so: the data should follow a Normal model pretty closely
- n between 15 and 40 or so: the data should be unimodal and reasonably symmetric
- $n > 40$: safe to use unless the data are extremely skewed