Welcome to STA 101!

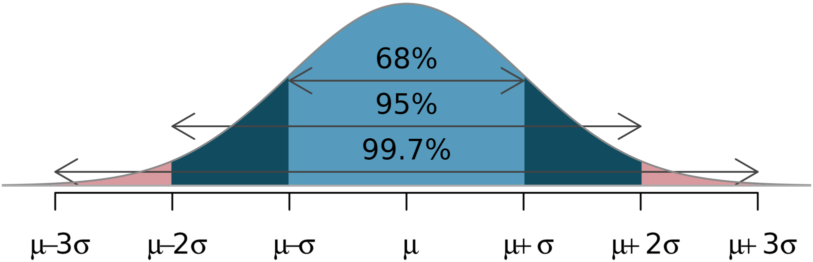
The normal distribution (bell curve)

Notation: $X \sim \mathrm{N}(\mu, \sigma)$.

Two parameters:

- μ: "mu." The mean. Controls location of the middle;
- σ: "sigma." The standard deviation. Controls spread.

The 68-95-99.7 rule:



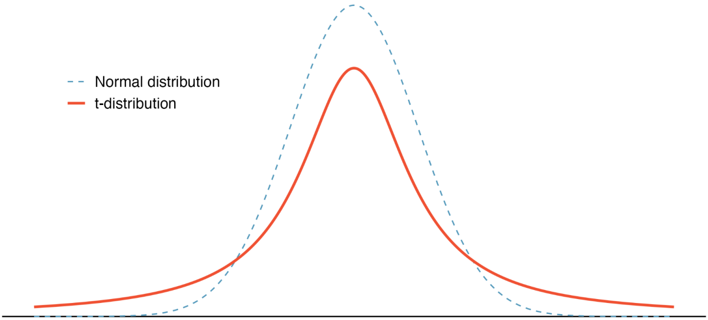
Student's t-distribution

Notation: $X \sim t_\nu$.

One parameter:

- ν: "nu." The degrees of freedom.

Heavier tails than the normal distribution:



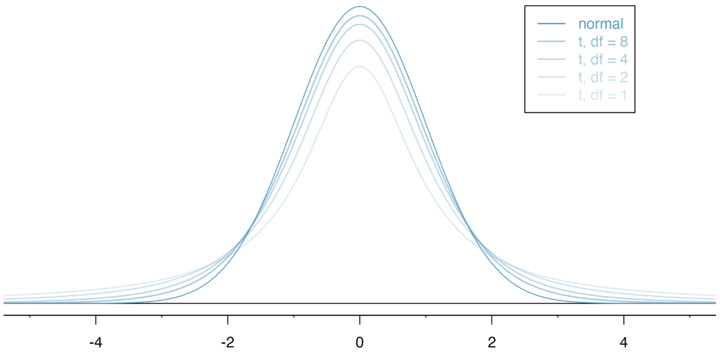
Student's t-distribution

Notation: $X \sim t_\nu$.

One parameter:

- ν: "nu." The degrees of freedom

Closer to standard normal as DoF increase:



Recap: one-sample t-test

One sample from normal distribution:

$$x_1,\, x_2,\, \ldots,\, x_n \sim \mathrm{N}(\mu,\, \sigma).$$

Point estimates:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2}.$$

How do we do interval estimation and testing for μ?

- We already know how to approximate with simulation;
- Assuming normality, can we do better than just approximate?

Recap: one-sample t-test

Since we assume

$$x_1,\, x_2,\, \ldots,\, x_n \sim \mathrm{N}(\mu,\, \sigma),$$

it turns out that

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

and so

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}\left(0,\, 1\right).$$

If instead we plug in the *estimate* of σ, we have

$$\frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}.$$

Heavier tails reflect extra estimation uncertainty from $\hat{\sigma}$.

Recap: one-sample t-test

Based on the fact that

$$\frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1},$$

you can construct an *exact* confidence interval

$$\bar{x} \pm t^{\star}_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}},$$

and perform an exact test of

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

using this test statistic and null distribution:

$$\frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}.$$

Today: beyond one-sample inference for a mean

We will consider other estimation problems based on data from the normal distribution, but in all cases, the template is the same:

1. Start here:

$$\frac{\texttt{estimate} - \text{true value}}{\widehat{\texttt{standard error}}} \sim t_{df};$$

2. Get a confidence interval:

$$\texttt{estimate} \pm t^\star_{1-\alpha/2}\widehat{\texttt{standard error}}.$$

3. Run a test based on this statistic and null distribution:

$$\frac{\texttt{estimate} - \text{null value}}{\widehat{\texttt{standard error}}} \sim t_{df};$$

Depending on the problem, you have slightly different formulas for $\texttt{estimate}$, $\widehat{\texttt{standard error}}$, and df. That's it.

Two-sample t-test

We want to compare the means of two independent samples:

$$x_1, x_2, \ldots, x_{n_1} \sim \mathrm{N}(\mu_x, \sigma_x)$$
$$y_1, y_2, \ldots, y_{n_2} \sim \mathrm{N}(\mu_y, \sigma_y).$$

Point estimates:

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \qquad \hat{\sigma}_x = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2}$$

$$\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \qquad \hat{\sigma}_y = \sqrt{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2}.$$

We can estimate $\texttt{diff} = \mu_x - \mu_y$ with $\widehat{\texttt{diff}} = \bar{x} - \bar{y}$.

What about interval estimation and testing?

Same starting place...

It turns out that

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\widehat{\text{SE}}} \sim t_{df}.$$

If you knew the true variances, then

$$\text{SE} = \sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}.$$

But you don't, so

$$\widehat{\text{SE}} = \sqrt{\frac{\hat{\sigma}_x^2}{n_1} + \frac{\hat{\sigma}_y^2}{n_2}}.$$

The true degrees of freedom formula is ugly, but this works:

$$df = \min\{n_1 - 1,\ n_2 - 1\}.$$

Two-sample t-test

Starting from

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_1}+\frac{\hat{\sigma}_y^2}{n_2}}} \sim t_{df},$$

You get an interval:

$$(\bar{x}-\bar{y}) \pm t^{\star}_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_x^2}{n_1}+\frac{\hat{\sigma}_y^2}{n_2}}.$$

You get a test of

$$H_0: \mu_x-\mu_y = 0$$
$$H_A: \mu_x-\mu_y \neq 0$$

based on

$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{\sigma}_x^2}{n_1}+\frac{\hat{\sigma}_y^2}{n_2}}} \sim t_{df},$$

Paired data

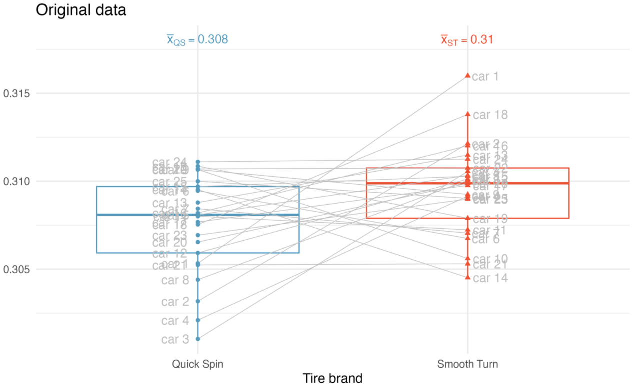


Figure 21.1: Box plots of the tire tread data (in cm) and the brand of tire from which the original measurements came.