Welcome to STA 101!

Final project proposal: due 5PM tomorrow

- Don't forget to pick a team name;
- Graded for completion;
- Compose the doc however you want;
- One member submits in Gradescope and tags everyone else;
- Use labs and OH to meet and get TA feedback;
- Make sure you link somewhere that I can directly download the data myself;
- You will receive detailed feedback from me about
  - technical advice;
  - which project is more interesting and/or feasible.

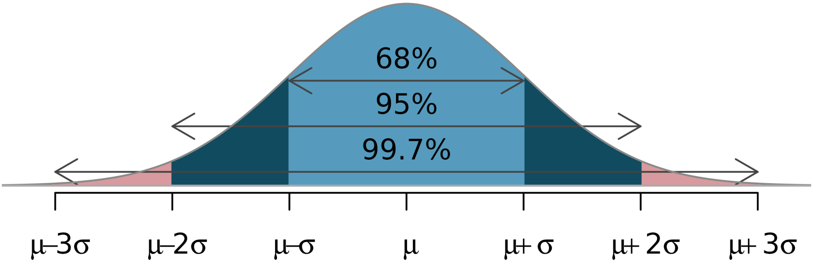
The normal distribution (bell curve)

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:



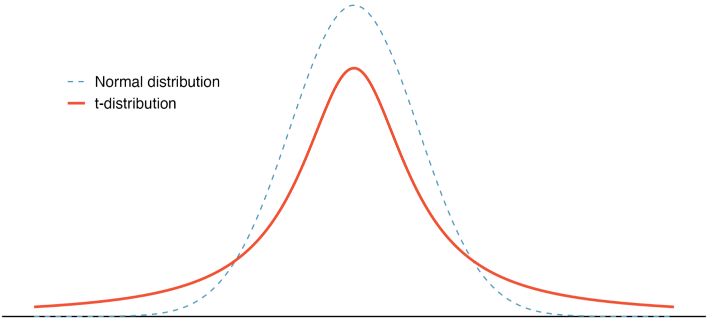
Student's t-distribution

Notation: $X \sim t_\nu$.

One parameter:

- ν: "nu." The degrees of freedom.

Heavier tails than the normal distribution:



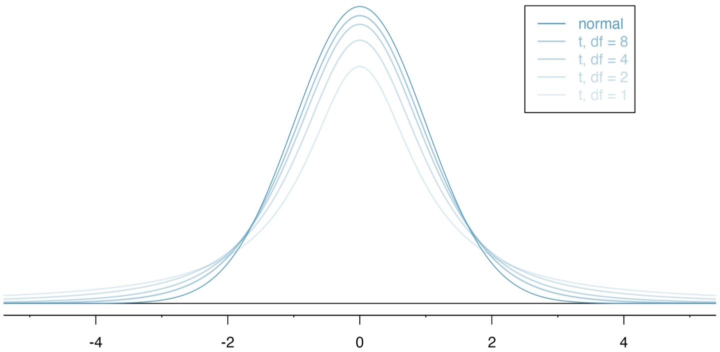
Student's t-distribution

Notation: $X \sim t_\nu$.

One parameter:

- ν: "nu." The degrees of freedom

Closer to standard normal as DoF increase:



Estimation problem: mean of normal data

Data: a list of numbers x_1, x_2, ..., x_n;

Unknown population: $\mathrm{N}(\mu, \sigma)$.

Point estimate: sample average

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

To do statistics, we must access the sampling distribution of $\bar{x}$.
We have seen two methods for doing this:

- histogram approximation via the bootstrap;
- normal approximation: $\mathrm{N}\left(\bar{x}, \hat{\mathrm{SE}} = \hat{\sigma}/\sqrt{n}\right)$.

Forget approximation. With normal data, we can do the real thing.

Sampling distribution of the sample average

The data are random:

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

The average is a function of the data, so it is random too:

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

This is its *sampling distribution*

The confidence interval formulas

We've seen several confidence intervals for μ:

$$\bar{x} \pm z^\star \frac{\sigma}{\sqrt{n}} \qquad \text{correct when } \sigma \text{ known}$$

$$\bar{x} \pm z^\star \frac{\hat{\sigma}}{\sqrt{n}} \qquad \approx \text{correct when } n \text{ large}$$

$$\bar{x} \pm t^\star \frac{\hat{\sigma}}{\sqrt{n}} \qquad \text{correct}$$

Where do these come from?

Standardization: subtract off the mean

If

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

then

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

So you make the mean zero.

Standardization: divide by the standard error

If

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

then

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

So you make the standard deviation 1.

Standardization: putting it together

If

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

then

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

So you make the mean zero and the standard deviation 1.

Reminder: standard normal quantiles

Reminder: standard normal quantiles

coverage	α	$1-\alpha/2$	$z^\star_{1-\alpha/2}$
80%	0.2	0.9	≈ 1.28
90%	0.1	0.95	≈ 1.64
95%	0.05	0.975	≈ 1.96
99%	0.01	0.995	≈ 2.58

Deriving the confidence intervals

Since

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

it must be that

$$\mathrm{Prob}\left(-z^{\star}_{1-\frac{\alpha}{2}} \quad < \quad \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad < \quad z^{\star}_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Now let's eat some spinach...

Power through, people

(**Goal**: get μ alone by itself in the middle.)

This

$$-z^{\star}_{1-\frac{\alpha}{2}} \quad < \quad \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad < \quad z^{\star}_{1-\frac{\alpha}{2}}$$

implies this:

$$-z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad < \quad \bar{x}-\mu \quad < \quad z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

We multiplied by the positive number $\sigma/\sqrt{n}$ everywhere.

It's almost over

(**Goal**: get μ alone by itself in the middle.)

This

$$-z^{\star}_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad < \quad \bar{x} - \mu \quad < \quad z^{\star}_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

implies this:

$$-\bar{x} - z^{\star}_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad < \quad -\mu \quad < \quad -\bar{x} + z^{\star}_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

We subtracted $\bar{x}$ everywhere.

Remember, it wasn't on the exam

(**Goal**: get μ alone by itself in the middle.)

This

$$-\bar{x} - z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad < \quad -\mu \quad < \quad -\bar{x} + z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

implies this:

$$\bar{x} + z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad > \quad \mu \quad > \quad \bar{x} - z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

We multiplied by -1 everywhere, which means we had to flip the direction of all the inequalities.

Done!

Because

$$\text{Prob}\left(-z^{\star}_{1-\frac{\alpha}{2}} \quad < \quad \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad < \quad z^{\star}_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

it must be that

$$\text{Prob}\left(\bar{x} - z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad < \quad \mu \quad < \quad \bar{x} + z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

So the interval $(L,\ U)$ with bounds

$$L = \bar{x} - z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$
$$U = \bar{x} + z^{\star}_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

is an *exact* $100 \times (1-\alpha)\%$ confidence interval for μ.

But we don't know σ, so who cares?

- We have
$$\bar{x} \pm z^{\star}_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$
To make it operational, we blithely plug in $\hat{\sigma}$.

- The central limit theorem (and some other things...) gives us permission to do this when n is "big enough", but for small or medium n, all of the math we did is just plain wrong, and this interval will *under cover*:
$$\bar{x} \pm z^{\star}_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}.$$

How do we fix this?

Revisiting the standardized average

We started with this:

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

If you plug in $\hat{\sigma}$, it turns out that

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

Why?

- $\bar{x}$ and $\hat{\sigma}$ are both random (depend on the random sample);
- When you go from one source of randomness on the left to two, things are "more random," and the tails get heavier.

Finite-sample interval for mean of normal data

If you start from

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

the same steps from before give an *exact* $100 \times (1 - \alpha)\%$ confidence interval for the mean of normal data:

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

This does not assume you know σ, and it has correct coverage no matter what n is.

Mathematical facts you are asked to take on faith

1. Sampling distribution of sample average of normal data:

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

2. Sampling distribution of "realistically" standardized average:

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

3. The central limit theorem.

A sliver of STA 240: using calculus to prove these things for real.

Whence the t-test?

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

Point estimate: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Hypotheses:

$$\begin{aligned} \mathrm{H}_0 &: \mu = \mu_0 \\ \mathrm{H}_A &: \mu \neq \mu_0. \end{aligned}$$

Null distribution: if the null is true, we know that

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

Instead of simulations and histograms and approximations, just use Student's t as the null distribution. Compute p-value based on that and proceed business-as-usual.

One-sample t-test

1. Collect data set of size n;
2. Compute $\bar{x}$, $\hat{\sigma}$, and **test statistic**

$$\frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

3. Locate the observed statistic under the t_{n-1} curve:
4. Compute p-value and decide:
   - if p-value $< \alpha$, reject null;
   - if p-value $\geq \alpha$, fail to reject null;