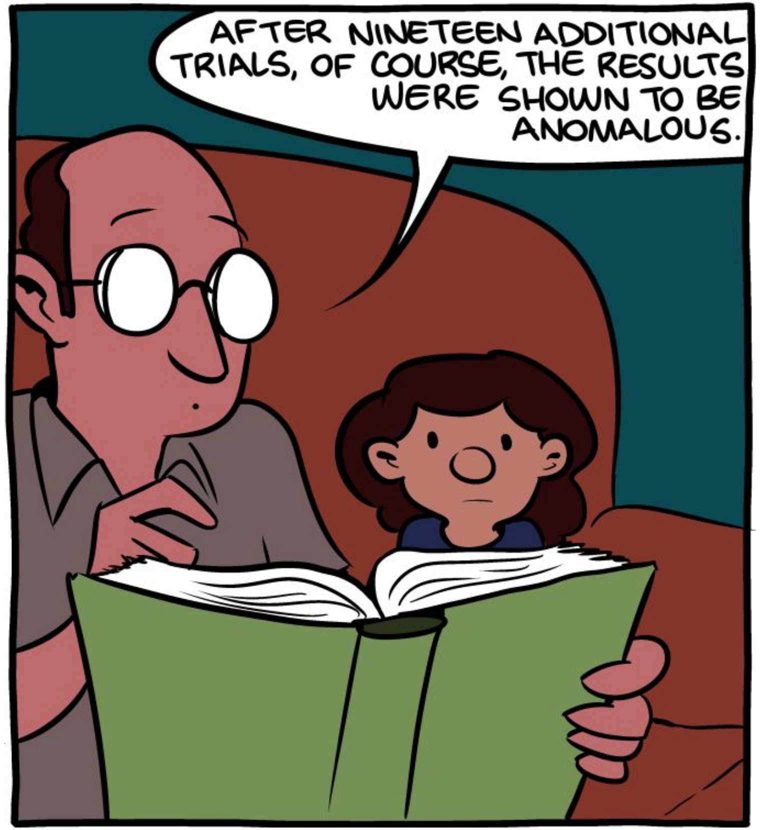
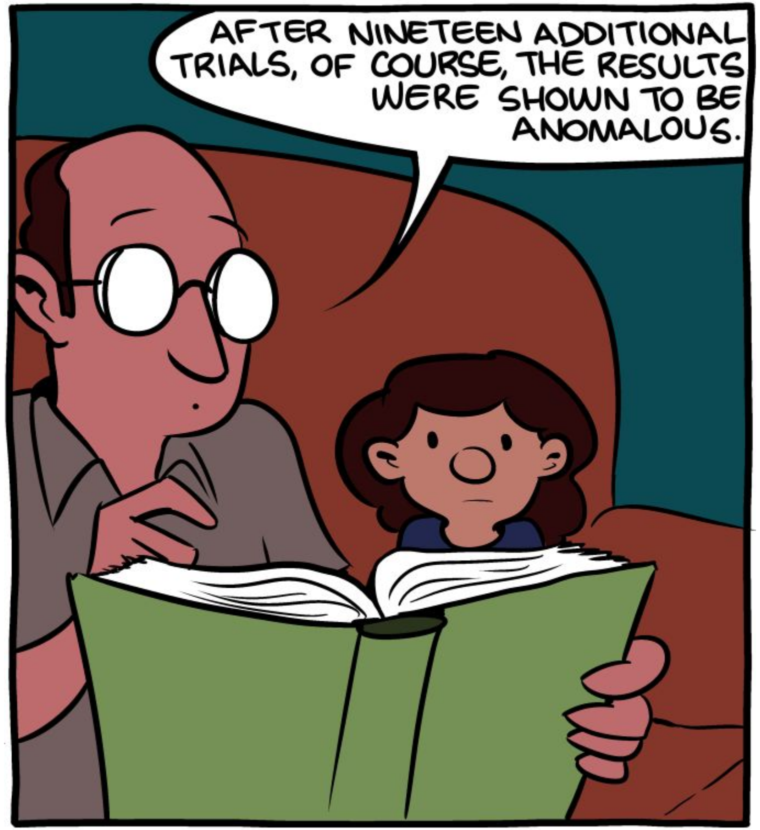
Lecture 18: Statistical Inference with Small Samples



"The Tortoise And The Hare" is actually a fable about small sample sizes.

Announcements and reminders

- HW 5 due **Friday 9 November at 5 PM**
- Check in on Arkaive!



"The Tortoise And The Hare" is actually a fable about small sample sizes.

Previously, on CSCI 3022…

Statistical inference for population mean when **data are normal** and $\boxed{\textbf{n is large}}$ and…

$$\boxed{E[\bar{X}] \;\&\; Var(\bar{X})}$$

σ is known: $$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

σ is unknown: $$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Previously, on CSCI 3022…

$(n \geq 30)$

Statistical inference for population mean when **data are NOT normal** and **n is large** and…

σ is known:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{CLT}{\sim} N(0,1)$$

σ is unknown:

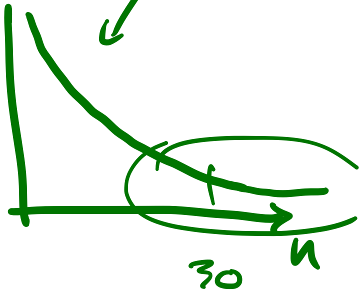
$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \overset{CLT}{\sim} N(0,1)$$

Previously, on CSCI 3022…

$n < 30$

Statistical inference for population mean when **data are normal** and **n is small** and…

$\frac{1}{n-1}$



σ is known: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

σ is unknown: ???

ϵ_i Normally ~~distributed~~

w/ $n < 30$

tempting… $\frac{\bar{X}-\mu}{s/\sqrt{n}} \overset{?}{\sim} N(0,1)$

$\sqrt{\frac{1}{n-1}\sum (x_i-\bar{x})^2}$

The story so far for Means

Thus far, we've talked about Hypothesis Testing / Confidence Intervals for the mean of a population in the following cases

	n ≥ 30	n < 30
Normal data, known σ ✓		
Normal data, unknown σ		
Non-normal data, known σ		
Non-normal data, unknown σ		

z-distribution (past)

t-distribution (today!)

Bootstrap (later)

Small-sample tests for μ

$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

est. of μ that we "bought" w/ 1 df

When n is small, we can't invoke the Central Limit Theorem

- If we don't even know if the data are Normal, then we can **bootstrap**
- But that can be expensive (producing lots of replicates takes **time** and **memory**)

If we have **small n** and **some reason to think our data are (approximately) Normal**, then…

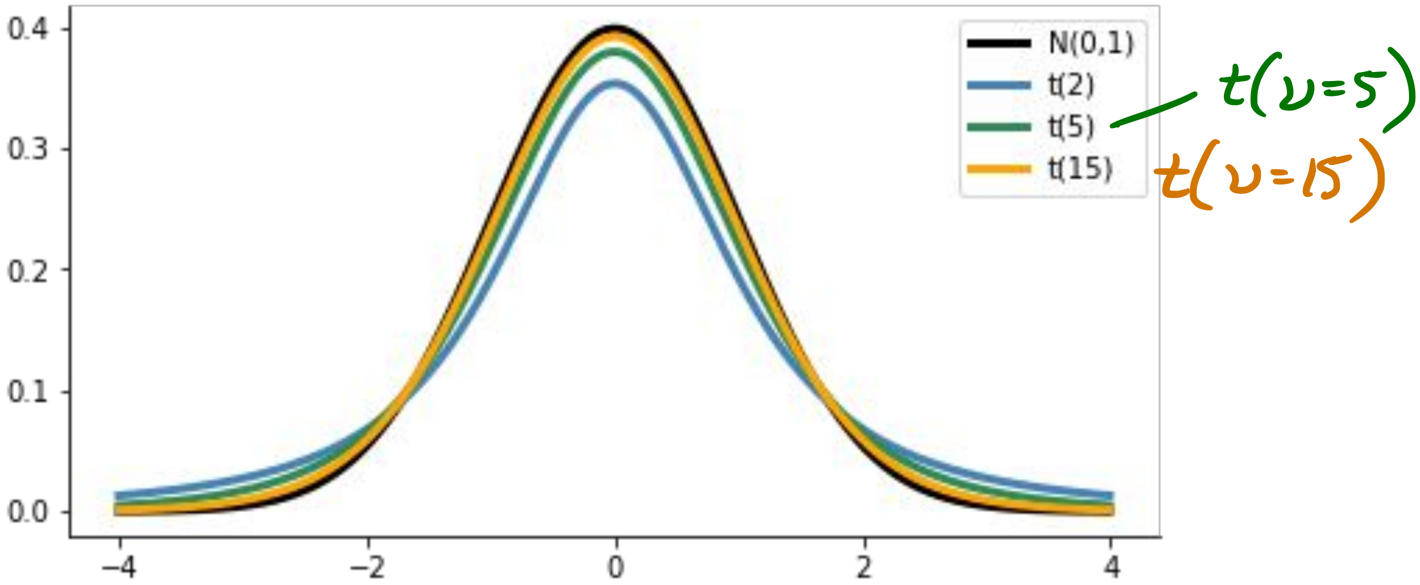
When $\bar{X}$ is the sample mean of a random sample of size n from a normal distribution with mean μ, the random variable

Test statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_\nu$

follows a probability distribution called a **t-distribution** with parameter $\boxed{\nu = \text{n-1}}$ degrees of freedom (df)

The t-distribution

Here are some members of the family of t-distributions, and the standard normal N(0,1)

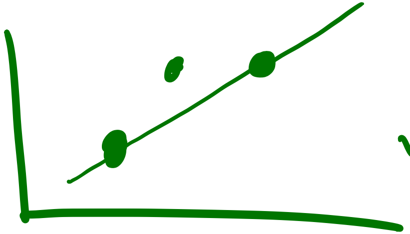


Properties of t-distributions

n = # data points (samples)

Let t_ν denote the t-distribution with parameter $\boxed{\nu = n - 1}$ df

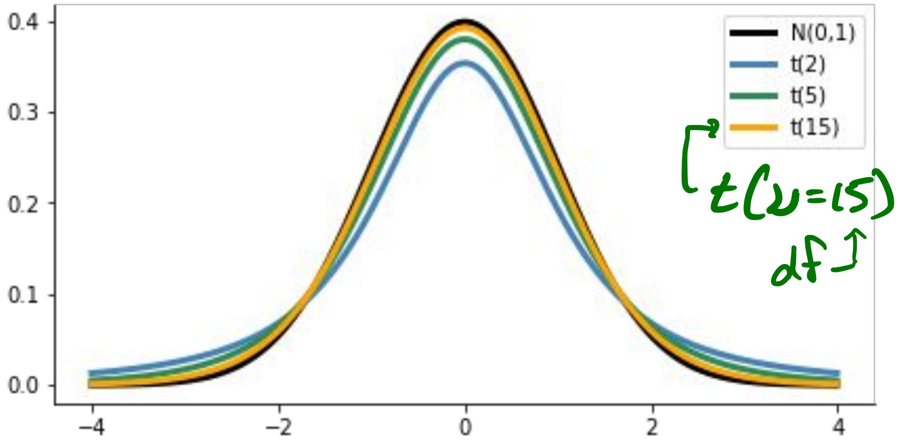
- Each t_ν curve is bell-shaped and centered at 0 & symmetric
- Each t_ν curve is more spread out than the standard normal distribution
- As ν increases, the spread of the corresponding t_ν curve decreases $\lim_{\nu \to \infty} t_\nu$ "=" $N(0,1)$
- As $\nu \to \infty$ the sequence of t_ν curves approaches the standard normal curve



df=2

$y = mx + b$

$y = ax^2 + bx + c$



$t(\nu = 15)$

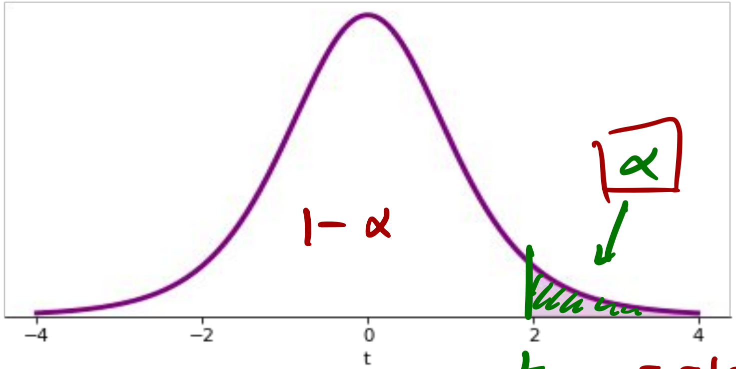
df ↑

The t-critical value

\nu

We can extend all of our inferential mechanics to the small-sample case by introducing the so-called t-critical value, which we denote as $t_{\alpha,\nu}$

Definition: The t-critical value, $t_{\alpha,\nu}$, is the point such that the area under the t_ν -curve to the **right** of $t_{\alpha,\nu}$ is equal to



ddof

$t_{\alpha,\nu}$ = stats.t.ppf(1−α, ν=n−1)

Example: $t_{0.05,\,6}$ is the t-critical value that captures the upper-tail area of 0.05 (5%) under the t-curve with 6 degrees of freedom.

$\nu = 6 = n-1$

→ Sample size = 7

The t-confidence interval for the mean

$$\left[\bar{x} - \boxed{z_{\alpha/2}} \cdot \frac{s}{\sqrt{n}}, \ \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] \leftarrow z\text{-dist.}$$

Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from a random sample of size n, from a normal population with mean μ.

Then a 100·(1-α)% t-confidence interval for the mean μ is given by:

$$\left[\bar{x} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}, \ \bar{x} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}\right]$$

Or, more compactly:

$$\bar{x} \pm t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$$

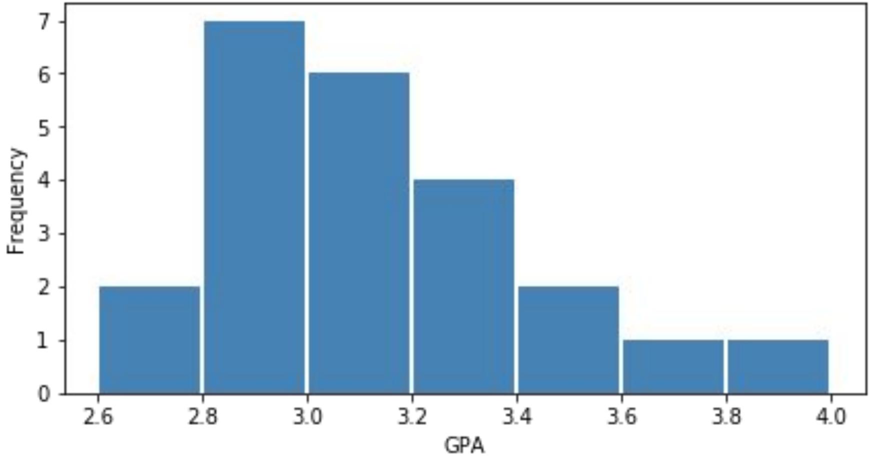
$\text{stats.t.ppf}(0.95, 22) = 1.717$
$[3.036, 3.256]$

The t-confidence interval for the mean

$\nu = 23 - 1 = 22$

Example: S'pose the GPAs for 23 students have the histogram shown here. The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Find a 90% CI for the mean GPA.

$\rightarrow \alpha = 0.1$

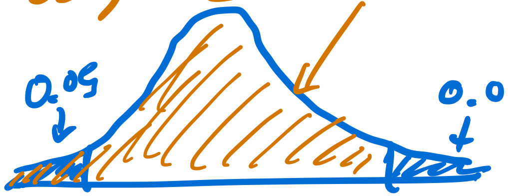


90% CI: $\bar{x} \pm t_{0.05,22} \cdot \frac{0.308}{\sqrt{23}}$

stats.t.ppf(0.95, ddof=22) = 1.717

$= 3.146 \pm 1.717 \cdot \frac{0.308}{\sqrt{23}}$

$= \boxed{[3.036, 3.256]}$



$t_{\alpha/2, n-1} = t_{0.05,22}$

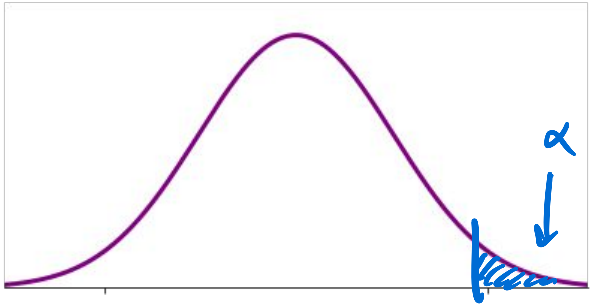
The t-test, critical regions and p-values

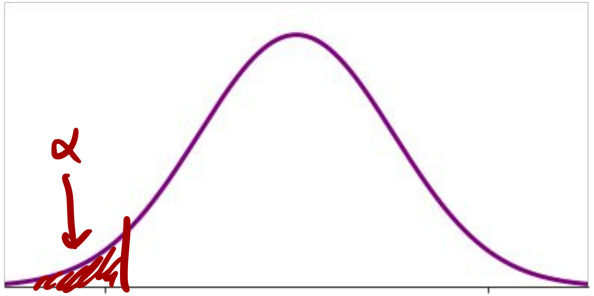
$$\frac{\bar{X}-\theta_0}{s/\sqrt{n}}$$

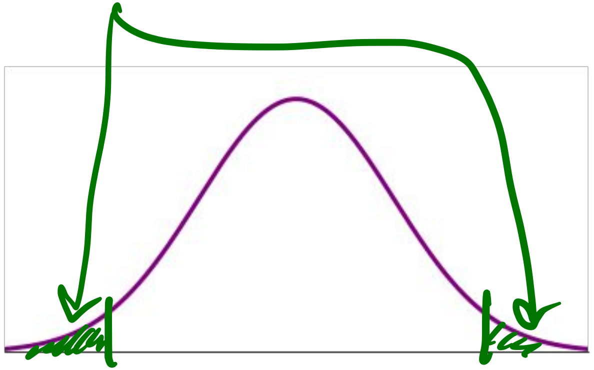
$P(\tau > t)$

$H_0: \theta = \theta_0$

Alternative hypothesis	Critical region level α test	p-value level α test
$H_1: \theta > \theta_0$	$t \geq t_{\alpha,\nu}$	p-value = 1 − t.cdf(t, ddof=n-1)
$H_1: \theta < \theta_0$	$t \leq t_{1-\alpha,\nu}$	p-value = t.cdf(t, ddof=n-1)
$H_1: \theta \neq \theta_0$	$(t \geq t_{\alpha/2,\nu})$ or $(t \leq -t_{\alpha/2,\nu})$	p-value = 2 · t.cdf(−|t|, ddof=n-1)

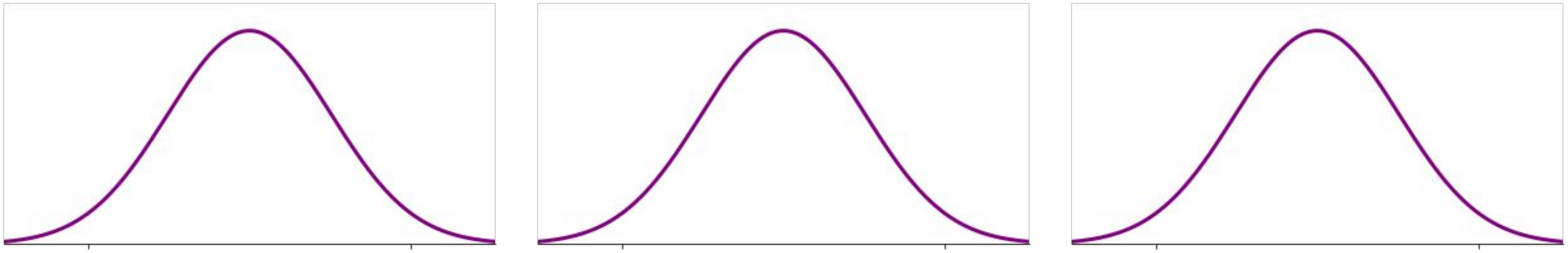






The t-test, critical regions and p-values

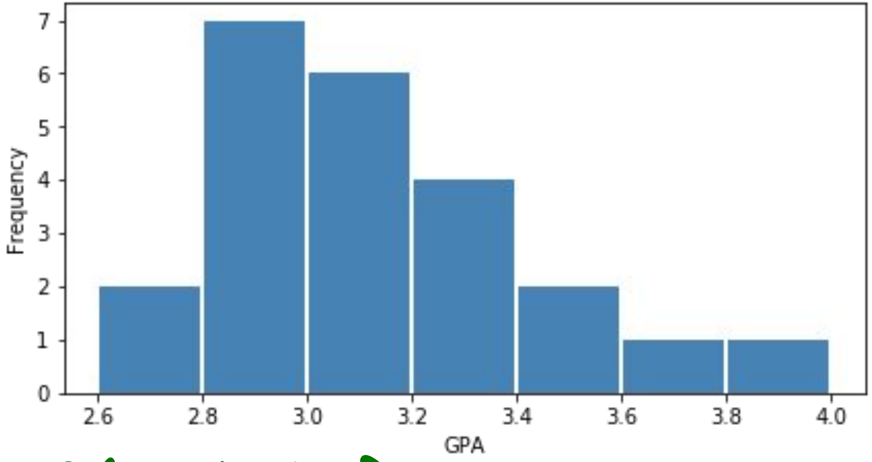
Alternative hypothesis	Critical region level α test	p-value level α test
H_1: $\theta > \theta_0$	$t \geq t_{\alpha,\nu}$	$P(T \geq t \mid H_0) \leq \alpha$
H_1: $\theta < \theta_0$	$t \leq t_{\alpha,\nu}$	$P(T \leq t \mid H_0) \leq \alpha$
H_1: $\theta \neq \theta_0$	$(\, t \geq t_{\alpha/2,\nu} \,)$ or $(\, t \leq -t_{\alpha/2,\nu} \,)$	$2 \cdot \min\{P(T \leq t \mid H_0), P(T \geq t \mid H_0)\} \leq \alpha$



Stats.t.cdf(-2.398, ν=22) = 0.0127

t-test for the mean, using p-values

Example: S'pose the GPAs for 23 students have the histogram shown here. The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Determine if there is sufficient evidence to conclude at the 0.10 (10%) significance level that the mean GPA is not equal to 3.30.

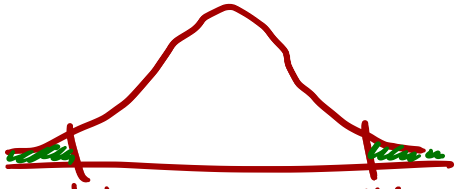


$\alpha = 0.1$

$H_0: \mu = 3.30$

$H_1: \mu \neq 3.30$

Two tailed



p-value = 2 · t.cdf(-|t|, ddof=22) = 2 · 0.0127

= 0.0254 < 0.1

Test statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3.146 - 3.30}{0.308/\sqrt{23}} = -2.398$$

REJECT H_0

Inference for variances

We’ve talked about confidence intervals for the **mean** and for **proportions**

Question: What does the sampling distribution of the **variance** look like when the population is **normally distributed**?

… if your population is **normally distributed**, it turns out we have some theory that gives us a **confidence interval** and works for both large ***and*** small samples!

Inference for variances

We've talked about confidence intervals for the **mean** and for **proportions**

Question: What does the sampling distribution of the **variance** look like when the population is **normally distributed**?

… if your population is **normally distributed**, it turns out we have some theory that gives us a **confidence interval** and works for both large ***and*** small samples!

Inference for variances

We’ve talked about confidence intervals for the **mean** and for **proportions**

Question: What does the sampling distribution of the **variance** look like when the population is **normally distributed**?

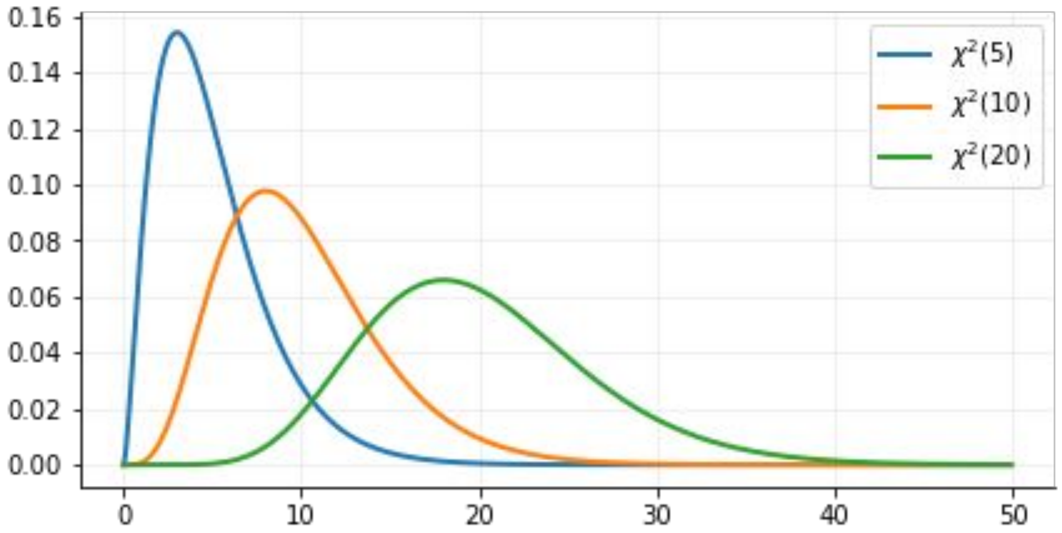
… if your population is **normally distributed**, it turns out we have some theory that gives us a **confidence interval** and works for both large ***and*** small samples!

The chi-squared distribution (χ^2 distribution)

\chi^2

The chi-squared (χ_ν^2) distribution is also parameterized by degrees of freedom ν = n-1

The pdfs of the family χ_ν^2 are pretty nasty, so let's just plot a few.



A confidence interval for the variance

HERE 3p

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and standard deviation σ. Define the sample variance in the usual way as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Then the random variable $\dfrac{(n-1)\,S^2}{\sigma^2}$ follows the distribution χ^2_{n-1}

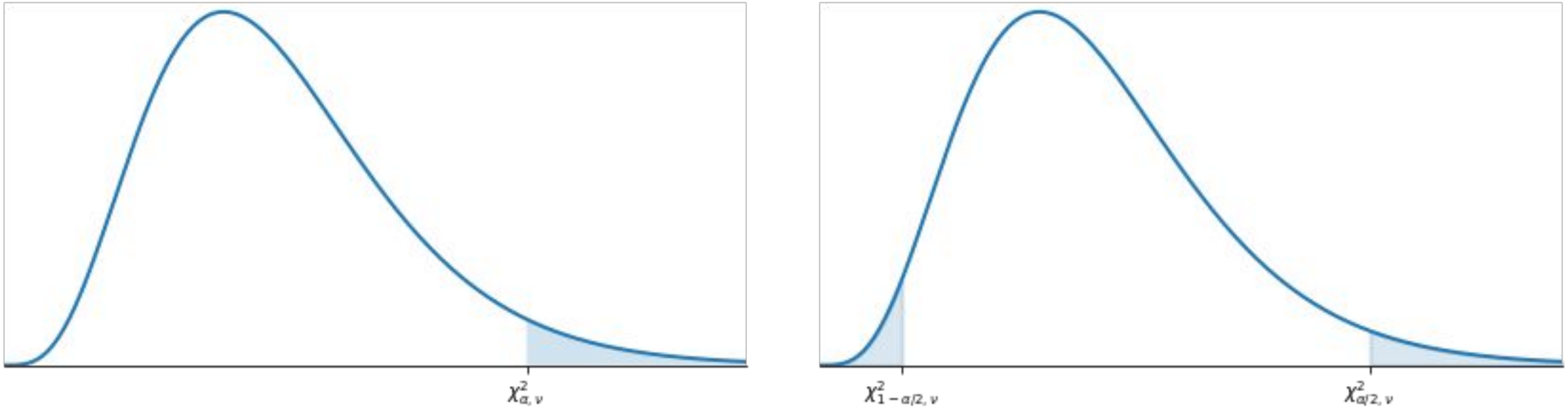
Then it follows that …

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

$$\text{ε}\quad P\left(\chi^2_{1-\alpha/2,\,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,\,n-1}\right) = 1-\alpha$$

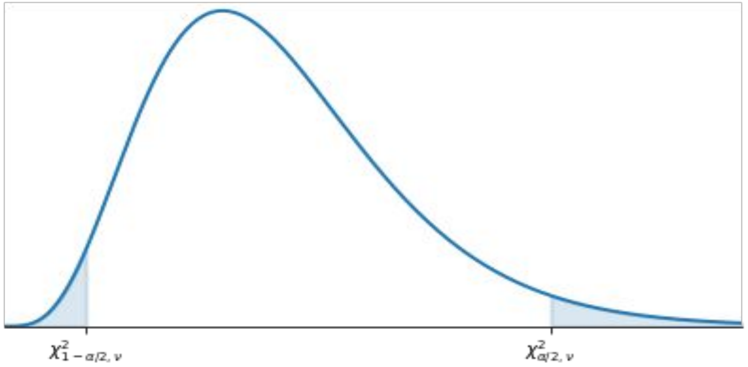
A confidence interval for the variance

Because the χ^2 distribution is not symmetric, we need to use two different critical values



A confidence interval for the variance

For a 100·(1-α)% CI, we choose the **two** critical values $\chi^2_{1-\alpha/2,\ n-1}$ and $\chi^2_{\alpha/2,\ n-1}$, which attributes α/2 probability to each the left and right tails. Then, with 100·(1-α)% confidence we can say that



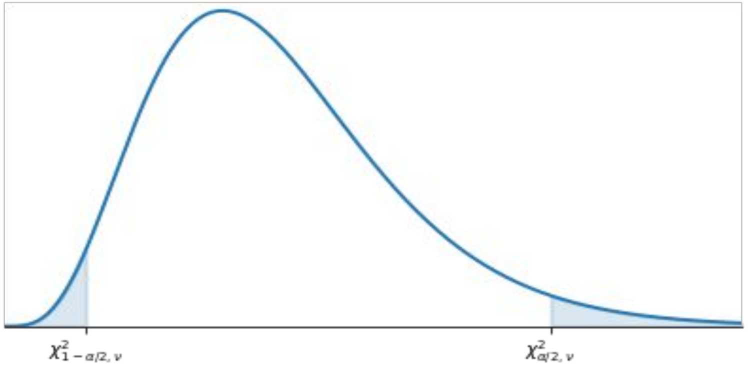
A confidence interval for the variance

HERE

For a 100·(1-α)% CI, we choose the **two** critical values $\chi^2_{1-\alpha/2,\ n-1}$ and $\chi^2_{\alpha/2,\ n-1}$, which attributes α/2 probability to each the left and right tails. Then, with 100·(1-α)% confidence we can say that

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\ n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\ n-1}}$$

Question: What, then, is a 100·(1-α)% CI for the SD?



A confidence interval for the variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52 g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation in the produced packages is larger than acceptable. In an attempt to estimate the variance he selects n=10 bags at random and weighs them. The sample yields a sample variance of 4.2 g^2. Find a 95% CI for the variance, and a 95% CI for the SD.

A confidence interval for the variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52 g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation in the produced packages is larger than acceptable. In an attempt to estimate the variance he selects n=10 bags at random and weighs them. The sample yields a sample variance of 4.2 g^2. Find a 95% CI for the variance, and a 95% CI for the SD.

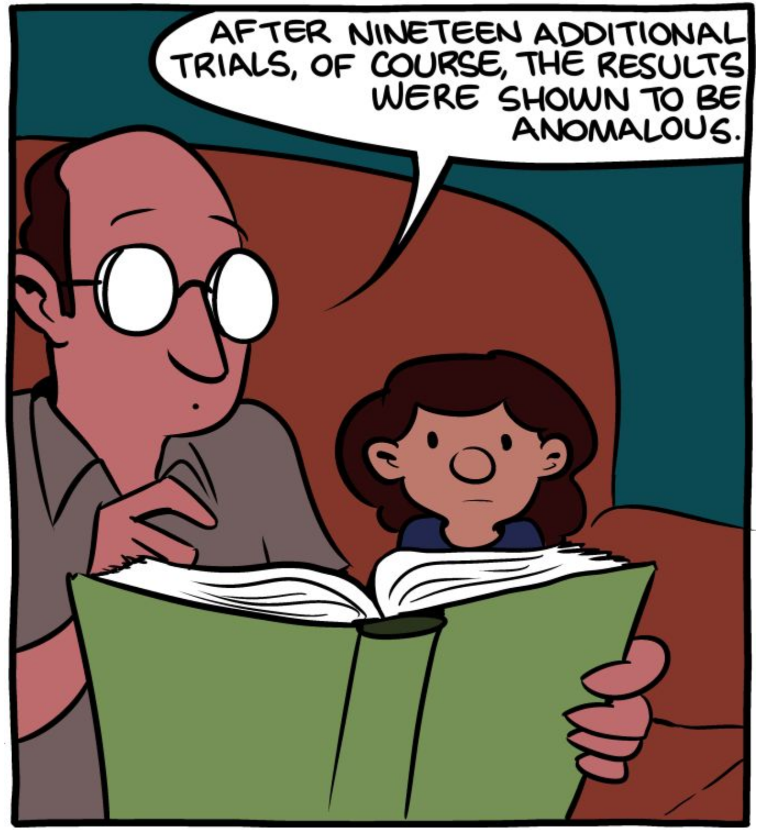
$$\alpha = 0.05, \quad \alpha/2 = 0.025, \quad n = 10, \quad s^2 = 4.2$$

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,9} = \text{stats.chi2.ppf}(0.025,\ 9) = 2.70$$
$$\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,9} = \text{stats.chi2.ppf}(0.975,\ 9) = 19.02$$

$$\frac{(10-1)\cdot 4.2}{19.02} < \sigma^2 < \frac{(10-1)\cdot 4.2}{2.70}$$
$$\Rightarrow \quad 1.99 < \sigma^2 < 14.0$$

What just happened?

- **Small samples** happened!
  - Learned what distributions (instead of standard normal) to use when our sample is too small for CLT to kick in
- **T-distributions** -- small sample CI/hypothesis testing for the **mean**
- **chi-squared distributions** -- small sample CI/hypothesis testing for the **variance**



"The Tortoise And The Hare" is actually a fable about small sample sizes.