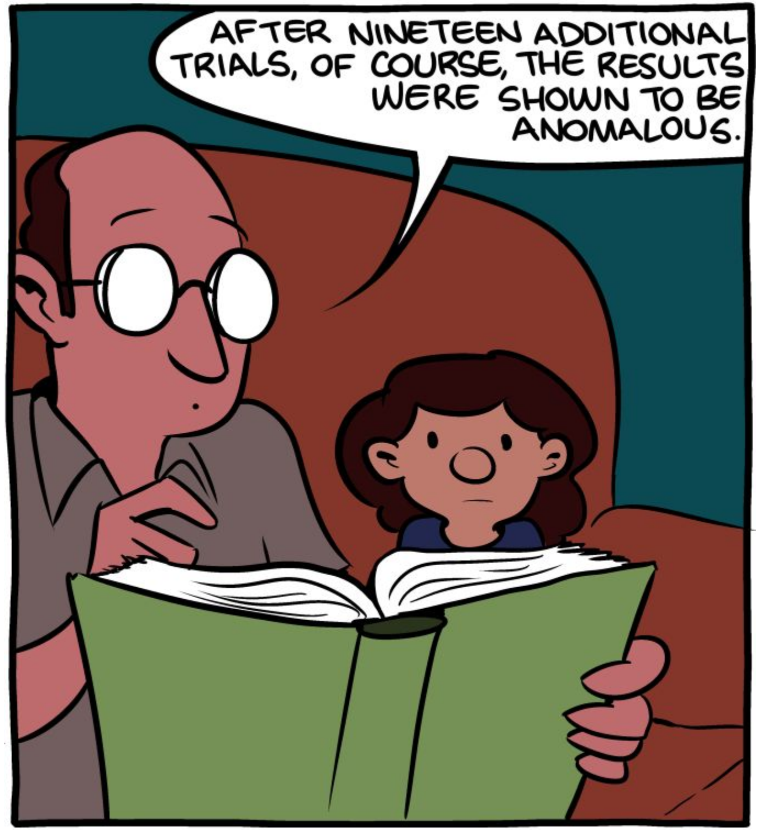
Lecture 18: Statistical Inference with Small Samples

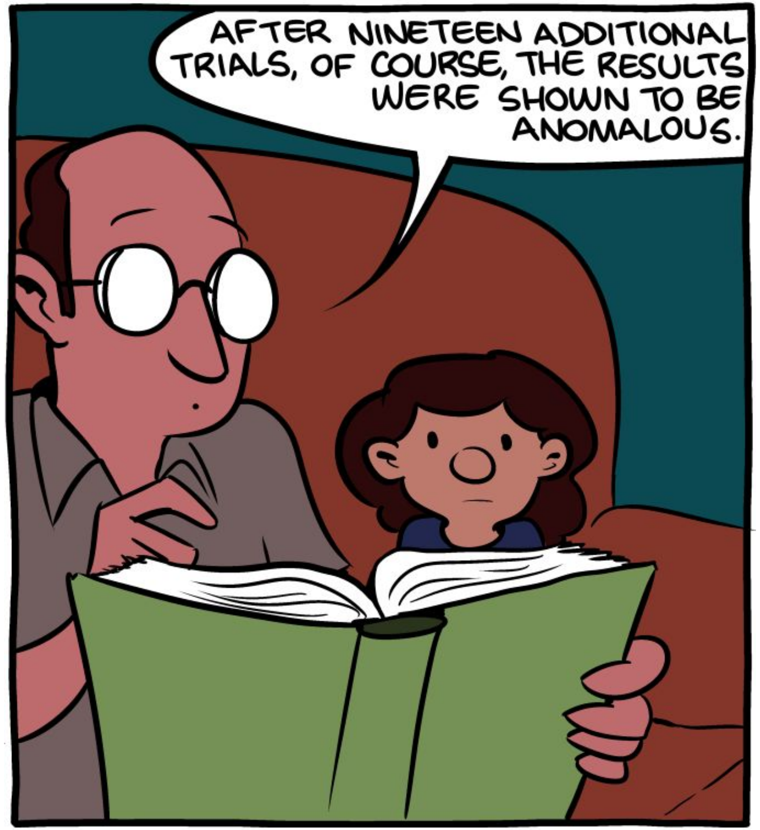


"The Tortoise And The Hare" is actually a fable about small sample sizes.

Announcements and reminders

- HW 4 posted!
  Due Friday 5 April at 5 PM

- Quizlet 8 due Friday 22 March at 10 AM



"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 **n is large** and…

σ is known:

σ is unknown:

Previously, on CSCI 3022…

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

σ is known:

σ is unknown:

Previously, on CSCI 3022…

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

σ is known:

σ is unknown:

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		

Small-sample tests for μ

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

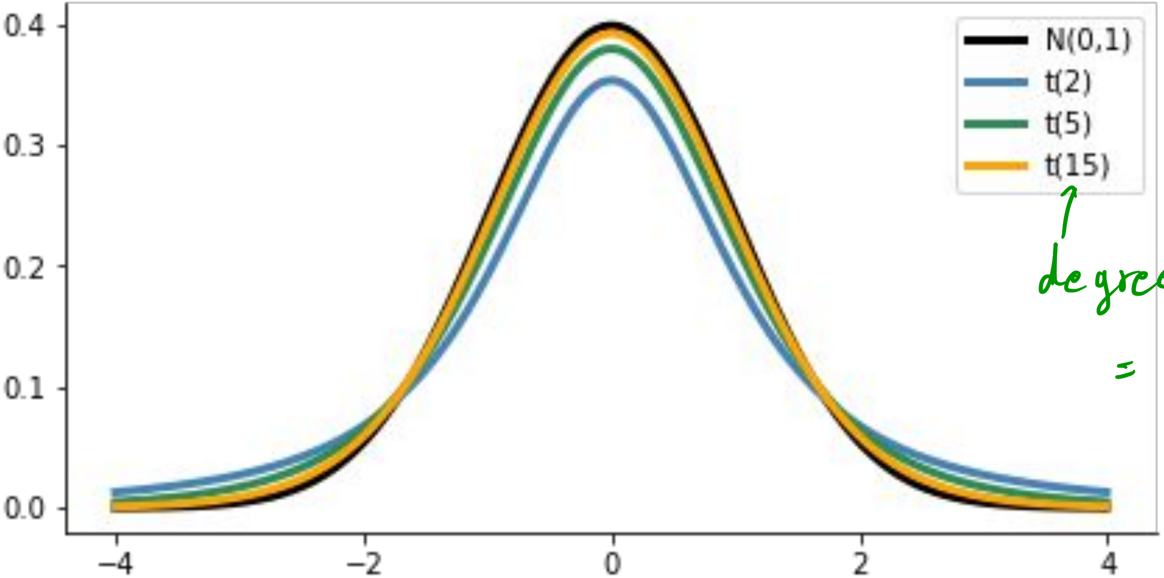
σ unknown

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

follows a probability distribution called a <u>**t-distribution**</u> with parameter ν = 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)



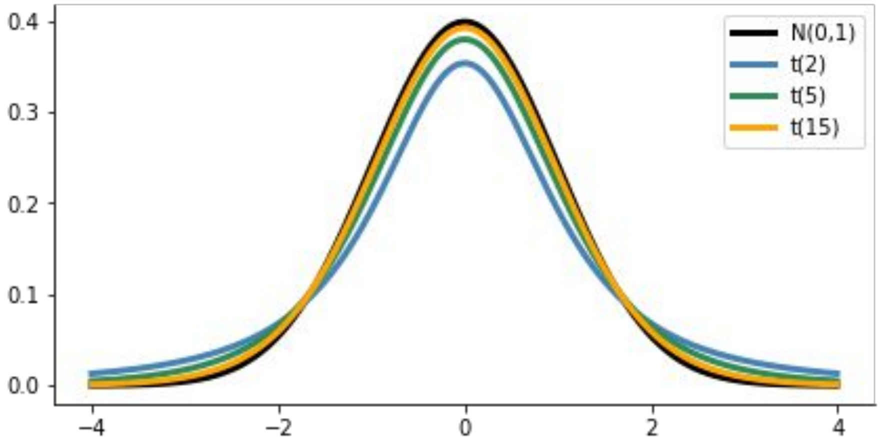
degree of freedom
= # data − 1

Properties of t-distributions

\nu ← latex

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

- Each t_ν curve is bell-shaped and centered at 0
- 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
- As $\nu \to \infty$ the sequence of t_ν curves approaches the standard normal curve



The t-critical value

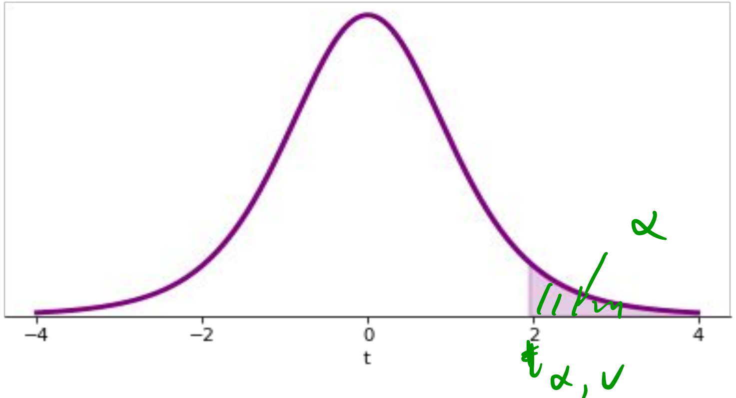
– small sample and unknown standard deviation
if we want to use t-critical value

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

ν degrees of freedom
= # data − 1

$t_{.05,6}$



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.

→ Sample size = 7 (degree of freedom + 1)

6 + 1 = 7

The t-confidence interval for the mean

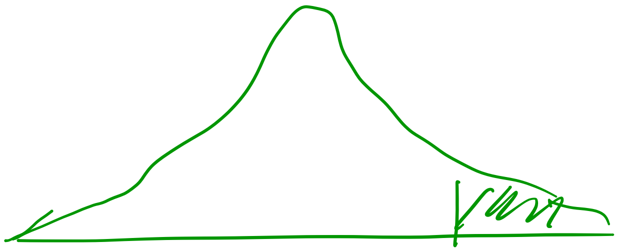
Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from a random sample of size n, from a normal population with mean μ.

for unknown popular mean

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

Z $CI : [\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + \ldots]$

t $CI : [\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}]$



$t_{\frac{\alpha}{2}, n-1}$

Or, more compactly:

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

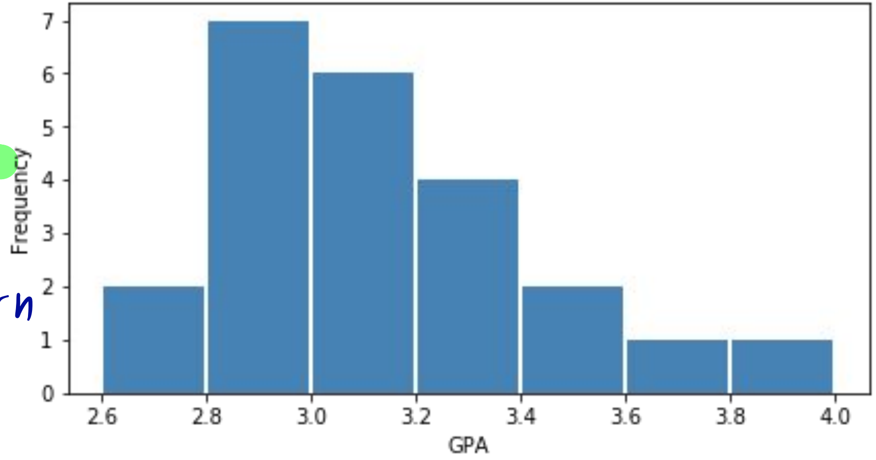
critical value z → stats.t.ppf$(1 - \frac{\alpha}{2}$, df = n-1)

The t-confidence interval for the mean

→ small sample

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.

unknown σ



Need to also assume: data are approximately normal.

$$\bar{x} \pm t_{0.05,22} \cdot \frac{s}{\sqrt{n}} = 3.145 \pm t_{0.05,22} \frac{0.308}{\sqrt{23}}$$

stats.t.ppf(.95, df=22)

The t-test, critical regions and p-values

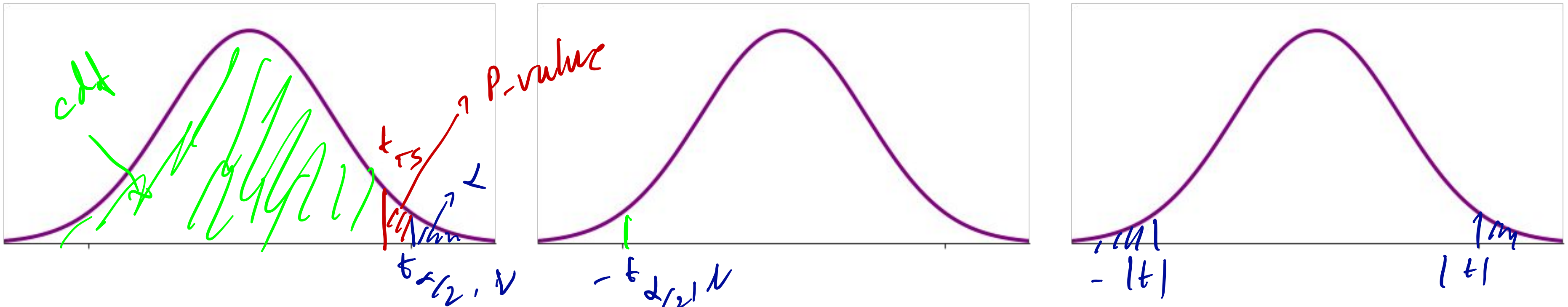
$$t_{.05} = \frac{\theta - \theta_0}{\text{STERR}(\theta)} \rightarrow t_{TS} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t_{TS} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$H_0 : \theta = \theta_0$ (eg H_0: $p = 0.5$)

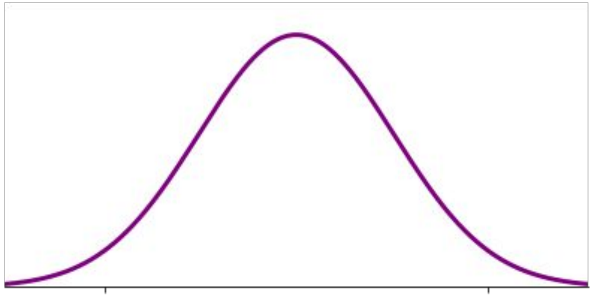
Alternative hypothesis	Critical region level α test	p-value level α test
H_1: $\theta > \theta_0$	$t \geq t_{\alpha,\nu}$	p-value $= 1 - \text{cdf}(t_{TS}, \nu)$
H_1: $\theta < \theta_0$	$t \leq t_{\alpha,\nu}$	p-value $= \text{cdf}(t_{TS}, \nu)$
H_1: $\theta \neq \theta_0$	$(t \geq t_{\alpha/2,\nu})$ or $(t \leq -t_{\alpha/2,\nu})$	p-value $= 2 \cdot \text{cdf}(-\lvert t_{TS} \rvert, \nu)$

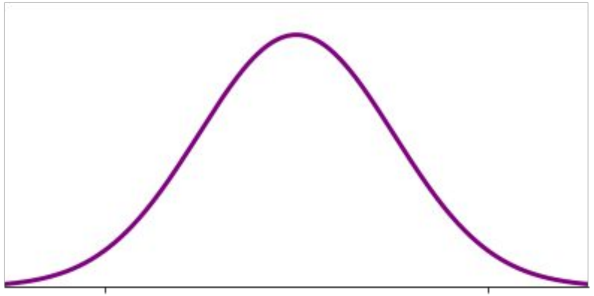
p-value $= 1 - \text{cdf}$

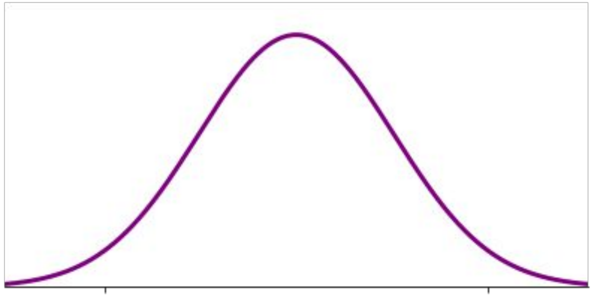


The t-test, critical regions and p-values

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



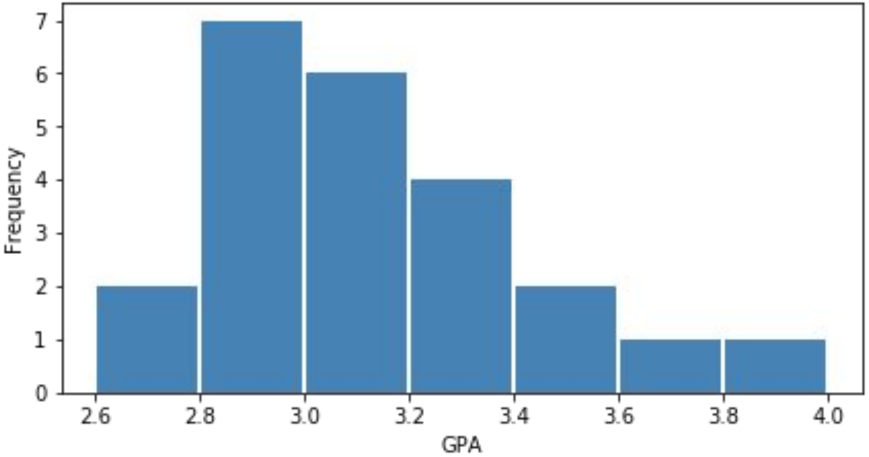




t-test for the mean, using p-values

1) write hypothesis
2) compute a test statistic

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.



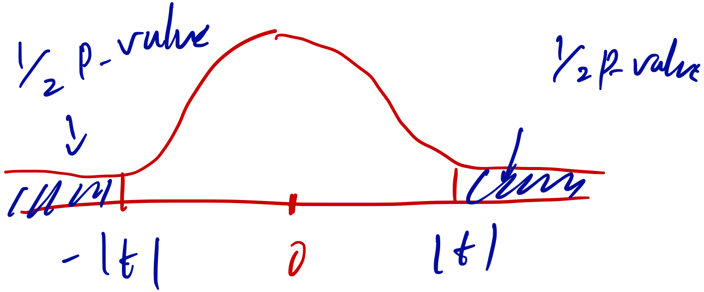
$H_0: \mu = 3.30$

$H_1: \mu \neq 3.30$

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

?

p-value = 2.stats.t.cdf(t, df = 22) $\boxed{< .1}$



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

never have negative b/c it's related to variance

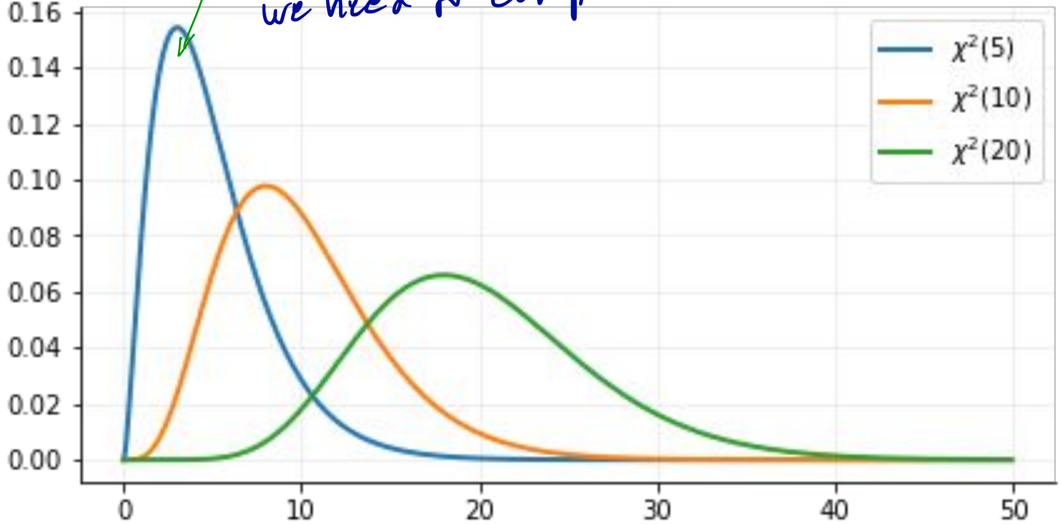
The chi-squared (χ_ν^2) distribution is also parameterized by degrees of freedom ν = n-1

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

$Var = \sum^{n} (dev)^2$

symmetric? No b/c one might go to infinite

we need to compute both critical value



A confidence interval for the variance

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 $\underbrace{\dfrac{(n-1)\,S^2}{\sigma^2}}_{\sim\,\chi^2_{n-1}}$ follows the distribution χ^2_{n-1}

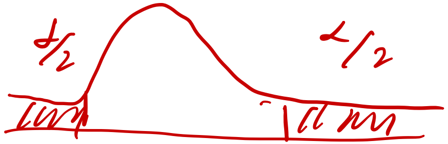
$$\left\{ \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \right.$$

Then it follows that …

$\chi^2_{\alpha/2,\,n-1}$ = value we plug into χ^2_{n-1} dist. to put $\frac{\alpha}{2}$ prob. on upper tail

$\text{cdf}(\chi^2_{\alpha/2,\,n-1}) = 1 - \frac{\alpha}{2}$

↳ $\text{ppf}\left(1 - \frac{\alpha}{2}\right) = \chi^2_{\alpha/2,\,n-1}$



A confidence interval for the variance

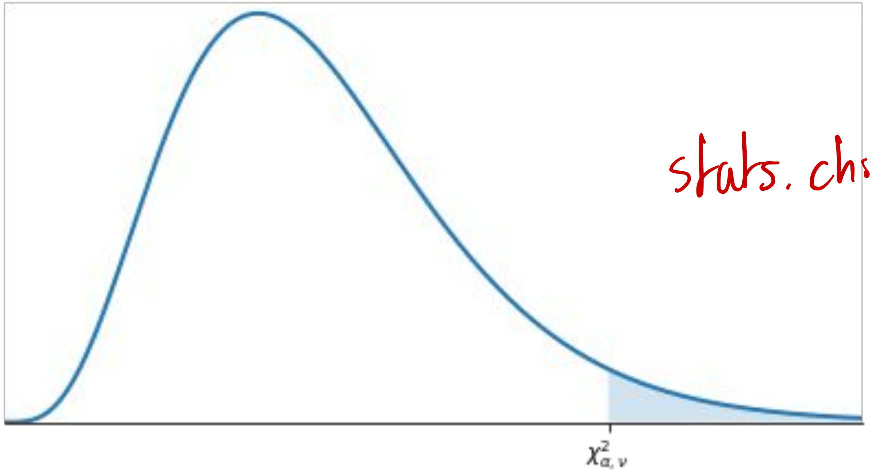
$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_\nu$$

→ is chi-sq var.

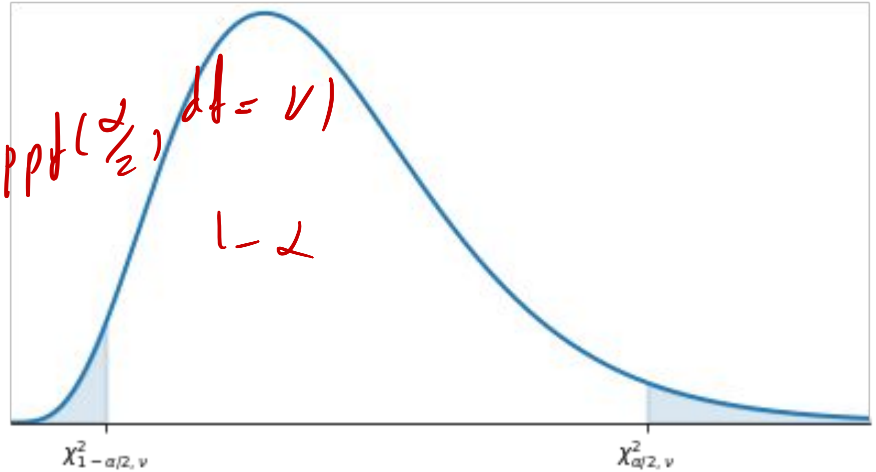
Because the χ^2 distribution is not symmetric, we need to use two different critical values

$$P\left(\chi^2_{1-\alpha/2,\nu} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2,\nu}\right) = 1-\alpha$$

CI for σ^2: $\_\_\_\_ \le \sigma^2 \le \_\_\_\_$



stats.chi2.ppf($\alpha/2$, df = ν)



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

"lower bound": $\chi^2_{1-\alpha/2, \nu} \le \dfrac{(n-1)s^2}{\sigma^2}$

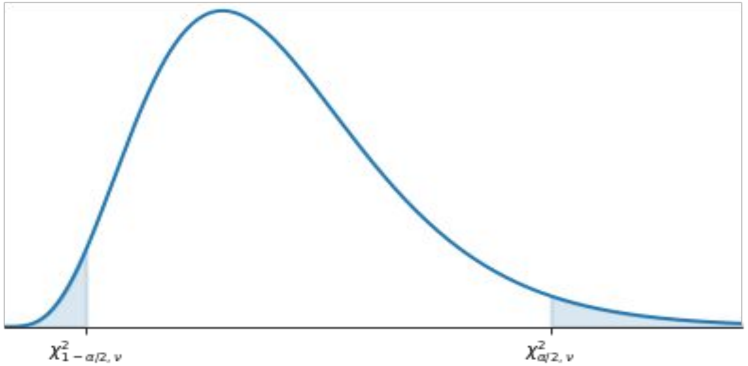
$\sigma^2 \le \dfrac{(n-1)s^2}{\chi^2_{1-\alpha/2,\nu}}$

"upper side": $\dfrac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2,\nu}$

$\dfrac{(n-1)s^2}{\chi^2_{\alpha/2,\nu}} \le \sigma^2$

100. (1-α)% CI for σ^2:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,\nu}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}$$



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

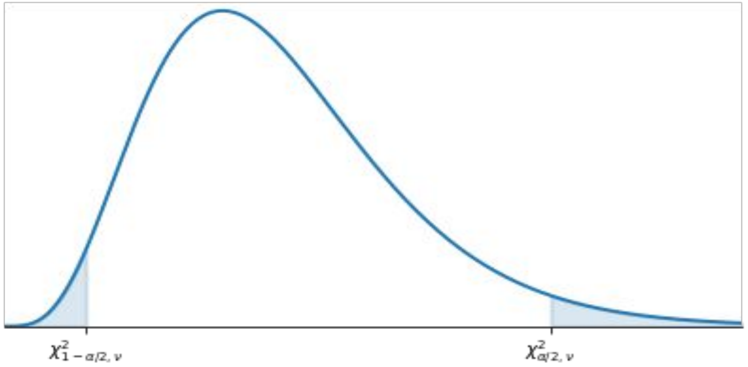
$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\ n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\ n-1}}$$

variance

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

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,\ n-1}}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,\ n-}}}$$

SD



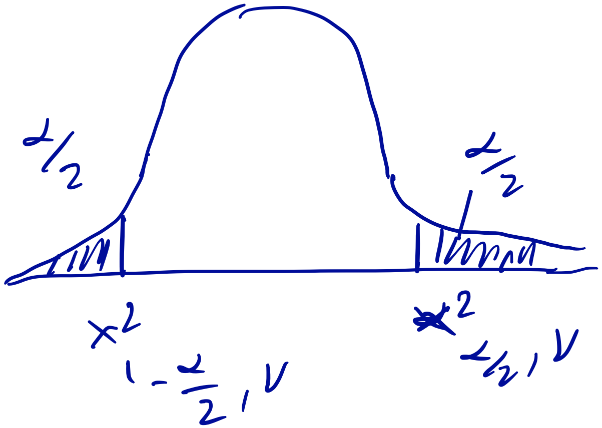
A confidence interval for the variance

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.

$s^2 = 4.2$

$n = 10$

$\hookrightarrow \alpha = 0.05 \rightarrow \frac{\alpha}{2} = 0.025$



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.

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

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

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

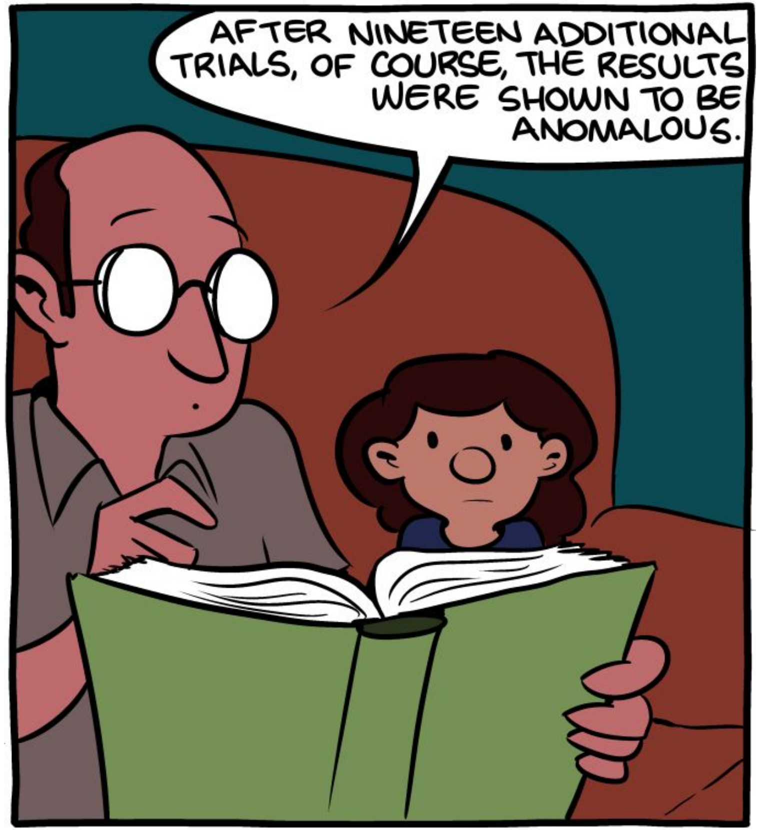
$$\frac{(10-1)\cdot 4.2}{19.02} < \sigma^2 < \frac{(10-1)\cdot 4.2}{2.70}$$

$$\Rightarrow \quad 1.99 < \sigma^2 < 14.0$$

$$\Rightarrow \sqrt{1.99} < \underbrace{\sigma}_{SD}{}^2 < \sqrt{14}$$

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.