Statistical Inference with Small Samples

Administrivia

- **Homework 5** due Friday Nov 10

Previously on CSCI 3022

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

σ is known: $\left(\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right) \sim N(0,1)$

Z-TEST

$\bar{x} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

σ is unknown: $\left(\dfrac{\bar{x}-\mu}{s/\sqrt{n}}\right) \sim N(0,1)$

Previously on CSCI 3022

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

σ is known: $\left(\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) \overset{CLT}{\sim} N(0,1)$

σ is unknown: $\left(\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\right) \overset{CLT}{\sim} N(0,1)$

Previously on CSCI 3022

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

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

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:

	$n \geq 30$	$n < 30$
Normal Data / Known σ		
Normal Data / Unknown σ		
Non-Normal Data / Known σ		
Non-Normal Data / Unknown σ		

Z-TEST BOOTSTRAP t-TEST

Small-Sample Tests for μ

- When n is small we cannot invoke the Central Limit Theorem

- When n is small and the variance is unknown we need to do something else …

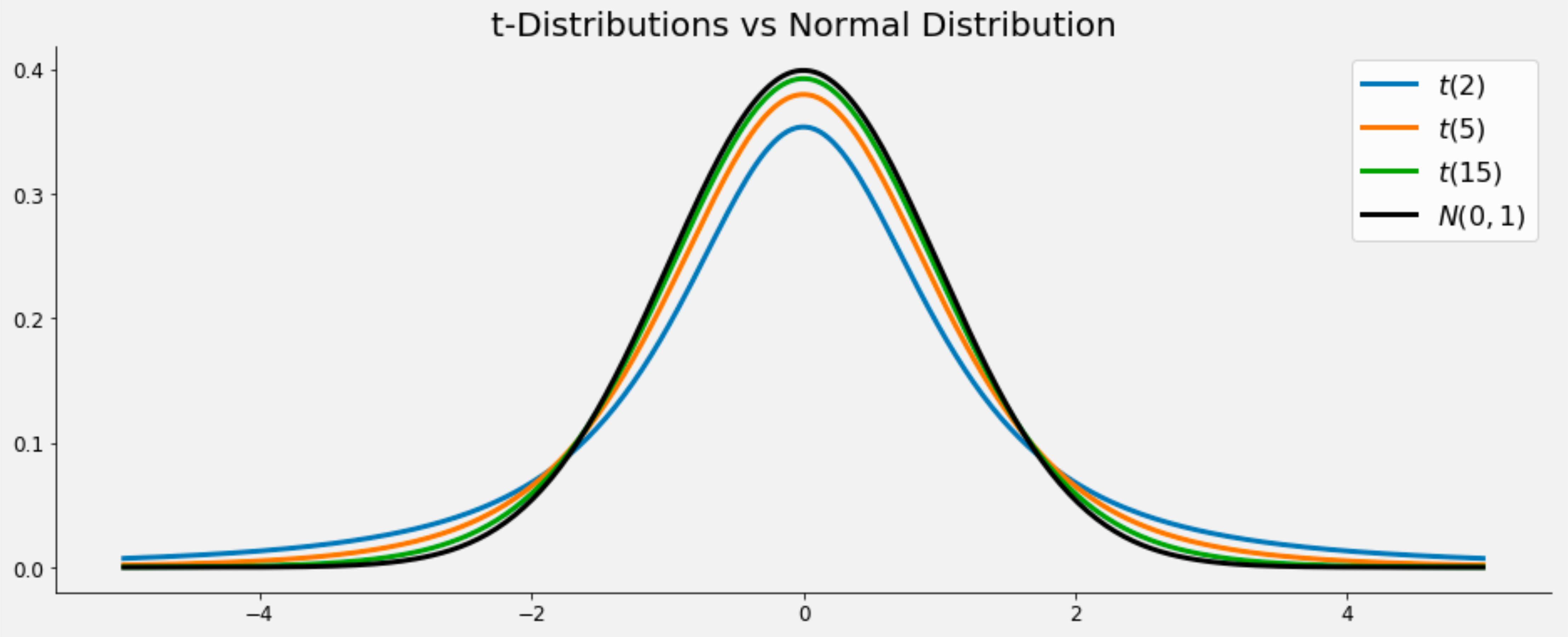
When $\bar{X}$ is the sample mean of a random sample of size n from a normal distribution with mean μ , the random variable

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

follows a probability distribution called a t-Distribution with parameter $\nu = n - 1$ degrees of freedom.

The t-Distribution

The following figure shows the pdf of some members of the family of t-Distributions



Properties of t-Distributions

$= n-1$

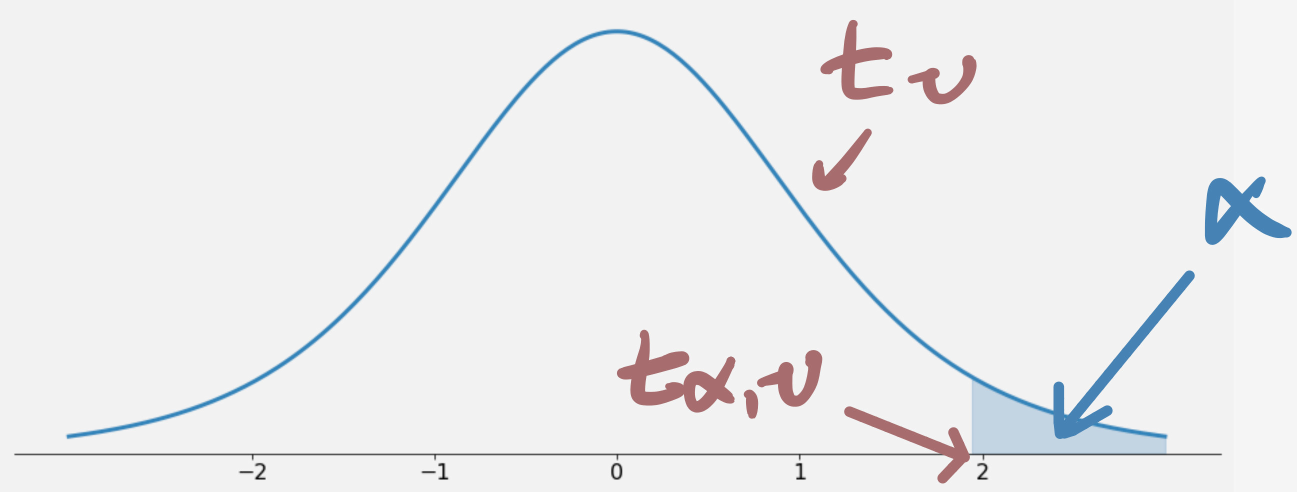
Let t_ν denote the t-Distribution with parameter ν degrees of freedom

- 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

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

Def: 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



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

The t-Confidence Interval for the Mean

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

Then a $100(1-\alpha)\%$ t-confidence interval for the mean μ is given by:

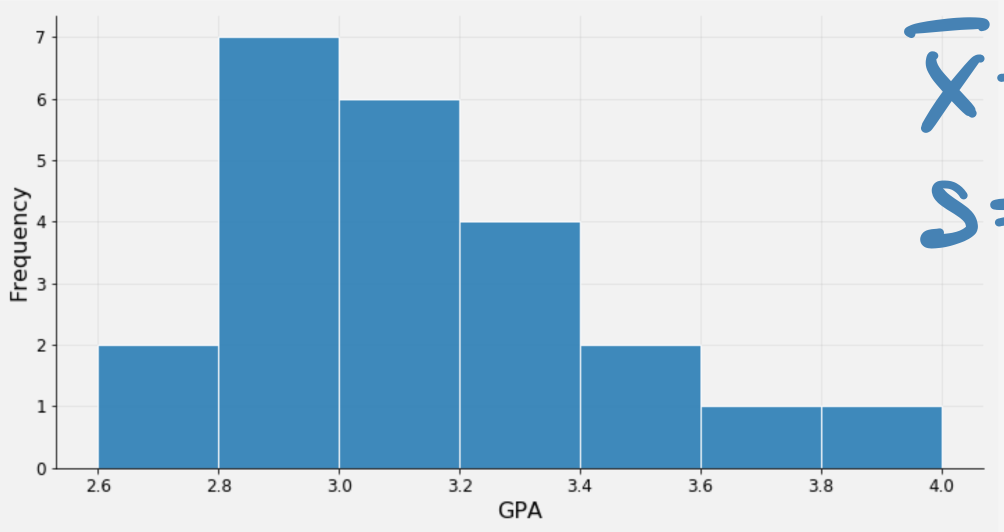
$$\left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

Or, more compactly:

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

t-Confidence Interval Example

Example: Suppose the GPAs for 23 students have a histogram that looks as follows:



$t_{.05,22}$

$\alpha = 0.1$

$\alpha/2 = .05$

$\bar{x} = 3.146$

$s = 0.308$

$n = 23$

$s/\sqrt{n} = \frac{.308}{\sqrt{23}} = .0642$

The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Find a 90% confidence interval for the mean GPA.

$t_{.05,22}$ = stats.t.ppf(.95,22) = 1.717

$3.146 \pm 1.717 * .0642 \Rightarrow [3.033, 3.259]$

The t-Test, Critical Regions and P-Values

Alternative Hypothesis	Critical Region Level α Test
$H_1: \quad \theta > \theta_0$	$t \geq t_{\alpha,\nu}$
$H_1: \quad \theta < \theta_0$	$t \leq -t_{\alpha,\nu}$
$H_1: \quad \theta \neq \theta_0$	$\left(t \leq -t_{\alpha/2,\nu}\right)$ or $\left(t \geq t_{\alpha/2,\nu}\right)$

STUDENTIZED TEST STATISTIC

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

The t-Test, Critical Regions and P-Values

Alternative Hypothesis	P-Value Level α Test
$H_1: \quad \theta > \theta_0$	$P(T \geq t \mid H_0) \leq \alpha$
$H_1: \quad \theta < \theta_0$	$P(T \leq t \mid H_0) \leq \alpha$
$H_1: \quad \theta \neq \theta_0$	$2\min(P(T \leq t \mid H_0), P(T \geq t \mid H_0)) \leq \alpha$

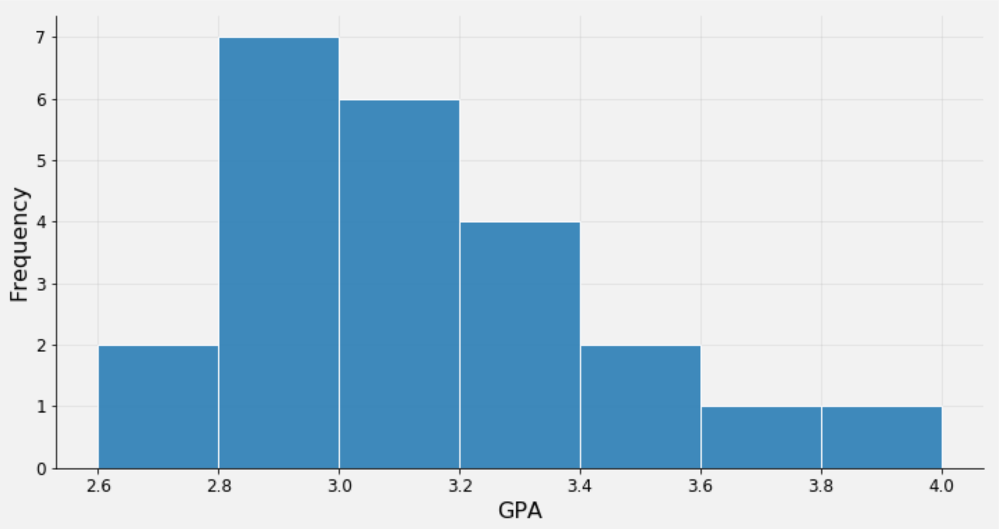
t-Test Example

Example: Suppose the GPAs for 23 students have a histogram that looks as follows:

$\bar{x} = 3.146$

$s = 0.308$

$n = 23$



$H_0: \mu = 3.30$

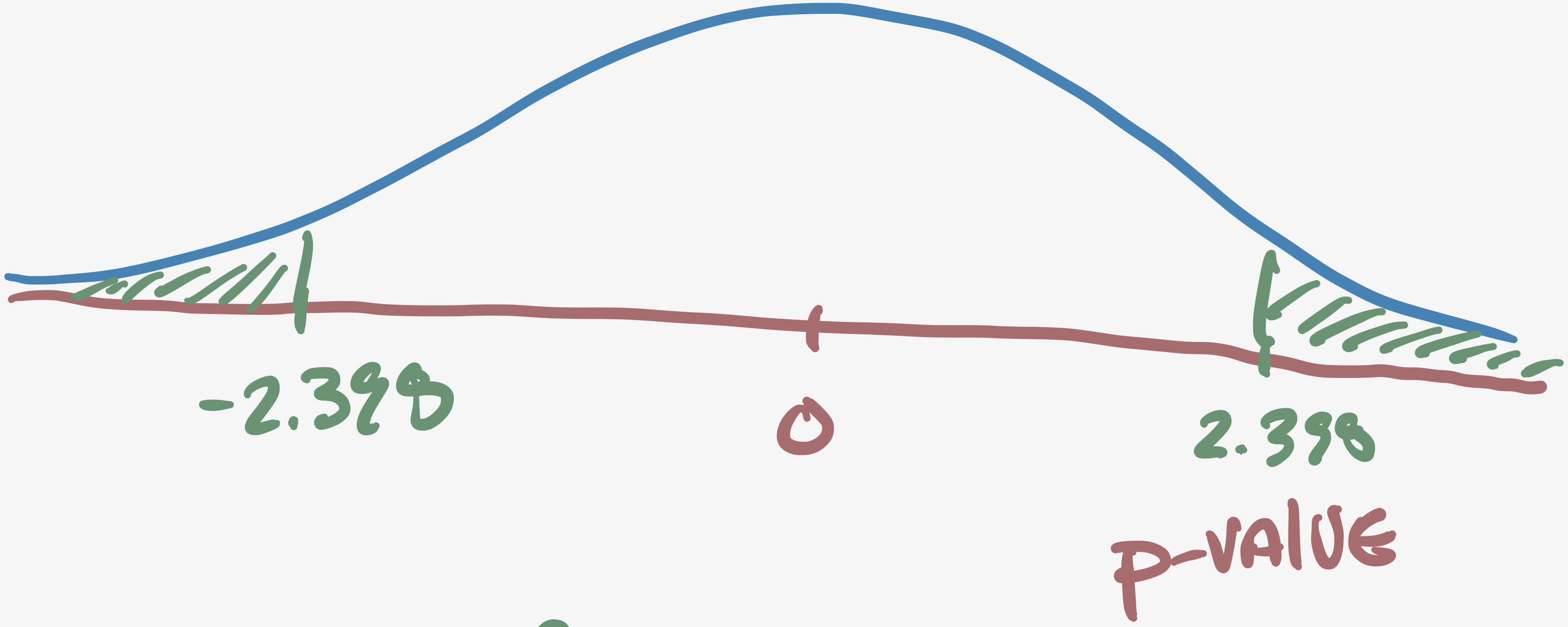
$H_1: \mu \neq 3.30$

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 significance level that the mean GPA is not equal to 3.30.

$$t = \frac{3.146 - 3.30}{0.308/\sqrt{23}} = -2.398$$

t-Test Example

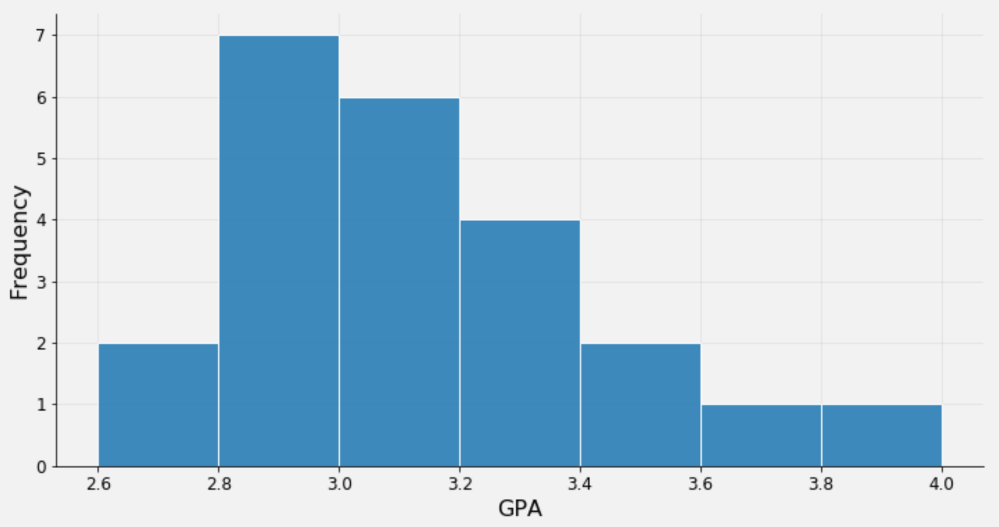
(w/ p-values)



2 × stats.t.cdf(-2.398, 22) = 0.0254 < 0.10

t-Test Example

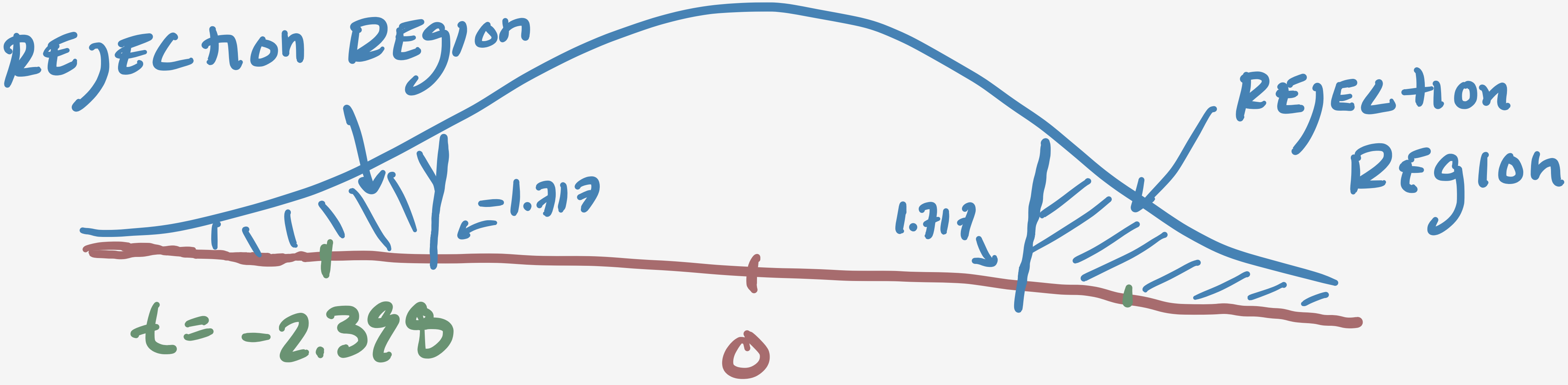
Example: Suppose the GPAs for 23 students have a histogram that looks as follows:



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 significance level that the mean GPA is not equal to 3.30.

t-Test Example

$t_{\alpha/2, 22} = t_{.05, 22} =$ stats.t.ppf(0.95, 22) $= 1.71$



Since $t = -2.398 < -t_{.05, 22} = -1.717$

We reject H_0

Inference for Variances

We've talked about estimating confidence intervals for the variance of a population using the Bootstrap

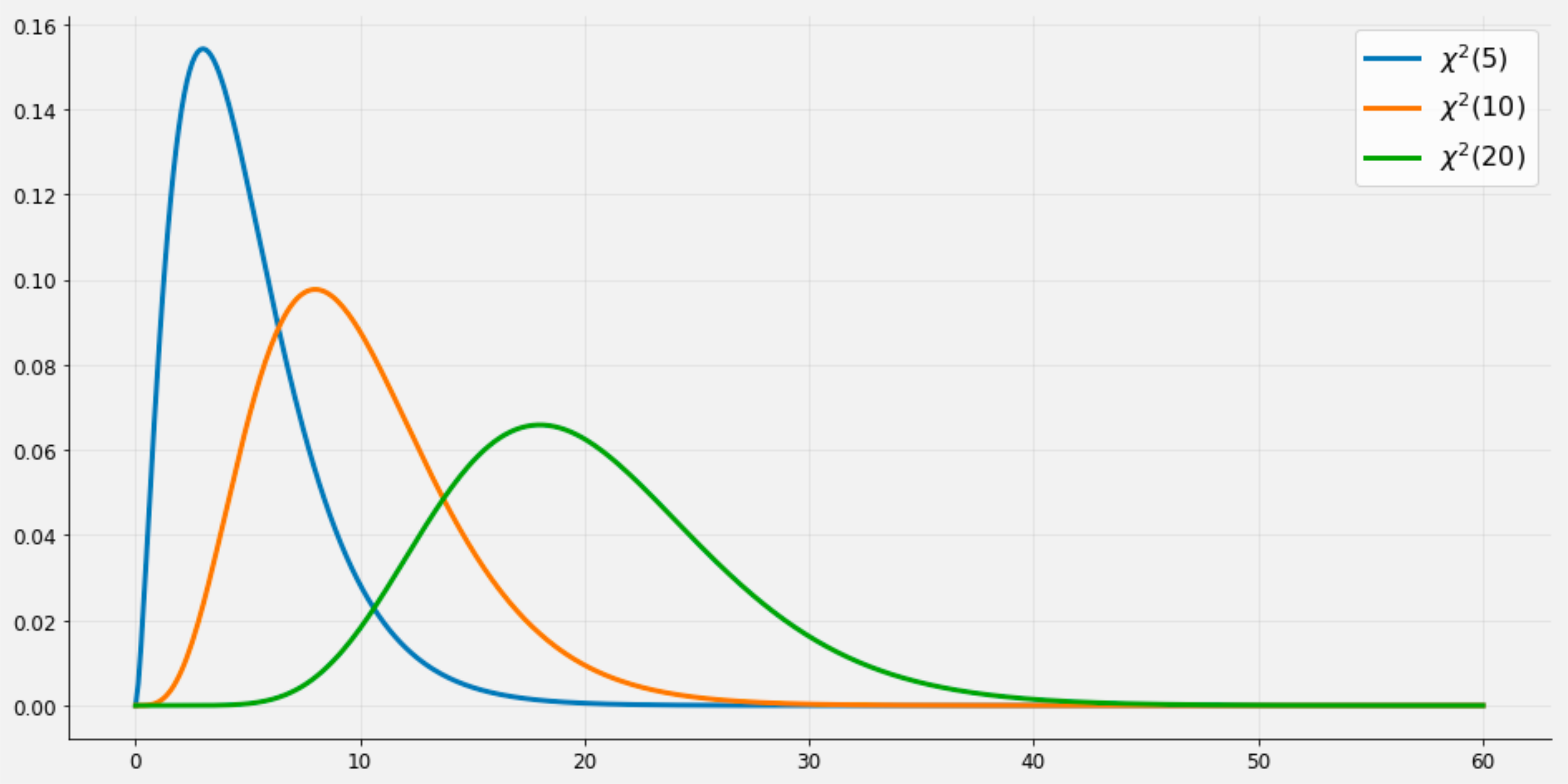
But if your population is **normally distributed,** we have some theory which gives us a better confidence interval and works for both large and small sample sizes

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

The Chi-Squared Distribution

The chi-squared (χ^2_ν) distribution is also parameterized by degrees of freedom $\nu = n - 1$

The pdfs of the family of χ^2_ν distributions are gross, so lets just draw them.



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

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})^2$$

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

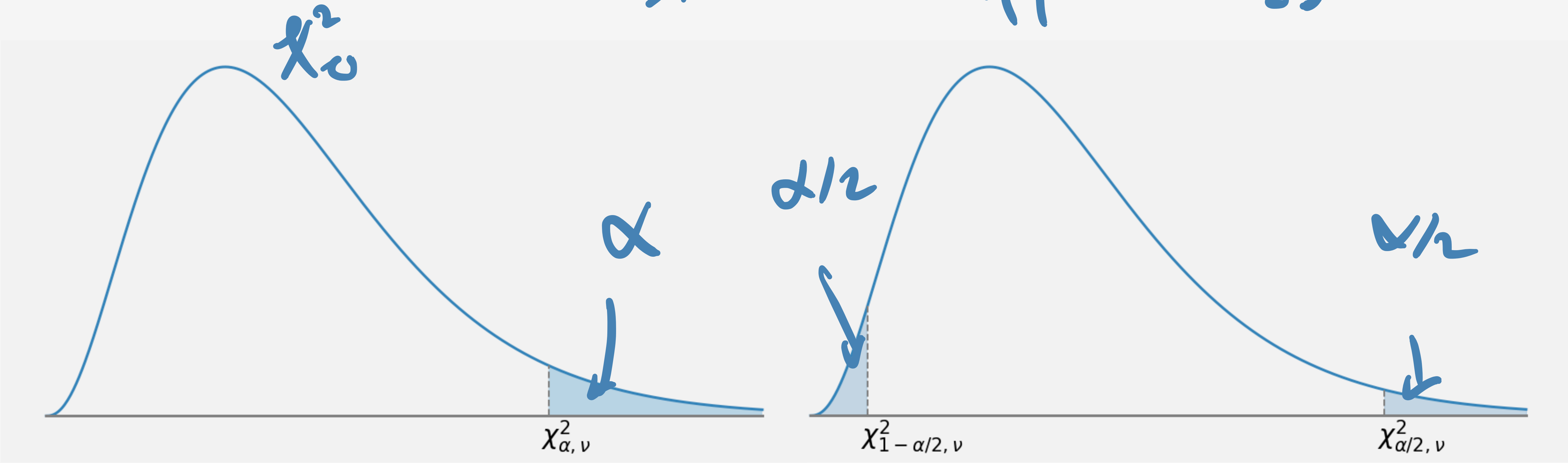
Then it follows that

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

The Chi-Squared Dist is Non-Symmetric

Because the distribution is non-symmetric, we need to use two different critical values.

STATS.CHI2.PPF(1-α/2)



A Confidence Interval for the Variance

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

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

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

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

A Confidence Interval for the Variance

For a $100(1-\alpha)\%$ confidence interval we choose the two critical values $\chi^2_{1-\alpha/2,n-1}$ and $\chi^2_{\alpha/2,n-1}$ which attributes $\alpha/2$ probability to each tail. Then, with $100(1-\alpha)\%$ 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: How can we use this to get a $100(1-\alpha)\%$ confidence interval for the standard deviation?

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

Variance CI Example

Example: A large candy manufacturer produces packages of candy targeted to weight 52g. 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 she selects n=10 bags at random and weighs them. The sample yields a sample variance of 4.2g. Find a 95% confidence interval for the variance and a 95% confidence interval for the standard deviation.

$\alpha = .05$ $\quad \alpha/2 = .025$ $\quad n = 10$ $\quad s^2 = 4.2$

$\chi^2_{1-\frac{\alpha}{2},n-1} = \chi^2_{.95,9} = $ STATS.CHI2.ppf(.05,9) = 2.70

$\chi^2_{\alpha/2,n-1} = \chi^2_{.05,9} = $ STATS.CHI2.ppf(.95,9) = 19.02

$$\frac{(10-1)4.2}{19.02} = 1.99, \quad \frac{(10-1)4.2}{2.70} = 14.0$$

95% CI FOR σ^2: $[1.99, 14.0]$, 95% CI FOR σ: $[1.41, 3.74]$

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 19 In-Class Notebook

Let's:

- Do some stuff!