

CSCI 3022

intro to data science with probability & statistics

Lecture 20
April 2, 2018

Small sample size hypothesis testing
and The Bootstrap

Stuff & Things

- HW5 due **this Friday.** ✓
- **New notetaker needed please!** Done
 - 1. Take notes as you normally would.
 - 2. Scan (with smartphone) after class and email to two of your peers.
- Questions about your HW grades? The graders are happy to explain!
 - sudeep.galgali@colorado.edu ✓
 - ajay.kedia@colorado.edu ✓

C+P
Overhead in OH
→ ASCII

H.C. O or F ✓
Course policy!

Previously on CSCI 3022

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 is large" $n \geq 30$	"n is small" $n < 30$
Normal Data / Known σ		
Normal Data / Unknown σ ^{use s}		
Non-Normal Data / Known σ		
Non-Normal Data / Unknown σ		

- z-test

- t-test (TODAY!)

Bootstrap
(after Spring Break)

The t-Test, Critical Regions and P-Values

$$H_0 : \theta = \theta_0$$

Alternative Hypothesis

$$H_1 : \theta > \theta_0$$

$$H_1 : \theta < \theta_0$$

$$H_1 : \theta \neq \theta_0$$

Critical Region Level α Test

$$t \geq t_{\alpha, \nu}$$

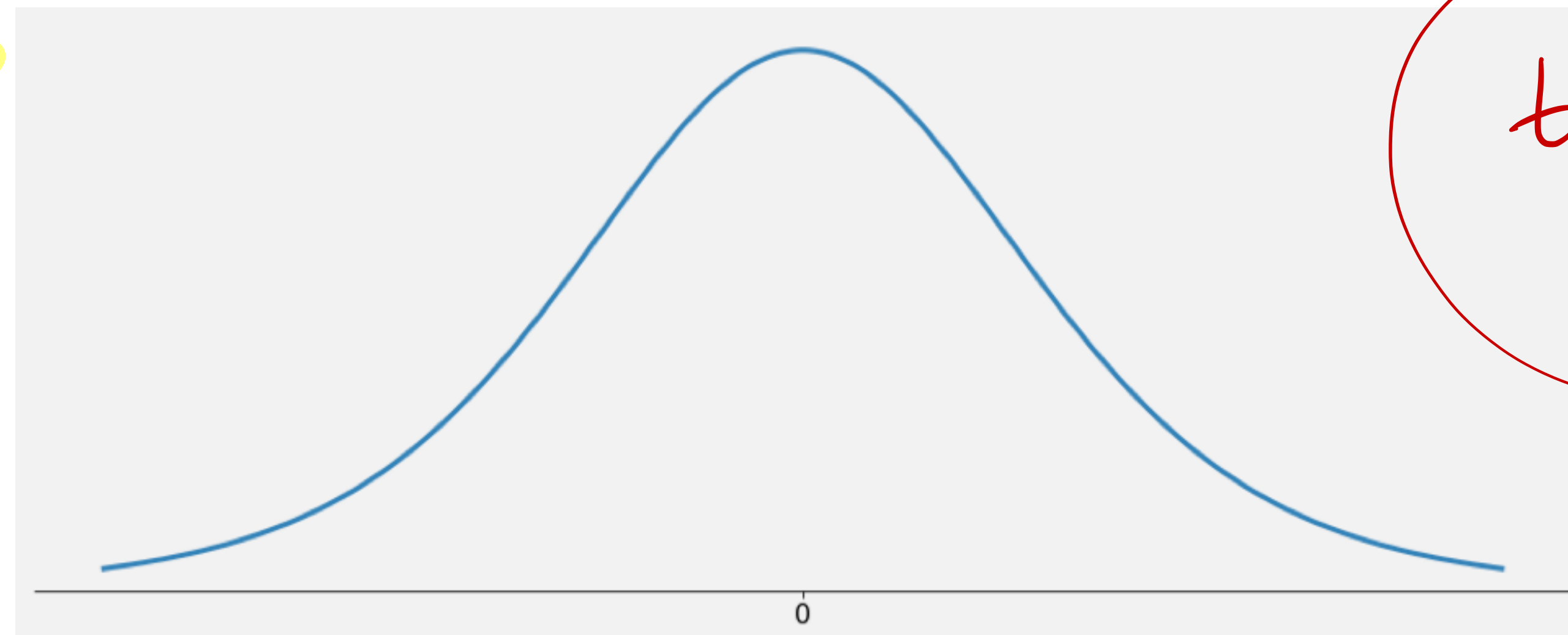
$$t \leq t_{\alpha, \nu}$$

$$(t \leq -t_{\alpha/2, \nu}) \text{ or } (t \geq t_{\alpha/2, \nu})$$

confidence degrees of freedom = $n-1$

t test statistic looks just like z test statistics!

The only difference is n is small ($n < 30$)



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

"standardized statistic"

The t-Test, Critical Regions and P-Values

Alternative Hypothesis

P-Value Level α Test

$$H_1 : \theta > \theta_0$$

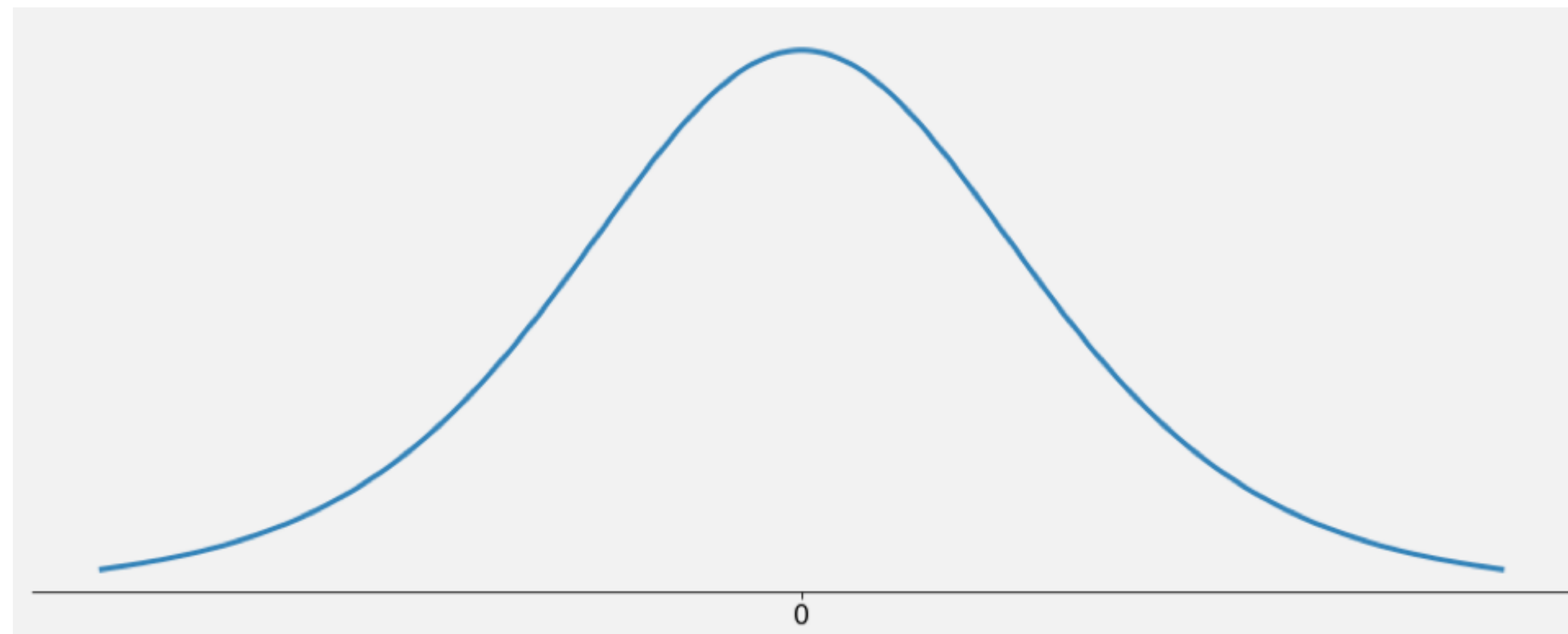
$$P(T \geq t \mid H_0) \leq \alpha$$

$$H_1 : \theta < \theta_0$$

$$P(T \leq t \mid H_0) \leq \alpha$$

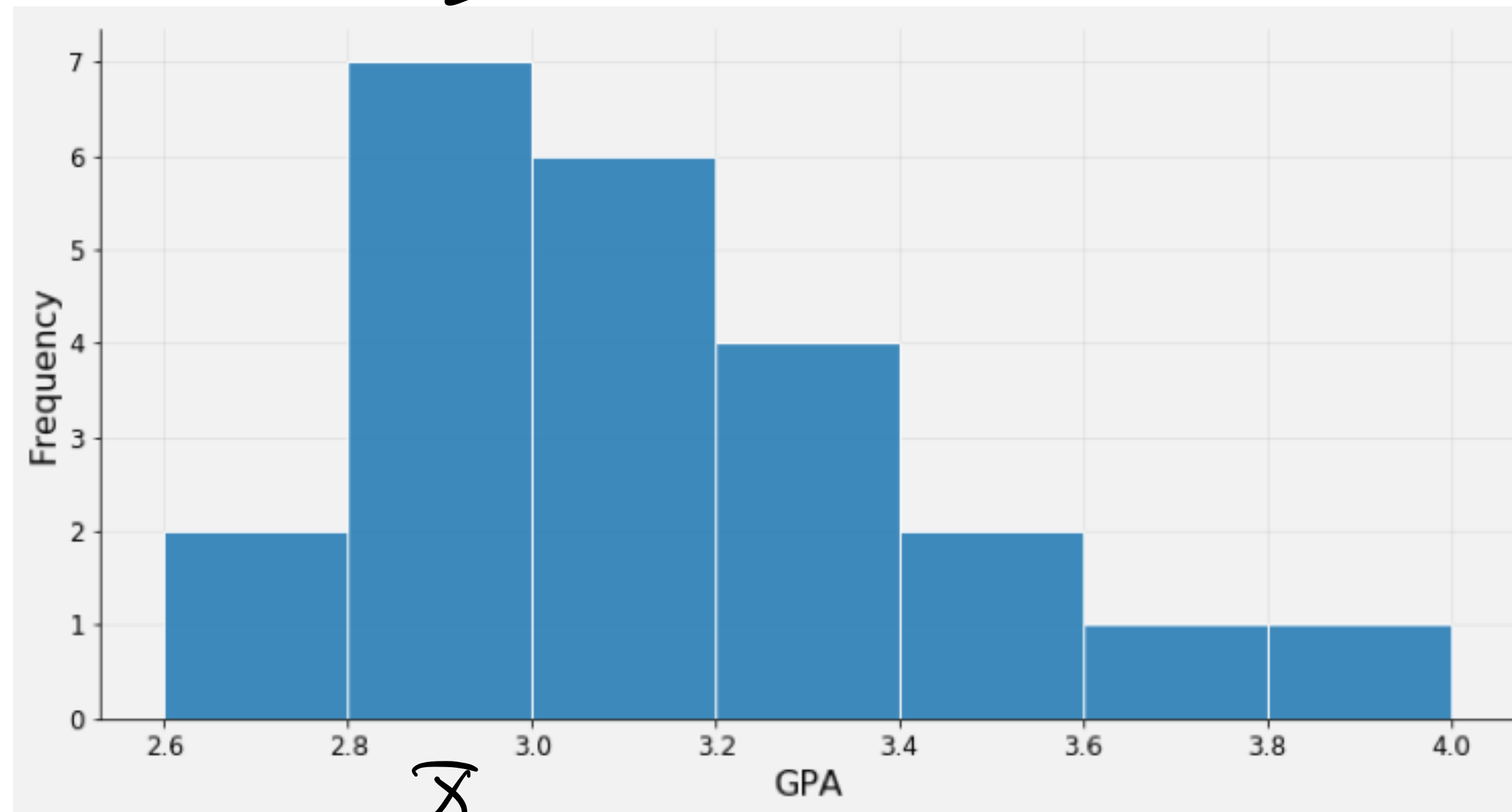
$$H_1 : \theta \neq \theta_0$$

$$2 \min \{P(T \leq t \mid H_0), P(T \geq t \mid H_0)\} \leq \alpha$$



t-Test example (p-value method)

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



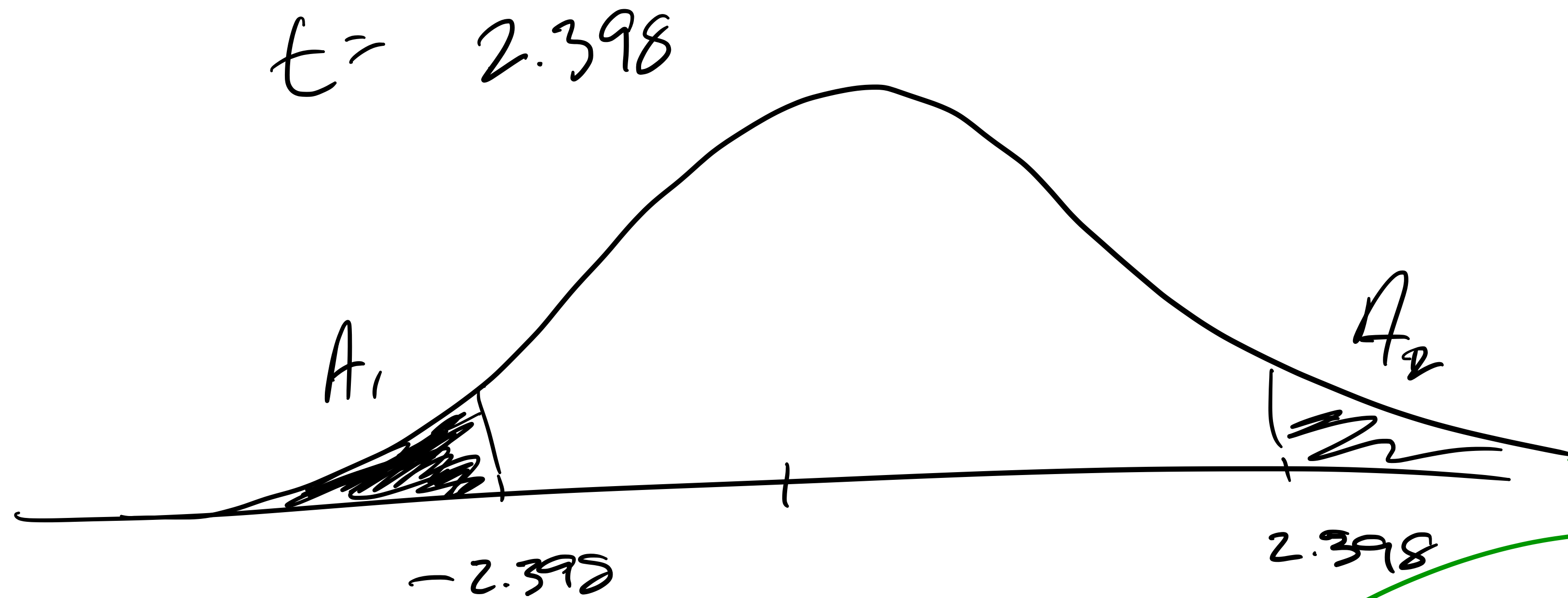
- The sample mean of the data is $\bar{x} = 3.146$ and the sample standard deviation is $s = 0.308$. Determine if there is sufficient evidence to conclude at the $\alpha = 0.10$ significance level that the mean GPA is not equal to 3.30.

$$H_1: \text{GPA} \neq 3.30$$

$$\alpha = 0.1$$

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

t-Test example (p-value method)



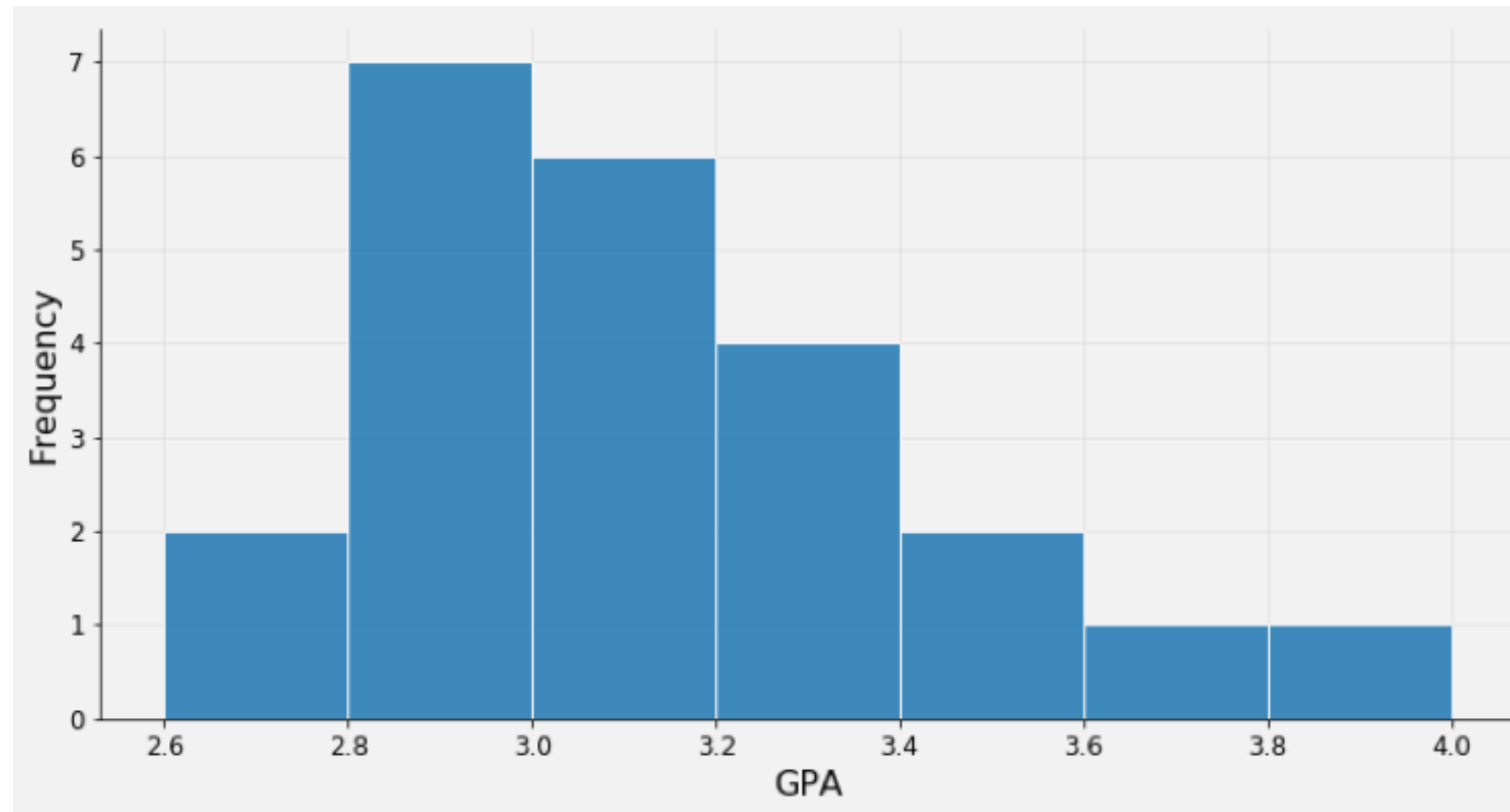
$$2 \times \text{stats.t.cdf}(\underset{\substack{\uparrow \\ \text{test} \\ \text{Statistic}}}{-2.398}, \underset{\text{dof}}{22}) = \boxed{0.0254} < 0.10$$

$p\text{-value}$

α

t-Test example (rejection region method)

- **Example:** Suppose the GPAs for $n = 23$ 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.

$$\mu = 3.30$$

$$\bar{x} = 3.146$$

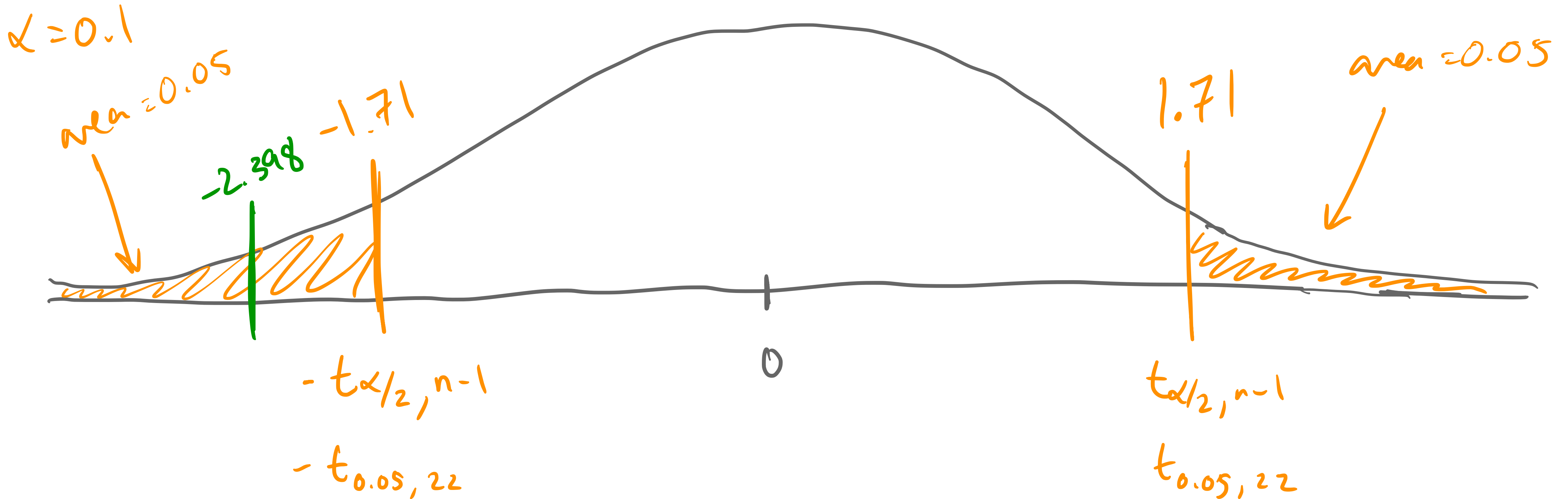
$$\alpha = 0.1$$

$$H_0: \mu = 3.30$$

$$H_1: \mu \neq 3.30$$

$$s = 0.308$$

t-Test example (rejection region method)



stats. t. ppf(0.95, 22) = 1.71

Prev. Slides $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = -2.398$ "test statistic"

In the rejection region! REJECT H_0 .

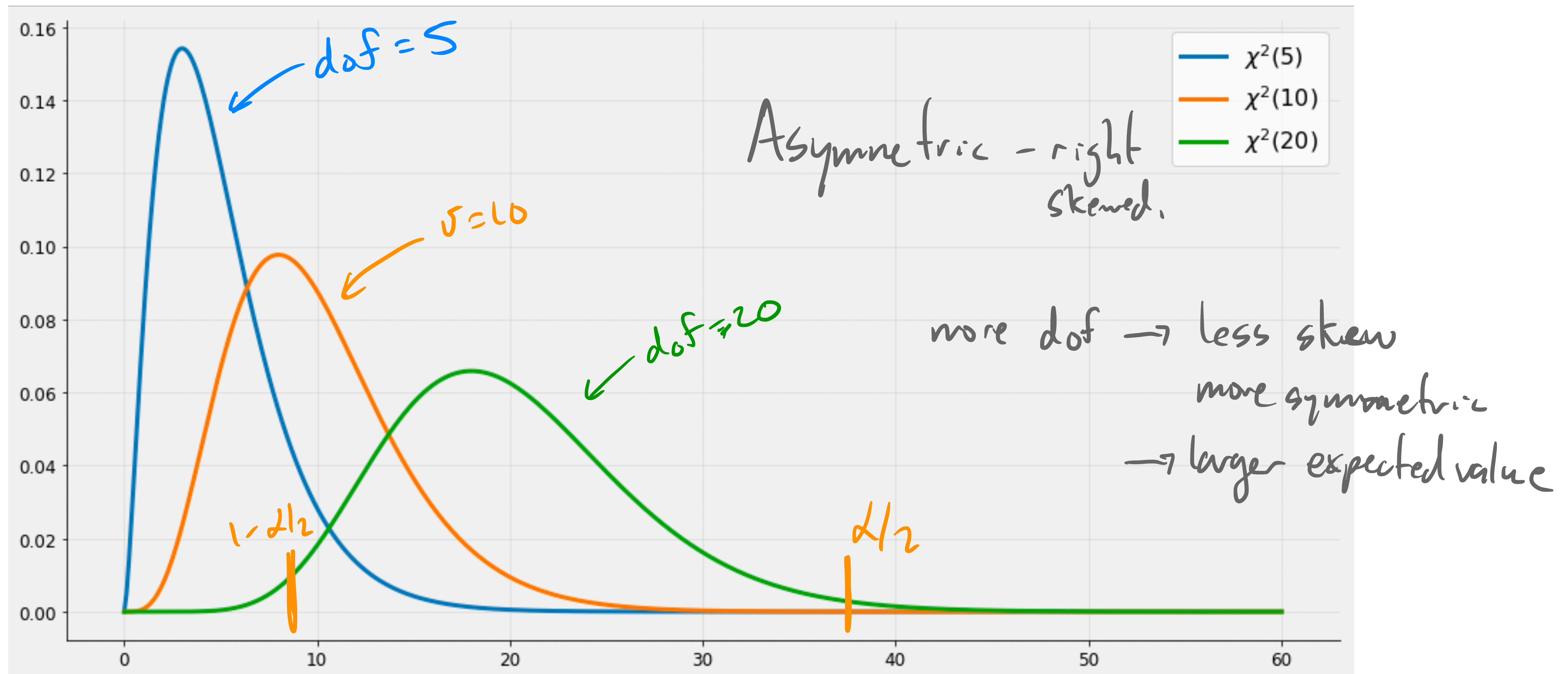
Inference for *variances*

Today

- ~~After Spring Break~~, we'll talk about estimating confidence intervals for the variance of a population using something [wonderful] called **The Bootstrap**.
- But if your population is normally distributed, we have some [wonderful] 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 χ^2

- The chi-squared distribution (χ^2_ν) is also parameterized by degrees of freedom $\nu = n - 1$
- The pdfs of the family of χ^2_ν distributions are gross, so let's just draw them! \sqrt{nu}



A confidence interval for the variance

- Let X_1, X_2, \dots, X_n be IID samples 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 $(n-1)S^2/\sigma^2$ follows the distribution χ_{n-1}^2 .

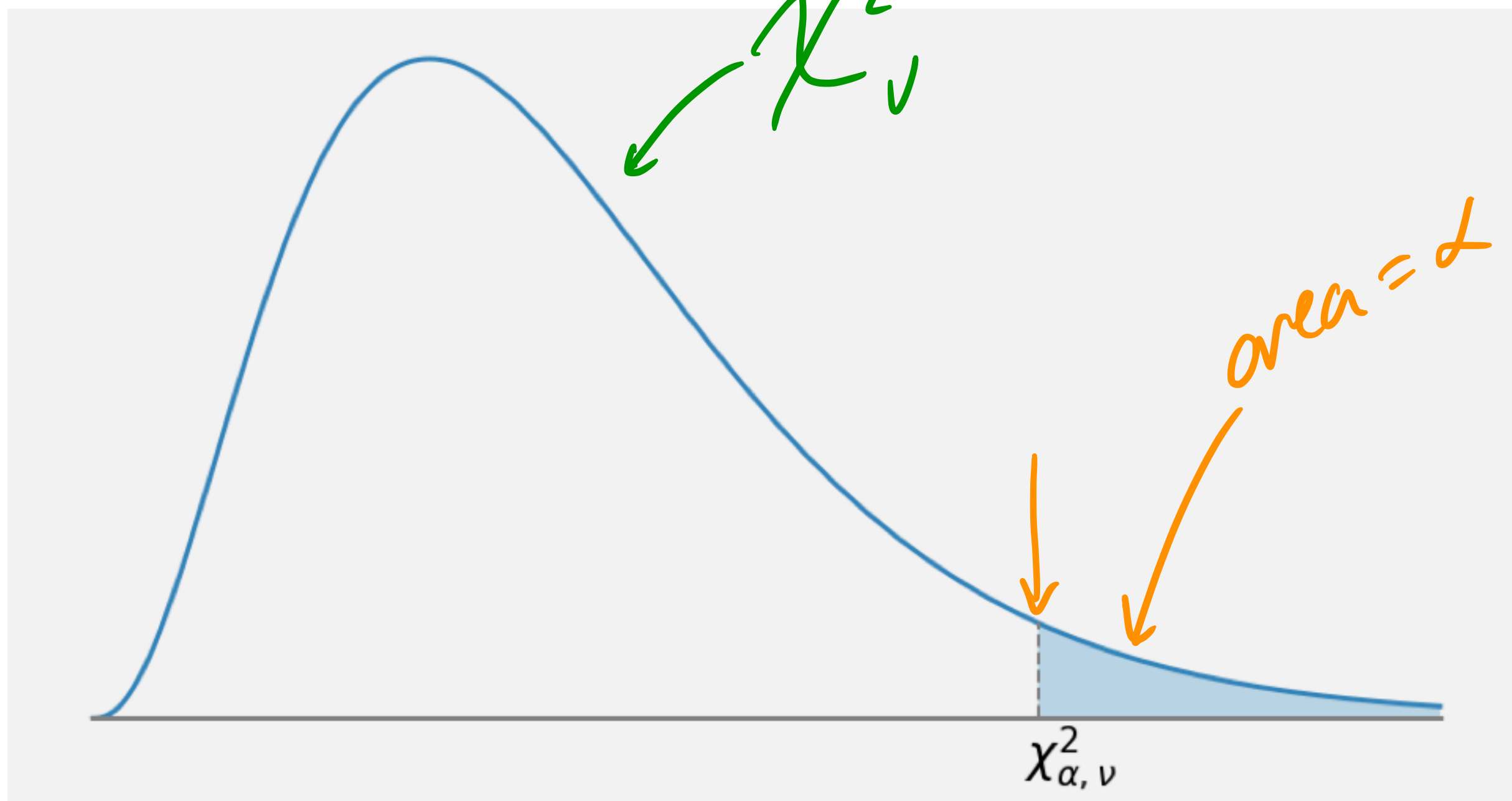
- Then it follows that

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

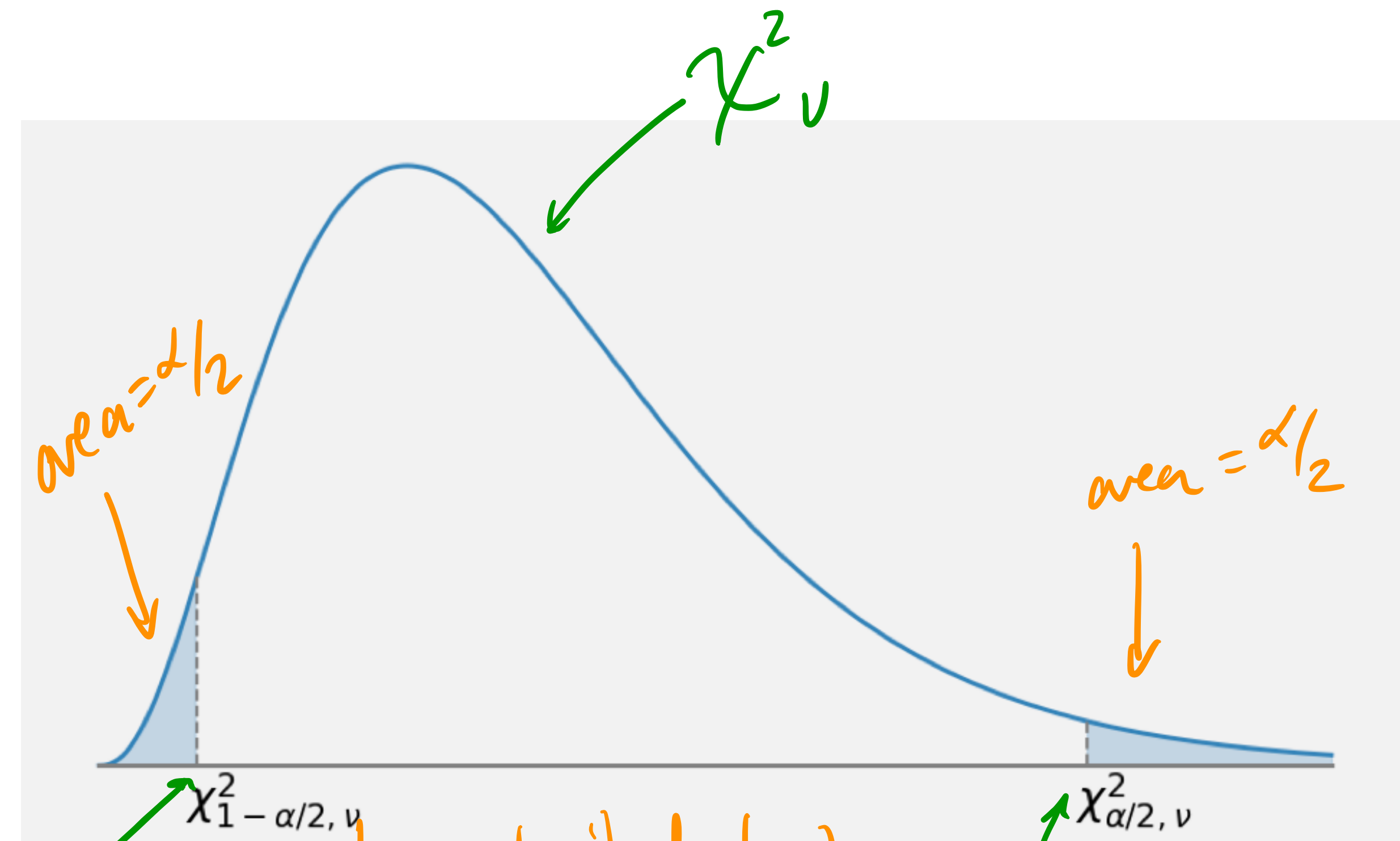
Left bound Right bound

The Chi-Squared Dist is Non-Symmetric

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



one-tailed test



two-tailed test

$\text{stats.chi2.ppf}(1-\alpha/2, v)$ $\text{stats.chi2.ppf}(\alpha/2, v)$

A confidence interval for the variance

$$\frac{x}{y} < \frac{s}{y}$$

$$x < sy$$

$$y > \frac{x}{s}$$

- 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 puts $\alpha/2$ probability in each tail. Then, with $100(1 - \alpha)\%$ confidence we can say that

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

Solve for this

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

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

Conf. Interval for variance σ^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 puts $\alpha/2$ probability in 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?

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

yay!

useless!

- 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.

$$s^2 = 4.2$$

$$n = 10$$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

$$\frac{(n-1)s^2}{\chi^2}$$

$$\frac{(10-1)4.2}{19.02} = 1.99$$


$$\frac{(10-1)4.2}{2.70} = 14.0$$

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

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

$$95\% \text{ CI for } \sigma^2: [1.99, 14.0]$$

$$\text{for } \sigma = [1.41, 3.74]$$

A red heart outline is centered on the page. The heart is drawn with a single continuous line and has a slightly irregular, hand-drawn appearance. It is open at the top and bottom, with the lines meeting at points near the top and bottom centers.

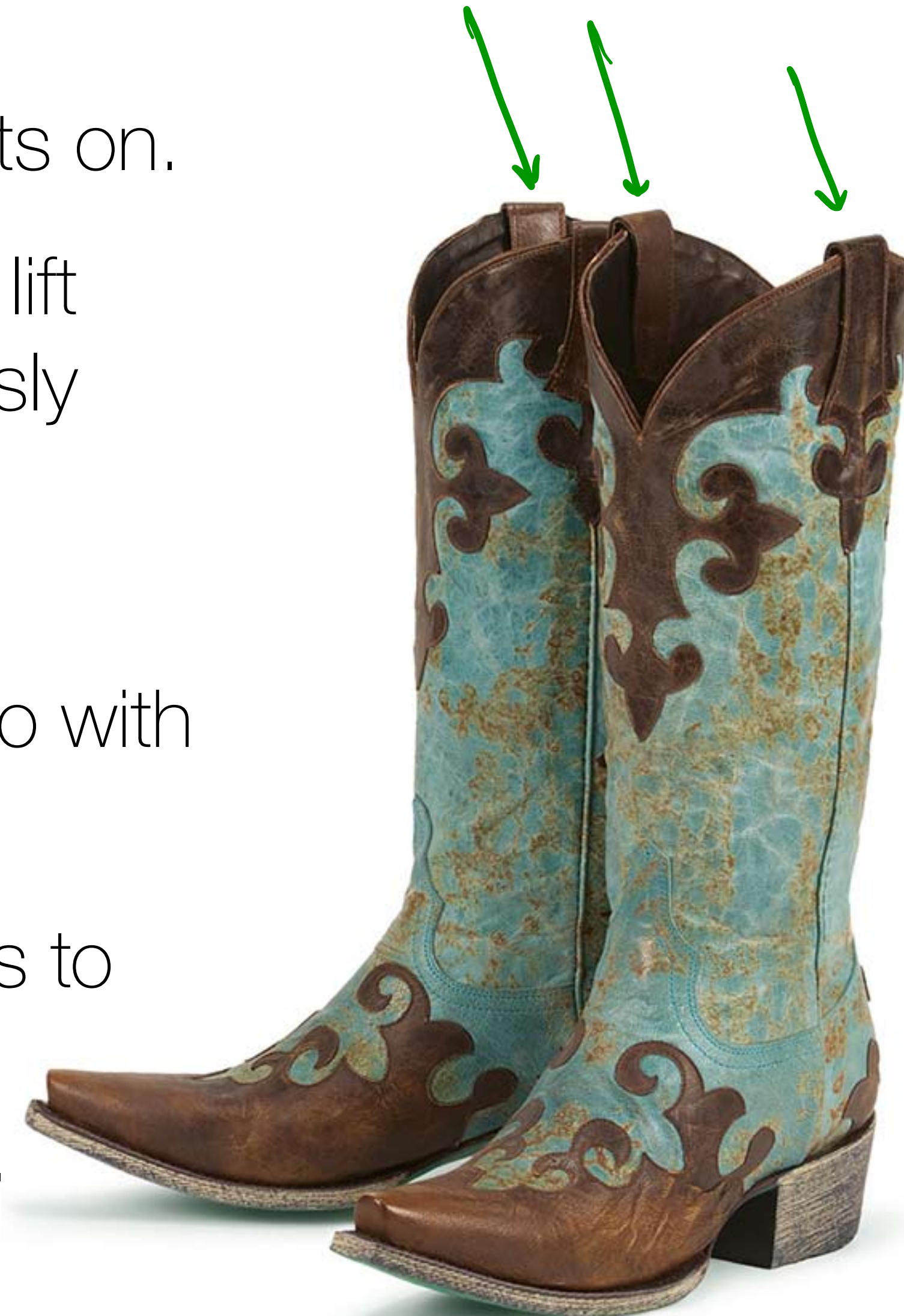
The Bootstrap

Not all datapoints come cheap...

- In real scenarios, **data can be expensive**...
 - in **money**. For example, data from an aircraft in a wind tunnel.
 - in **time**. For example, polling people in surveys is time consuming.
 - in **privacy tradeoffs**. For example, storing another person's genome in the database incurs ethical risk or cost, even when it does not cost much time or money.
- Today, we'll learn a technique that enables us to learn from small amounts of data to compute confidence intervals: **the bootstrap**

What are bootstraps?

- Bootstraps are the straps that you use to pull your boots on.
- To “pull yourself up by your bootstraps” is to somehow lift yourself upward by pulling on your own shoes. Obviously impossible.
- Now, however, bootstrapping means to accomplish something without aid. To accomplish what you need to with what you’ve got.
- The statistical bootstrap is in this last sense. It allows us to really **make the most of a small dataset** without sacrificing statistical rigor or collecting more \$ samples.



A confidence interval for the mean

- **Recall:** if we have n samples from a distribution that is normal or non-normal, then by the Central Limit Theorem, the confidence interval for the mean is given by $\bar{X} \pm z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}}$ or for an unknown variance $\bar{X} \pm z_{\alpha/2} \sqrt{\frac{s^2}{n}}$
- The bootstrap is a different approach. Consider the same set of samples as above, X_1, X_2, \dots, X_n , but instead of computing a CI analytically from this sample, instead *re-sample* your sample many times and examine (?) those!
- **Definition:** a bootstrapped resample is a set of n draws from the original set, sampled *with replacement*.
of n values

A confidence interval for the mean

- **Definition:** a bootstrapped resample is a set of n draws from the original dataset (drawn IID from X), sampled *with replacement*.
- **Example:** suppose we have the data $[2, 4, 6, 7, 9]$
 - Resample 1 might be: $[4, 6, 7, 4, 9]$
 - Resample 2 might be: $[2, 6, 7, 9, 2]$
 - Resample 3 might be: $[9, 7, 7, 7, 4]$
- Given the example above, what does “sample with replacement” mean?

A confidence interval for the mean

- **Definition:** a bootstrapped resample is a set of n draws from the original dataset (drawn IID from X), sampled *with replacement*.
- **Proposition:** a suitable estimate of the 95% confidence interval for the mean of the distribution X is given by $[a, b]$, where a and b are the 2.5 percentile and 97.5 percentile of the means of a large number of bootstrapped resamples.
- **In plain English:** resample your original data many times. Compute the mean for each resample. Compute the 2.5 and 97.5 percentiles of those means.

Magic!