CSCI 3022

intro to data science with probability & statistics

Lecture 19
March 19, 2018

Small sample size hypothesis testing

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

CSCI 3022

intro to data science with probability & statistics

Lecture 19
March 19, 2018

Small sample size hypothesis testing

(Sam Way)

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- HW5 posted tonight. Due the Friday *after* Spring Break.
- Dan’s OH cancelled this Weds & Fri.

Previously on CSCI 3022

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

"z tests"

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

$$\underbrace{\bar{x}}_{\text{center}} \pm \underbrace{z_{\alpha/2}\,\sigma/\sqrt{n}}_{\text{window}}$$

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

"empirical std. dev."

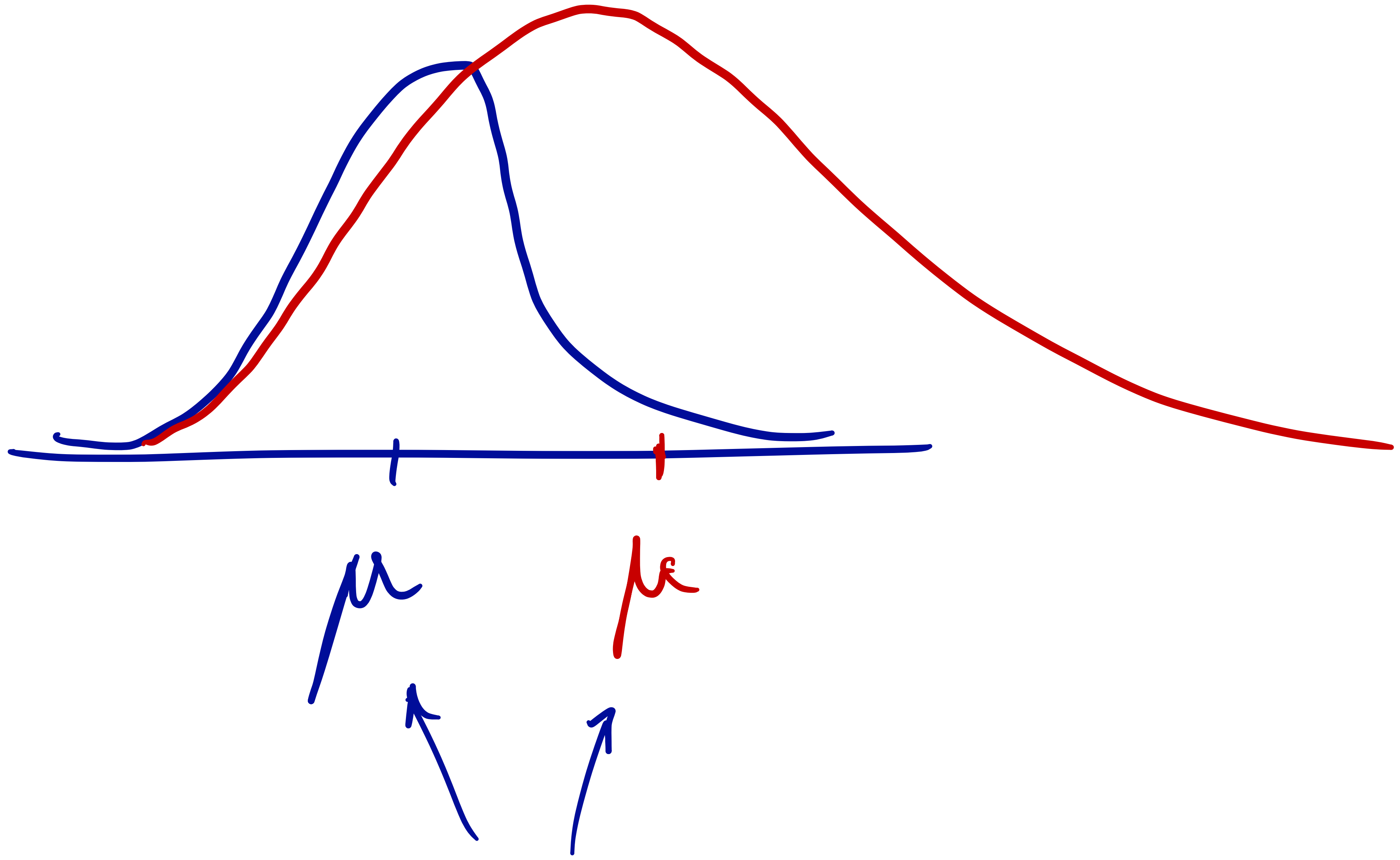
Previously on CSCI 3022

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

"Thanks, CLT!"

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

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



Previously on CSCI 3022

$n < 30$

- Statistical inference for population mean **when data is normal** and $\boxed{\text{n is small}}$ and…

- σ is known:

$$\underbrace{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}_{???} \sim N(0,1)$$

- 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 is large"	"n is small"
	$n \geq 30$	$n < 30$
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)

Small-sample tests

- 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 …

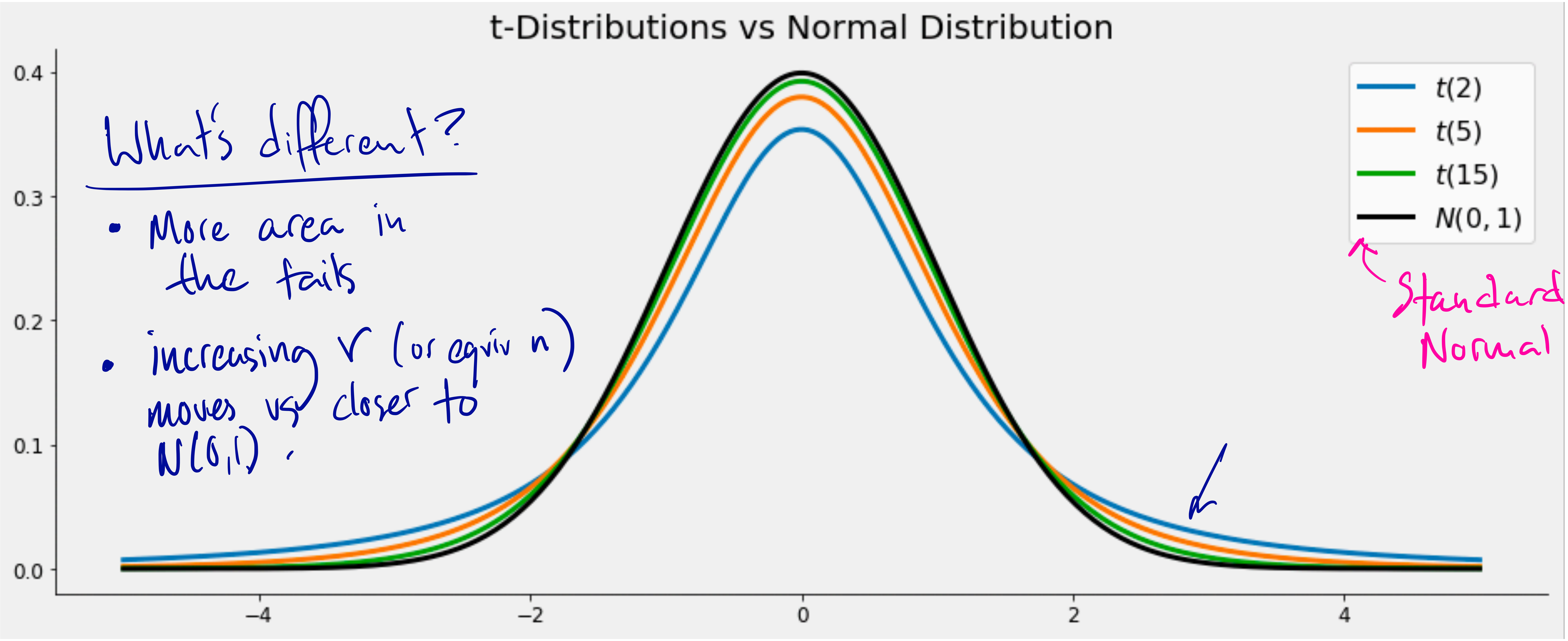
- When $\bar{X}$ is the sample mean of a random sample of size *n* from a normal distribution with mean μ, the random variable

$$\left(\frac{\bar{X}-\mu}{\underset{\uparrow}{s}/\sqrt{n}}\right)$$

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

The t-Distribution

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



What's different?

- More area in the tails
- increasing ν (or equiv n) moves us closer to N(0,1).

Standard Normal

- What do you notice about these t-Distributions, compared with the Standard Normal curve?

Properties of t-Distributions

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

$\nu = n-1$

(Aside: \nu in LaTeX.)

- 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

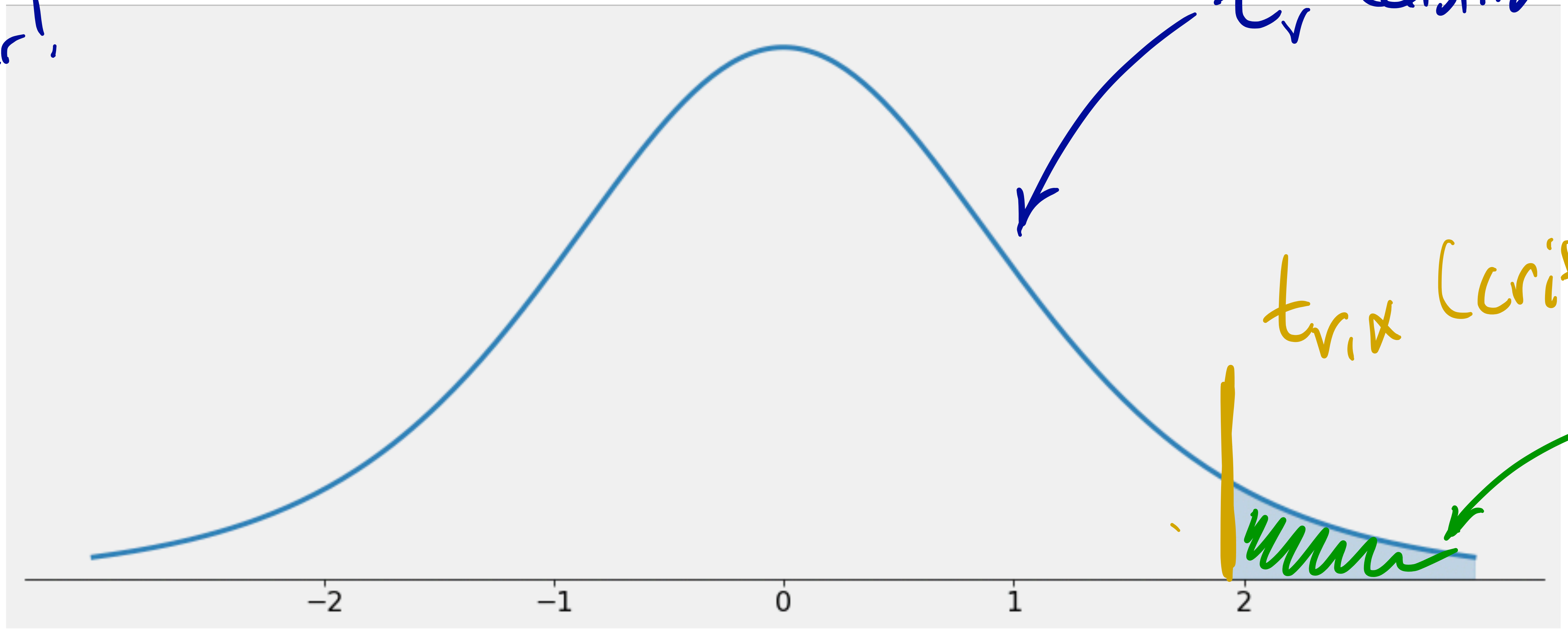
The t-critical value

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

- **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…

This should look familiar!

Very similar to our old friend, the z-test, and using z_α critical values



- 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:

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

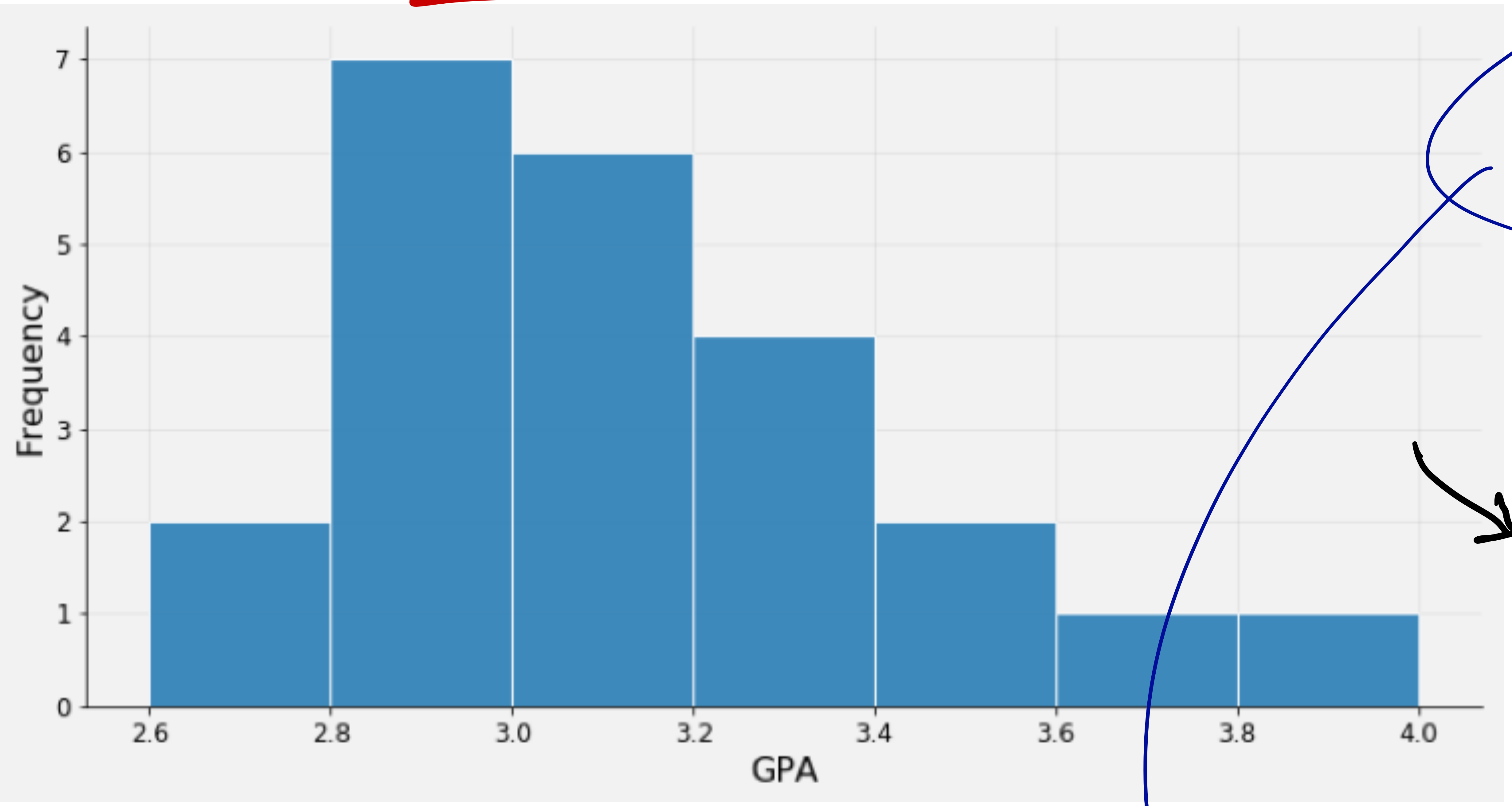
- Or more compactly:

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

CI

t-confidence interval example

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



$n = 23$

$\bar{x} = 3.146$

$s = 0.308$

$\alpha = 0.1$

$\alpha/2 = 0.05$

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

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

stats.t.ppf(0.95, 23-1)

$= 1.717$

$n-1$

- 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.

$\rightarrow \alpha = 0.1$

"$(1-\alpha) \cdot 100$" % CI

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

$\Rightarrow [3.033, 3.259]$

The t-Test, Critical Regions and P-Values

$H_0 : \quad \theta = \theta_0$

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

confidence

degrees of freedom

t test statistic looks just like z test statistics!

The only difference ... is n is small ($\approx n < 30$)

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

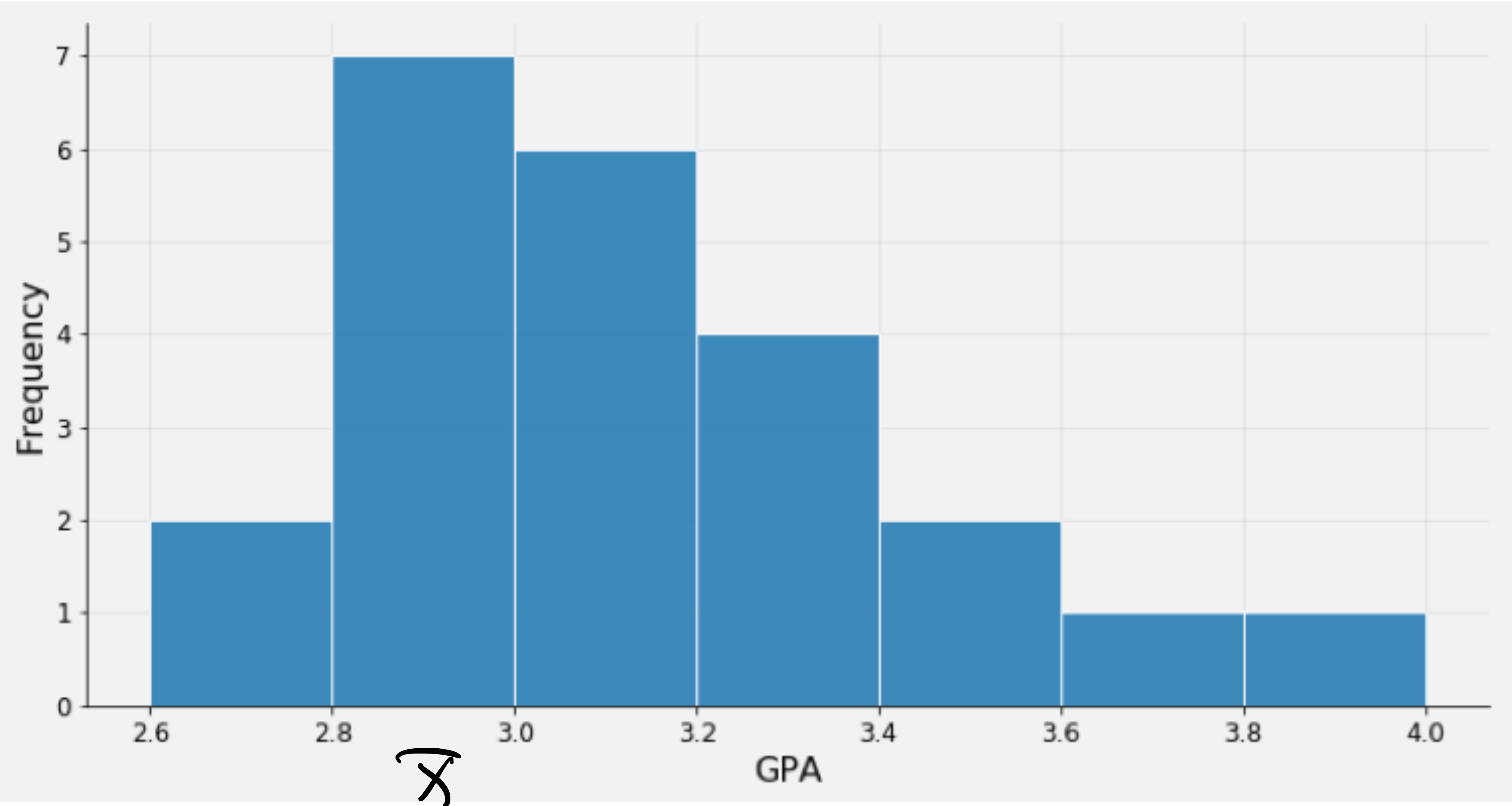
"standardized statistic"

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

H_1: GPA $\neq$ 3.30

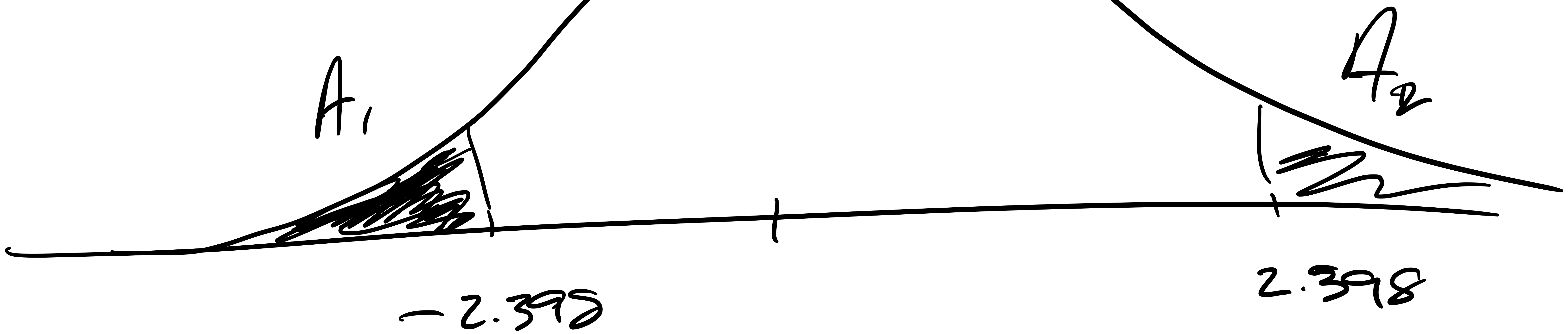
$\alpha = 0.1$

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

$$= -2.398$$

t-Test example (p-value method)

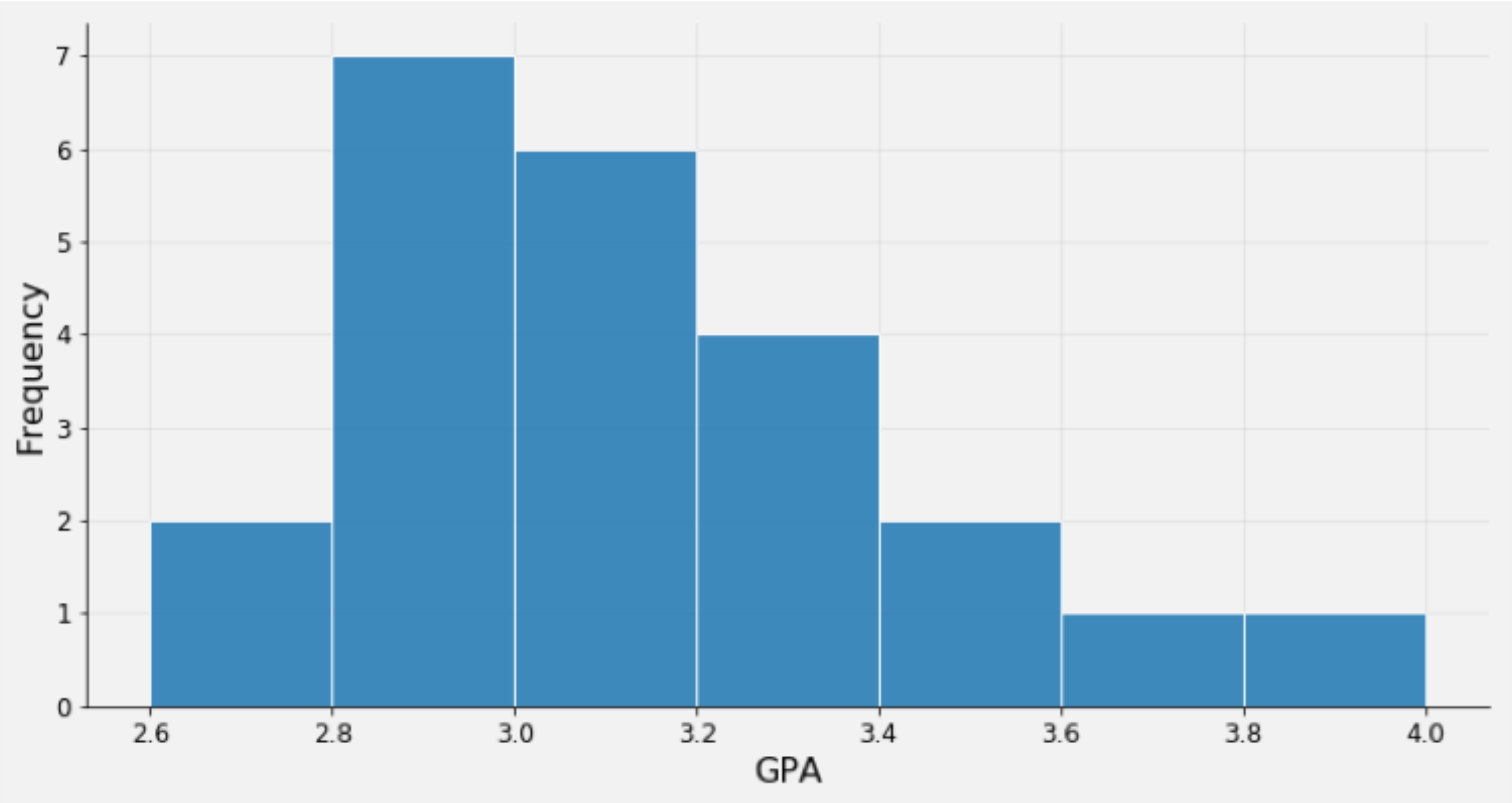
$t = 2.398$



$$2 \times \text{stats.t.cdf}(\underset{\substack{\uparrow \\ \text{test} \\ \text{statistic}}}{-2.398}, \underset{\text{dof}}{22}) = \underbrace{\boxed{0.0254}}_{\text{p-value}} < \underset{\alpha}{0.10}$$

t-Test example (rejection region method)

- **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 (rejection region method)

Inference for *variances*

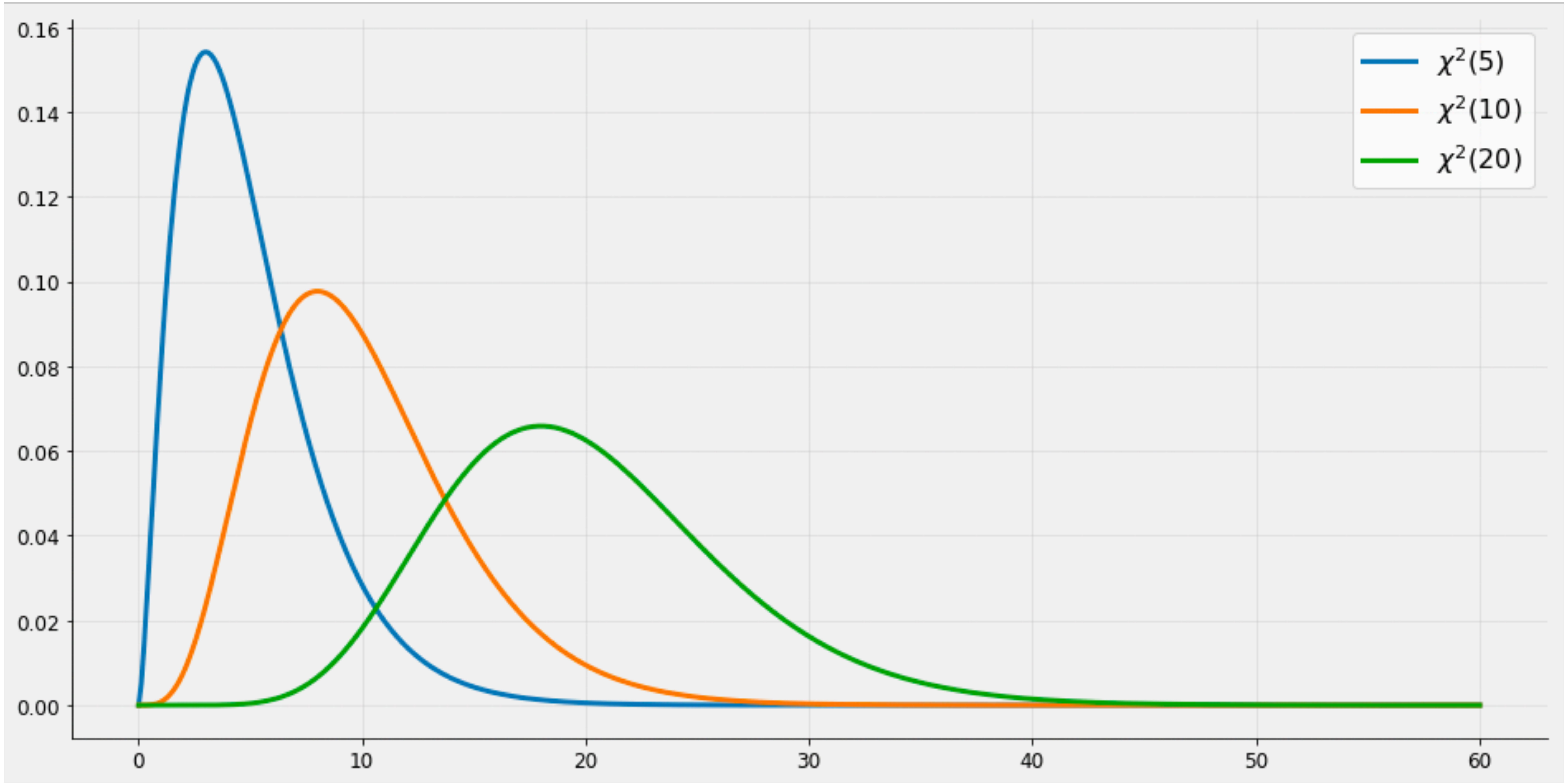
- After Spring Break, we'll talk about estimating confidence intervals for the variance of a population using something [wonderful] called **The Bootstrap**.

- 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

- 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 lets just draw them!



A confidence interval for the variance

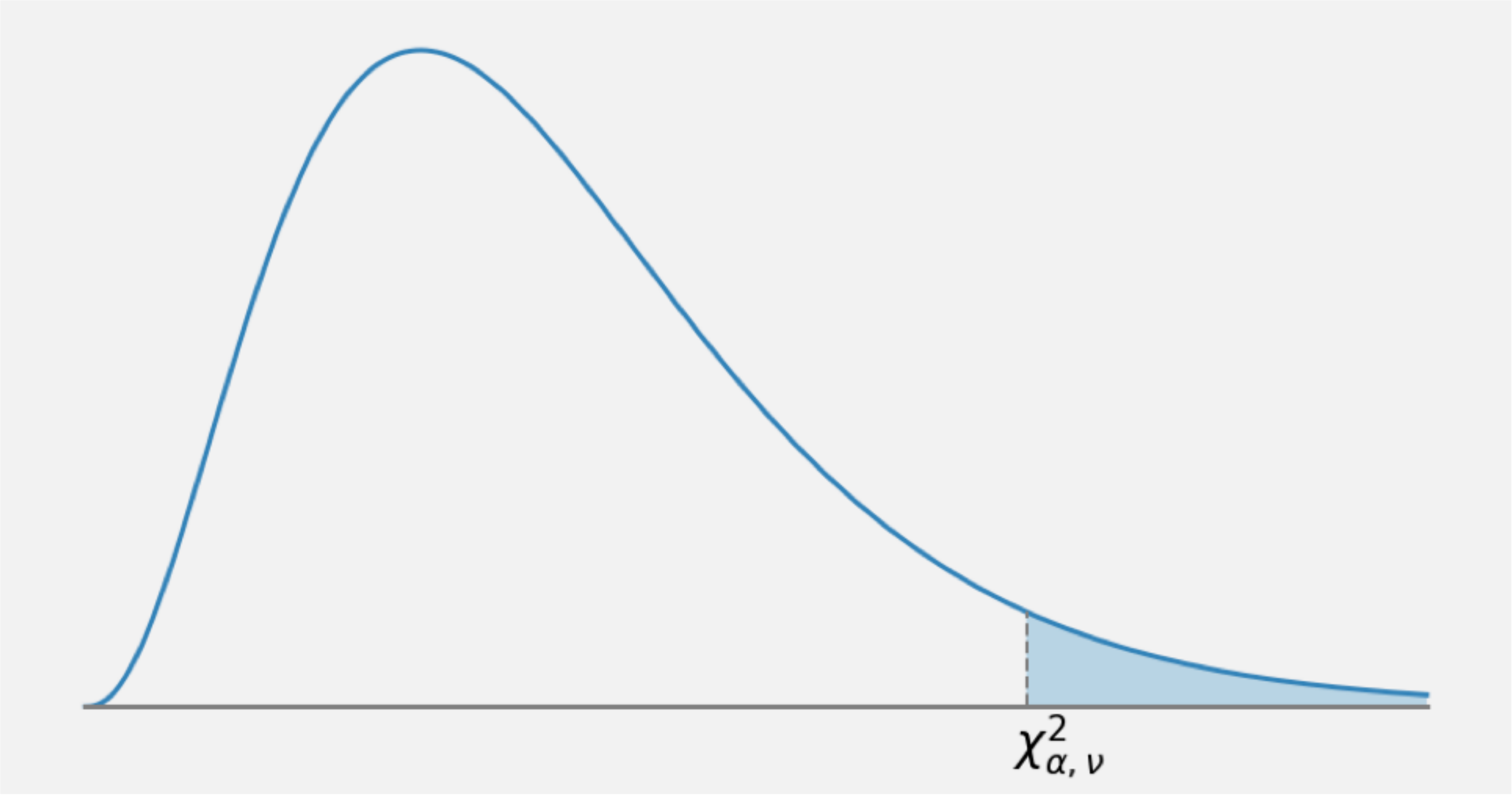
- Let $X_1, X_2, \ldots X_n$ be IID samples from a normal distribution with mean μ and standard deviation σ . Define the *sample variance* in the usual way as

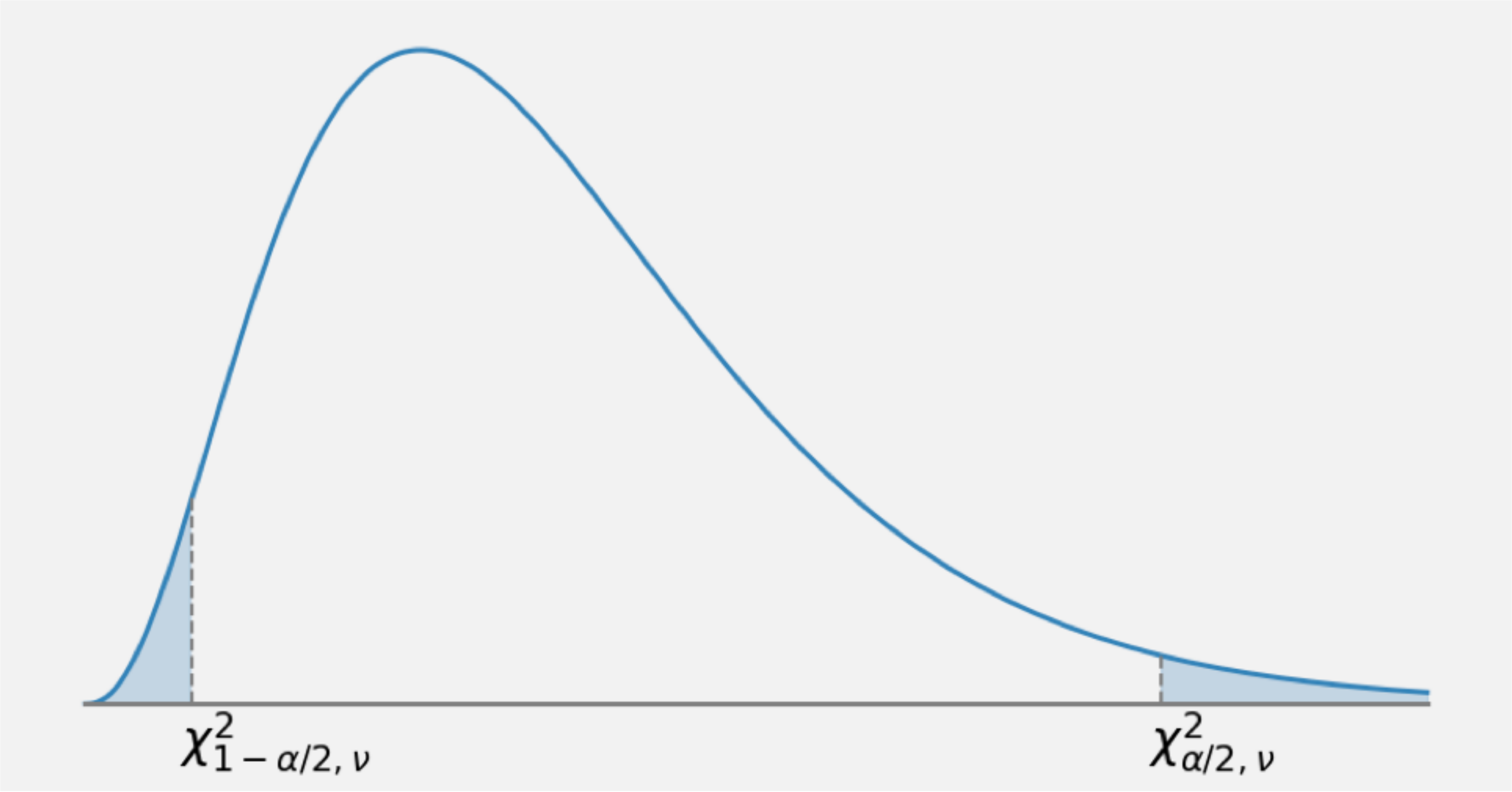
- Then the random variable $(n-1)S^2/\sigma^2$ follows the distribution χ^2_{n-1} .

- Then it follows that

The Chi-Squared Dist is Non-Symmetric

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





A confidence interval for the variance

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

A confidence interval for the variance

- For a $100(1-\alpha)\%$ confidence interval we choose the two critical values $X^2_{1-\alpha/2,n-1}$ and $X^2_{\alpha,n-1}$ 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?

- 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.