intro to data science with probability & statistics

CSCI 3022

Lecture 19
Nov 2, 2018

Hypothesis testing for variance or SD
The Bootstrap

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Inference for *variances*

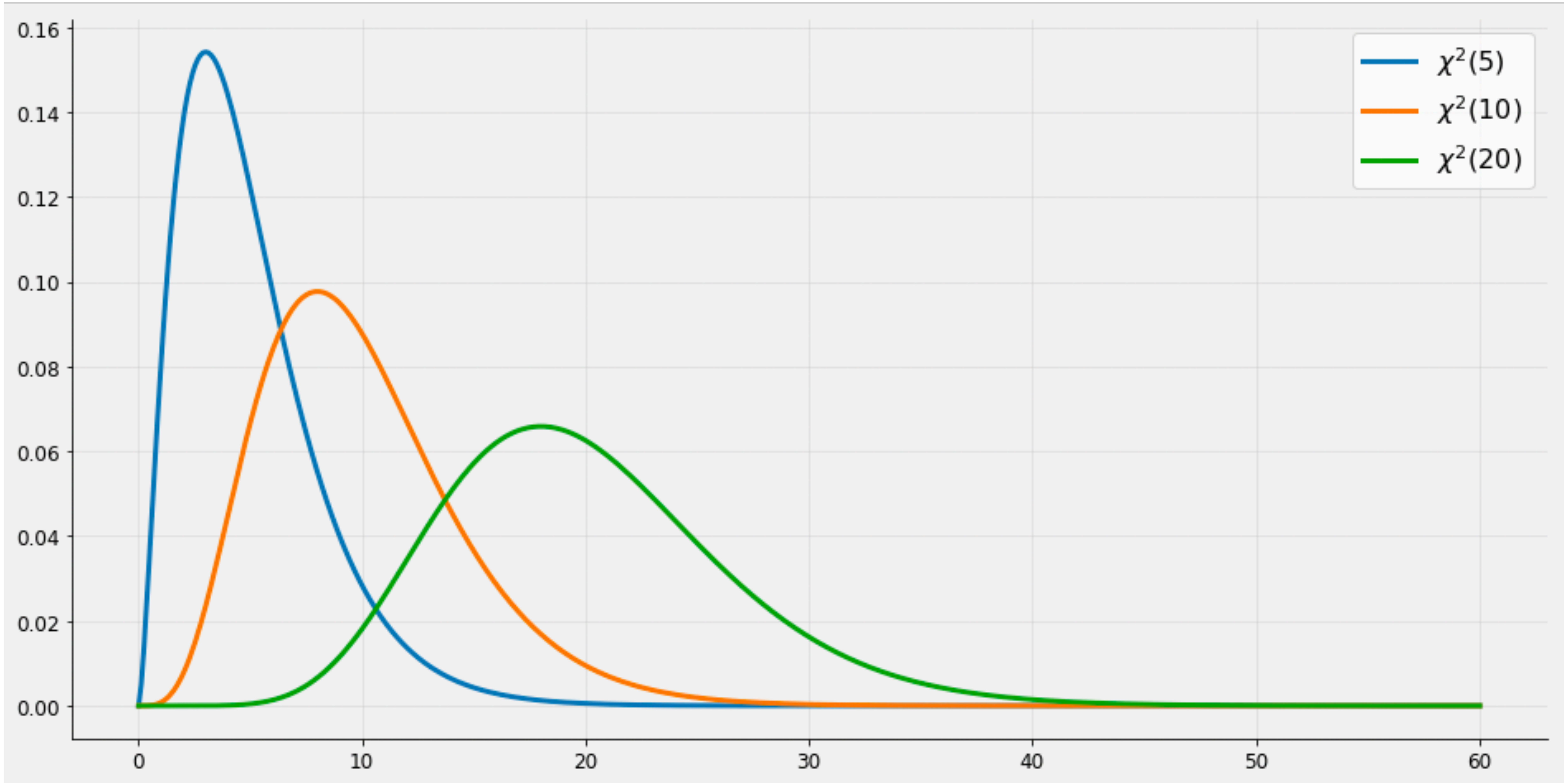
- After Spring Break, we'll talk about estimating confidence intervals for the variance of a population using something [wonderful] called **The Bootstrap**.

- 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

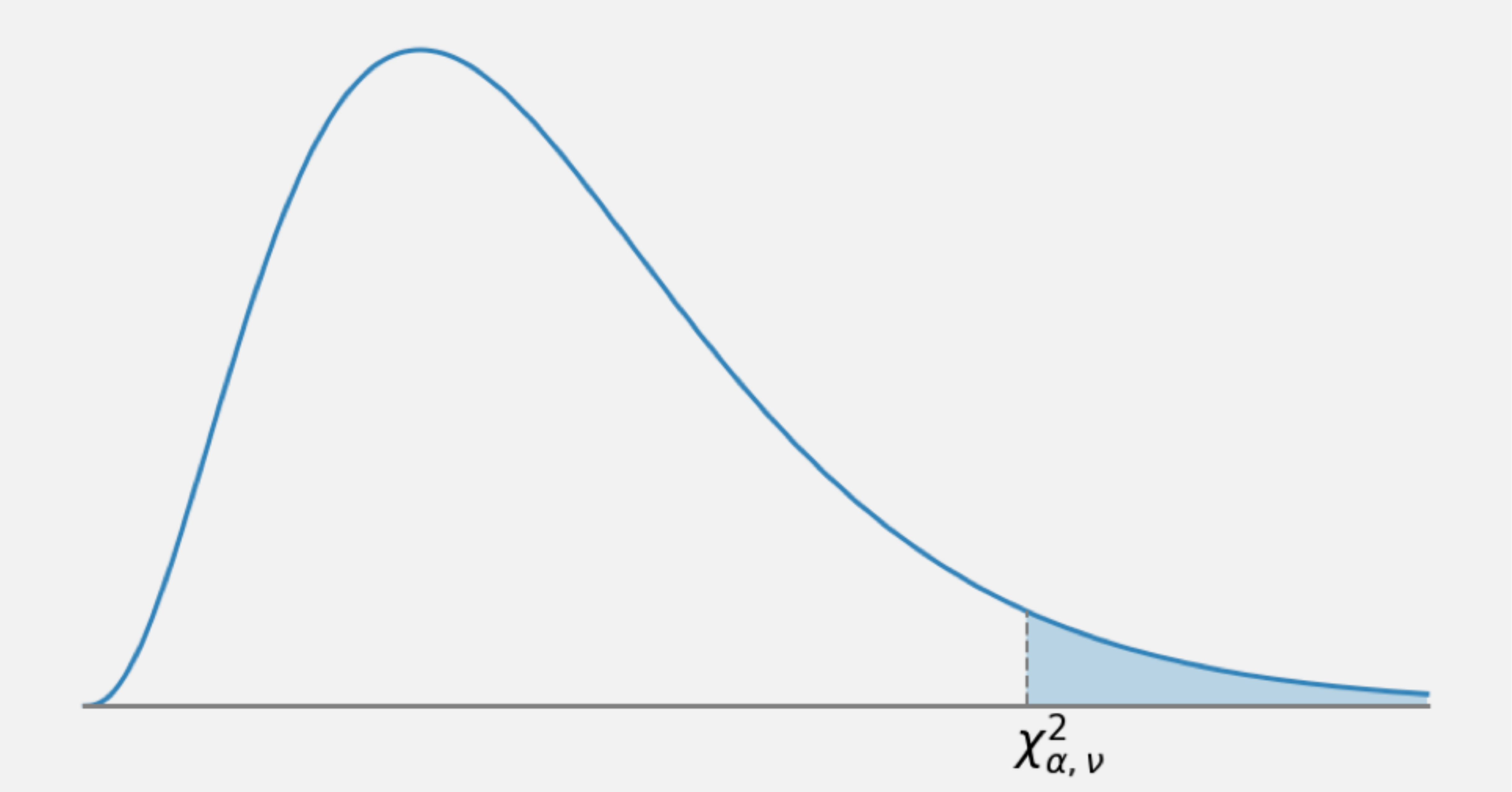
- Let $X_1, X_2, \ldots X_n$ be IID samples from a normal distribution with mean μ and standard deviation σ. Define the *sample variance* in the usual way as

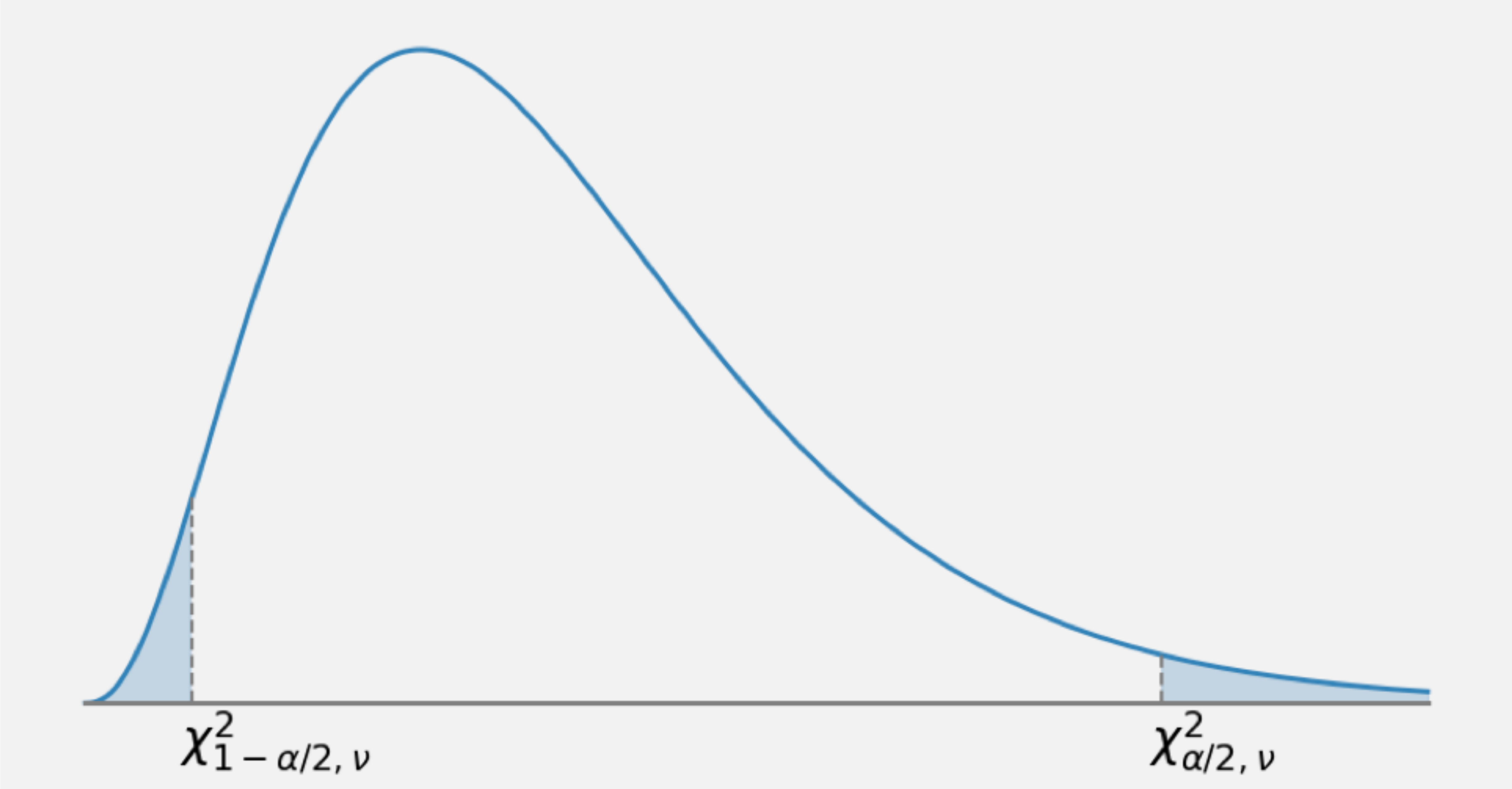
- Then the random variable $(n-1)S^2/\sigma^2$ follows the distribution χ^2_{n-1}.

- 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.

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, $X_1, X_2, \ldots, X_n$, 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*.

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:
  - Resample 2 might be:
  - Resample 3 might be:
- 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.

Bootstrap: why we like it

- The bootstrap for a confidence interval around the mean is convenient, particularly when there are **not enough samples** to use the CLT.
- Of course, if we *can* use the CLT, we should. So why is the bootstrap so exciting?

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		

Bootstrap: why we like it

- The bootstrap for a confidence interval around the mean is convenient, particularly when there are **not enough samples** to use the CLT.
- Of course, if we *can* use the CLT, we should. So why is the bootstrap so exciting?

We can bootstrap CIs for things other than the mean!

- Median.
- Standard Deviation.
- Other statistical measures that we don't have a theory for.

Bootstrap for the median

- Let’s write down the recipe for how we would bootstrap a CI for the median:

Bootstrap for the variance

- Let’s write down the recipe for how we would bootstrap a CI for the variance:

The Non-Parametric Bootstrap

- In the literature—your book, the Wikipedia, etc—you may read about a "non-parametric bootstrap." What is this?

The Non-Parametric Bootstrap

- In the literature—your book, the Wikipedia, etc—you may read about a "non-parametric bootstrap." What is this?
- Let's decode this word, "non-parametric"
- **Definition**: *parametric statistics* assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters.
- Can you name some **examples** of distributions with parameters?
- Can you name a *non*-parametric distribution we've talked about in class?

The Parametric Bootstrap

- We call the bootstrap discussed in class today the non-parametric bootstrap because it doesn't assume any parametric distribution. What you resample is what you get.
- **Definition**: the parametric bootstrap estimates a CI for a desired property in two steps: (1) repeatedly estimate the parameter(s) of the known distribution, and then (2) compute a CI for the desired property by sampling from the known known distribution using the parameters that you inferred.

The Parametric Bootstrap

- We call the bootstrap discussed in class today the non-parametric bootstrap because it doesn't assume any parametric distribution. What you resample is what you get.
- **Definition**: the parametric bootstrap estimates a CI for a desired property in two steps: (1) repeatedly estimate the parameter(s) of the known distribution, and then (2) compute a CI for the desired property by sampling from the known known distribution using the parameters that you inferred.
- **Why**? The parametric bootstrap can be shown to do a better job than the non-parametric bootstrap in various scenarios.
- Why not use the parametric bootstrap all the time?