Confidence Intervals (One Sample)

8.1, 8.2.1, 8.2.2, 8.4.1

Learning Objectives

At the end of this unit, students should be able to:

1. Derive the X% confidence interval for the mean of a population when the sample standard deviation is known.

2. Identify the center and the length of a confidence interval for the mean.

3. Properly interpret confidence intervals and identify common misinterpretations.

4. Compute confidence intervals for the mean of a variable of interest in a given dataset.

5. Find the sample size necessary to ensure that a resulting confidence interval has width of at most *w*.

6. Derive the X% confidence interval for the mean of a population when the sample standard deviation is unknown.

7. Describe why the (pre-calculated) confidence interval endpoints are random (for all CIs).

8. Describe the t-distribution, its properties, its parameter, and its use in confidence interval calculations. Use R to calculate critical values/percentiles of a t-distribution.

9. State the normal approximation of a binomial distribution.

10. Derive, calculate, and interpret the X% confidence interval for a population proportion.

11. Derive, calculate, and interpret the X% confidence interval for a population variance.

$\bar{x} \to \mu \leftarrow$ pop. parameter

sample statistic ("estimator")

Confidence Interval for the Mean (SD known)

Let's start with a simple example. Suppose that we have a simple random sample of *n* measurements from a normal population, and that the population standard deviation is known.

Standardizing the sample mean by first subtracting its expected value and then dividing by its standard deviation yields the standard normal variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

How big does our sample need to be if the underlying population is normally distributed? $n \geq 1$

pnorm(1.96) - pnorm(-1.96) = 0.95

Confidence Interval for the Mean (SD known)

Because the area under the standard normal curve between –1.96 and 1.96 is 0.95, we know:

$$P(-1.96 \le Z \le 1.96) = 0.95 \iff P\left(-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95$$

$$\iff P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Updownarrow$$

This is equivalent to:

$$P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

random

fixed

Confidence Interval for the Mean (SD known)

The interval

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

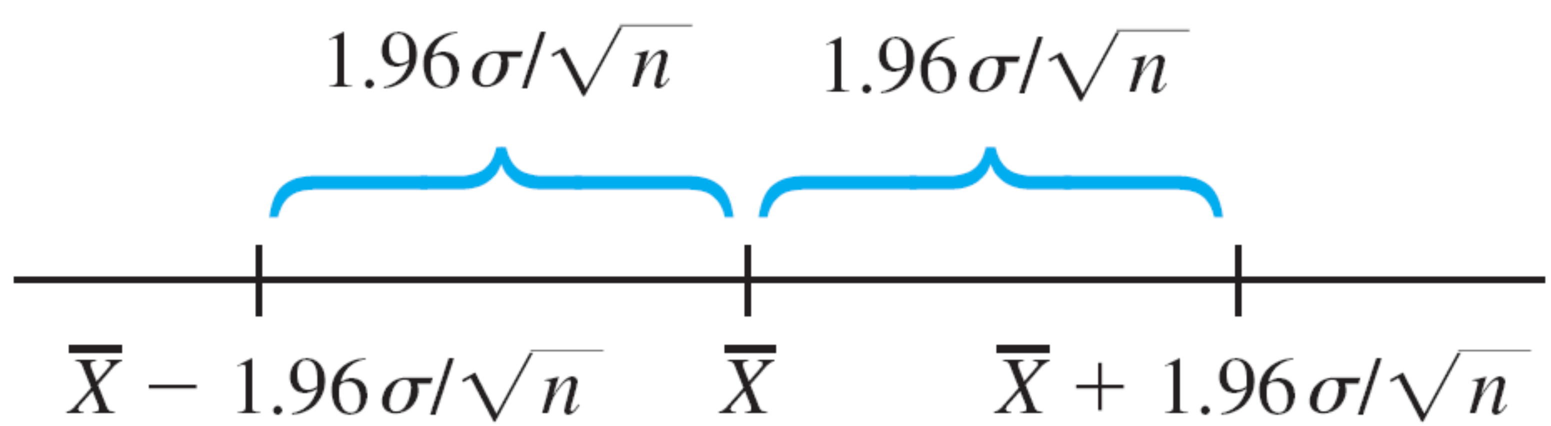
Is called the 95% confidence interval for the mean.

This interval varies from sample to sample, as the sample mean varies. So, the interval itself is a random interval.

Confidence Interval for the Mean (SD known)

The CI interval is centered at $\underline{\quad \bar{X} \quad}$ and extends $\underline{\quad 1.96 \frac{\sigma}{\sqrt{n}} \quad}$ to each side of $\bar{X}$.

The interval's width is $\underline{\quad 2\left(1.96 \frac{\sigma}{\sqrt{n}}\right) \quad}$ which is not random; only the location of the interval (its midpoint X) is random.



Confidence Interval for the Mean (SD known)

As we showed, for a given sample, the CI can be expressed as

" "

A concise expression for the interval is

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

where the left endpoint is the lower limit and the right endpoint is the upper limit.

Interpreting a Confidence Interval

$(3, 5)$

"We are 95% confident that the true parameter is in this interval."

What does that mean??

If we repeatedly took samples and calc. CIs, 95% would cover μ.

Interpreting a Confidence Interval

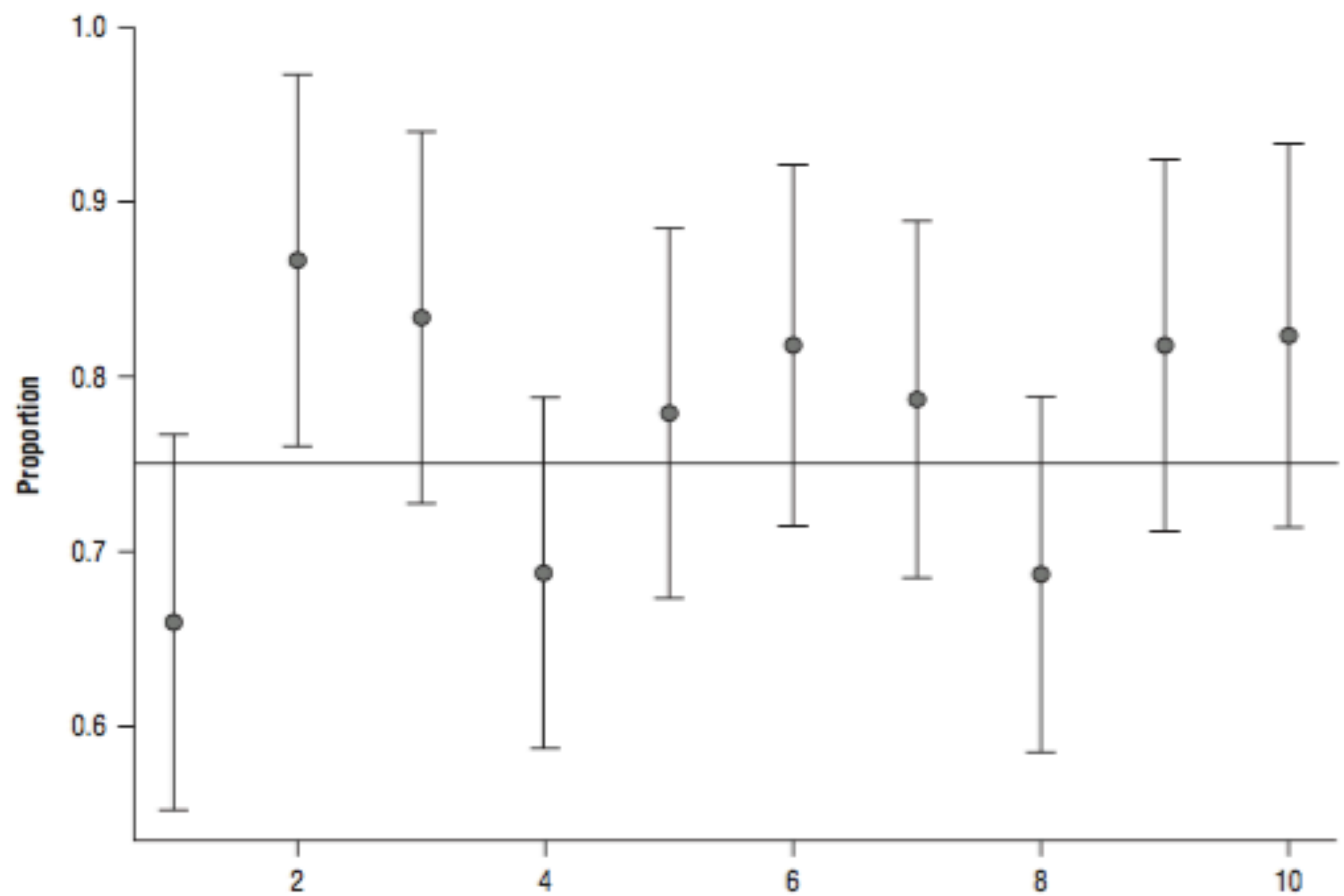
A correct interpretation of “95% confidence” relies on the long-run relative frequency interpretation of probability.

In repeated sampling, 95% of the confidence intervals obtained from all samples will actually contain μ. The other 5% of the intervals will not.

The confidence level is *not a statement about any particular interval* instead it pertains to what would happen if a very large number of like intervals were to be constructed using the same CI formula.

Interpreting a Confidence Interval

Figure 1: Confidence Interval



Note: Suppose that the true proportion of believers in climate change among French citizens is 0.75, as represented by the horizontal black line near the middle. This figure shows ten 90% confidence intervals used to estimate the proportion of believers in climate change among French citizens. Each circle represents a point estimate, p, calculated from a different sample of n French citizens; for each confidence interval, the length of the vertical line is twice the margin of error, E, for that interval. Notice that the second interval fails to cover the true proportion. For the 90% confidence interval procedure, it is expected that about one in every ten intervals will fail to cover the true proportion.

Interpreting a Confidence Interval

Show them this! http://www.ejwagenmakers.com/inpress/HoekstraEtAlPBR.pdf

Other Levels of Confidence

significance level (in hypothesis Testing)

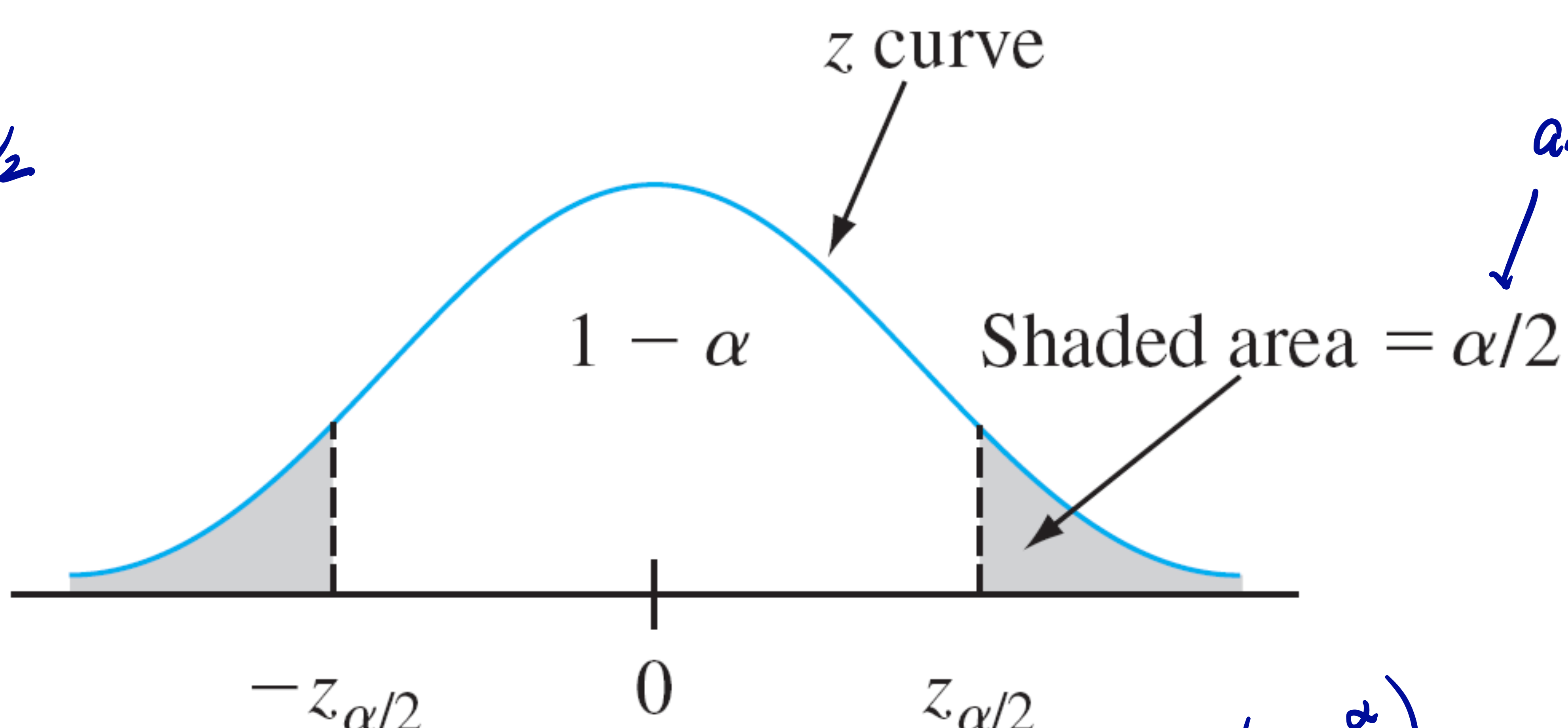
Probability of 1 − α is achieved by using $\boxed{z_{\alpha/2}}$ in place of

$z_{.025} = 1.96$

$= z_{\alpha/2}$

$= z_{0.05/2}$

area



qnorm(1 - α/2)

Other Levels of Confidence

A $\boxed{\textbf{100(1} - \alpha\textbf{)\%}}$ **confidence interval** for the mean when the value of σ is known is given by:

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right)$$

Or, equivalently, by:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Confidence Interval for the Mean (SD Known)

Example: A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm. $\boxed{n=40}$. $\boxed{\bar{x}=5.426}$. $\boxed{\sigma=0.1}$

(a) Calculate a confidence interval for true average hole diameter using a confidence level of 90%.

$Z_{\alpha/2} = Z_{0.1/2} = Z_{0.05} = 1.64.$

$\left(5.426 - 1.64\left(\frac{0.1}{\sqrt{40}}\right), 5.46 + 1.64\left(\frac{0.1}{\sqrt{40}}\right)\right)$

$= (5.4, 5.452)$

(b) What about the 99% confidence interval?

$Z_{\alpha/2} = 2.57$

$(5.385, 5.467)$

(c) What are the advantages and disadvantages to a wider confidence interval?

more conf.

less precision

Sample Size Computation

For a desired confidence level and interval width, we can determine the necessary sample size.

Example: For a given computer model, memory fetch response time is normally distributed with standard deviation of 25 milliseconds. A new computer has been purchased, and we wish to estimate the true average response time. What sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 units?

$$10 = w = 2(1.96)\left(\frac{25}{\sqrt{n}}\right) \Rightarrow n \approx 96.04 \Rightarrow n \geq 97$$

Large Sample Confidence Interval for the Mean

A difficulty in using our previous equation for confidence intervals is that it uses the value of σ which will <u>rarely be known</u>. Also, we may want a CI for a mean from <u>some other non-normal distribution</u>.

$E(s) = \sigma$ $P\left(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right) = 1-\alpha$, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Large Sample Confidence Interval for the Mean

In this instance, we need to work with the **sample standard deviation** *s*. Remember from our first lesson that the standard deviation is calculated as:

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

With this, we instead work with the standardized random variable:

approx

$$Z = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$$

Confidence Interval for the Mean (SD Unknown)

Previously, there was randomness only in the numerator of *Z* by virtue of the estimator ___$\bar{x}$___.

In the new standardized variable, both ___$\bar{x}$___ *and* ___s___ vary in value from one sample to another.

When *n* is large, the substitution of *s* for σ adds little extra variability, so nothing needs to change.

When *n* is smaller, the distribution of this new variable should be wider than the normal to reflect the extra uncertainty. (We talk more about this in a bit.)

$z_{\alpha/2}$

Confidence Interval for the Mean (SD Unknown)

Large Sample CI:

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha$$

If *n* is sufficiently large *(n* ≥*30)*, the standardized random variable

$$Z = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$$

has approximately a standard normal distribution. This implies that

$$\left(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

is a large-sample confidence interval for with confidence level approximately 100(1 – α)%. This formula is valid regardless of the population distribution for sufficiently large *n*.

	n >= 30	n < 30
pop. Underlying normal distribution	σ known	σ known
	σ unknown (previous slide)	σ unknown
pop Underlying non-normal distribution	σ known	σ known
	σ unknown (previous slide)	σ unknown

	n >= 30	n < 30
Underlying normal distribution	**σ known**	**σ known**
	σ unknown	**σ unknown**
Underlying non-normal distribution	**σ known**	**σ known**
	σ unknown	**σ unknown**

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown
Underlying non-normal distribution	σ known	σ known
	σ unknown	σ unknown

Small Sample Interval for the Mean

The CLT cannot be invoked when *n* is small, and we need to do something else when $n < 30$.

When $n < 30$ and the underlying distribution is normal, we have a solution!

t-Distribution

The results on which large sample inferences are based introduces a new family of probability distributions called *t distributions.*

When $\underline{\bar{x}}$ is the mean of a random sample of size *n* (≤ 30) from a **normal distribution** with mean $\underline{\mu}$, the random variable

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

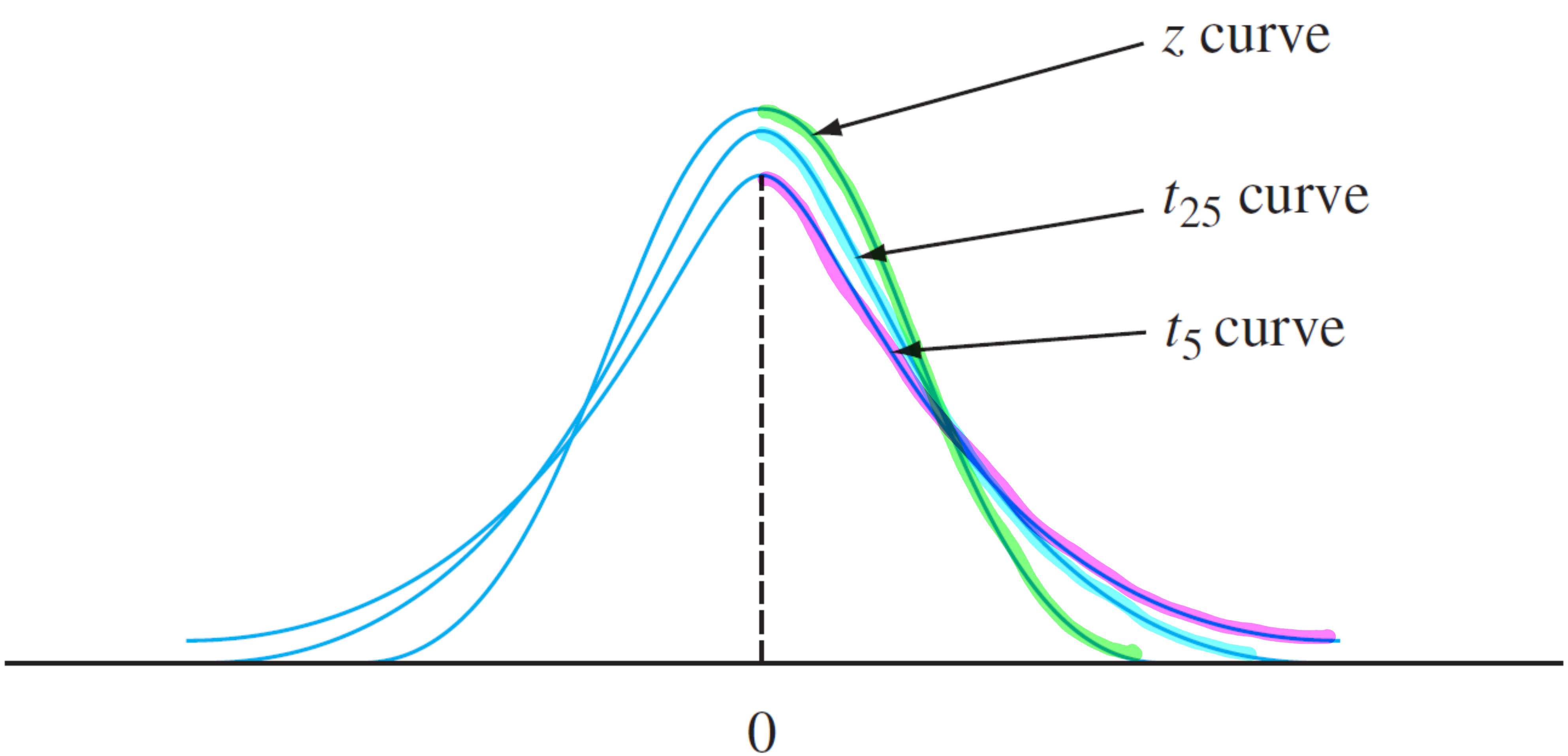
has a probability distribution called a *t* Distribution with *n*–1 degrees of freedom (df).

$P(T \le t) = pt(t, df)$

$v = n - 1$

t-Distribution

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



Properties of the t-Distribution

Let t_ν denote the t distribution with ν df.

1. Each t_ν curve is bell-shaped and centered at 0.

2. Each t_ν curve is more spread out than the standard normal (z) curve.

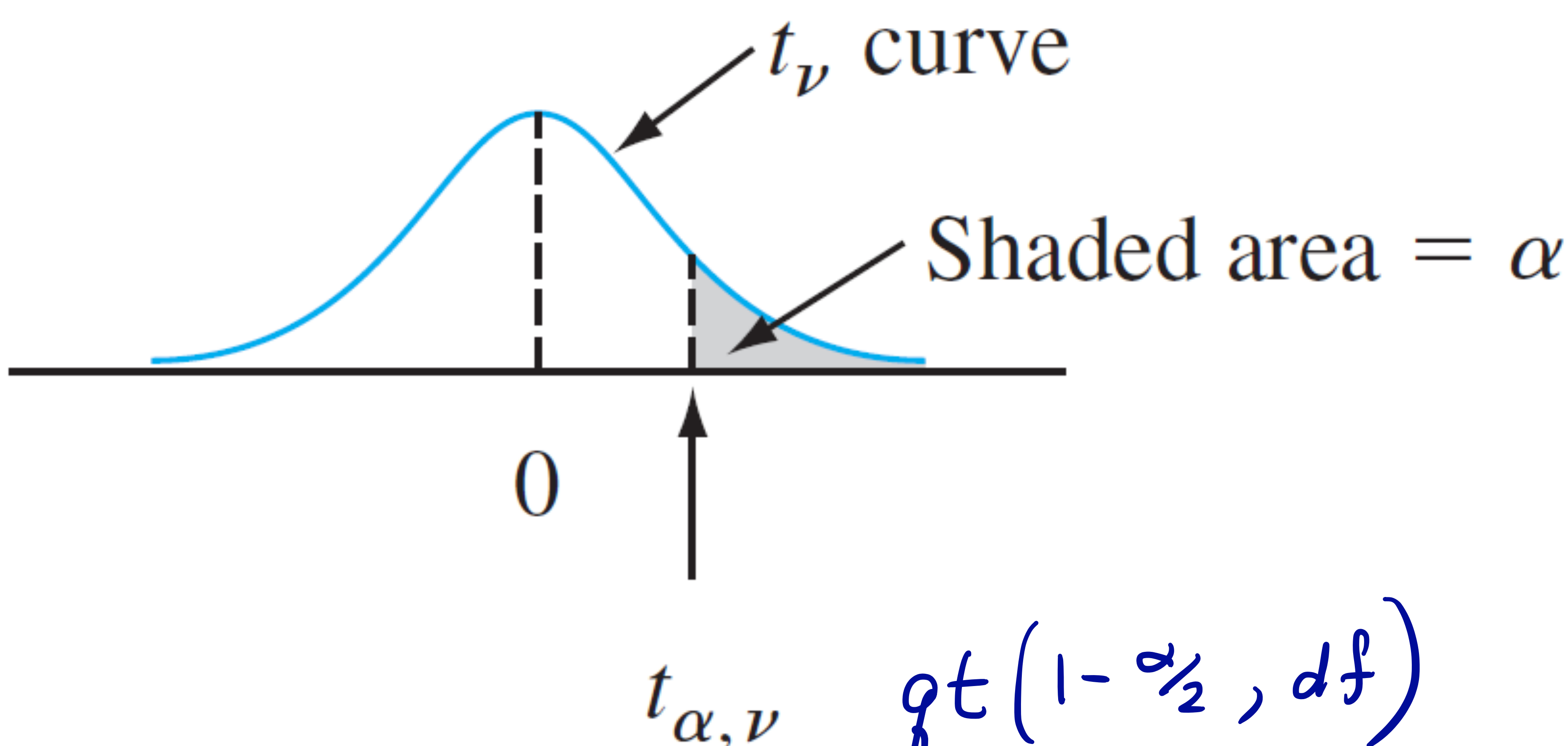
3. As ν (df) increases, the spread of the corresponding t_ν curve decreases.

4. As $\nu \to$ ___∞___ the sequence of t_ν curves approaches the standard normal curve (so the z curve is the t curve with df = ___∞___)

$z_{\alpha/2}$

Properties of the t-Distribution

Let $t_{\alpha,\nu}$ = the number on the measurement axis for which the area under the t curve with ν df to the right of $t_{\alpha,\nu}$ is ; $t_{\alpha,\nu}$ is called a ***t* critical value**.



qt(1 - α/2 , df)

For example, $t_{.05,6}$ is the t critical value that captures an upper-tail area of .05 under the t curve with 6 df.

Finding t-Values

The probabilities of *t* curves are found in a similar way as the normal curve.

0.05

df

Example: obtain $t_{.05,15}$

$=$

2.131

$s = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$

The t-Confidence Interval

Let ___$\bar{x}$___ and ___s___ be the sample mean and sample standard deviation computed from the results of a random sample from a <u>normal</u> <u>population</u> with mean μ. Then a $100(1-\alpha)\%$ t-confidence interval for the mean μ is

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

or, more compactly:

The t-Confidence Interval

Example: Suppose that the GPA measurements for 23 students follow a normal distribution. The sample mean is 3.146. The sample standard deviation is 0.308. Calculate a 90% CI for the mean GPA.

$= s$ $\quad \bar{x}$

$\alpha = 0.1$

$t_{\alpha/2, n-1} = t_{0.05, 22} = 1.717$

$CI: \quad 3.146 \pm 1.717 \left(\frac{0.308}{\sqrt{23}} \right) = (3.036, 3.256)$

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown ✓
Underlying non-normal distribution	σ known	σ known
	σ unknown	σ unknown

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown
Underlying non-normal distribution	σ known	σ known
	σ unknown	σ unknown

Special Cases

Special Cases

When $n < 30$ and the underlying distribution is unknown, we have to:

1. Make a specific assumption about the form of the population distribution and derive a CI based on that assumption.

2. Use other methods (such as bootstrapping) to make reasonable confidence intervals.

Confidence Interval for Population Proportion

parameter

Let p denote the proportion of "successes" in a population (e.g., individuals who graduated from college, computers that do not need warranty service, etc.). A random sample of n individuals is selected, and X is the number of successes in the sample.

Then, X can be modeled as a **Binomial rv** with mean np and

$$Var(X) = np(1-p)$$

If both $np \geq 10$ and $n(1-p) \geq 10$, X has approximately a normal distribution

$$X \underset{\text{approx}}{\sim} N(np, np(1-p))$$

$s^2 \to \sigma^2$

$s \to \sigma$

$E(s^2) = \sigma^2$

$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha$

Confidence Interval for Population Proportion

The estimator of *p* is: $\hat{p} = \frac{X}{n}$

This estimator is approximately normally distributed and:

$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}\cdot np = p$

$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

Standardizing the estimator yields:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \quad , \text{ for } np \ge 10$$

Thus, the CI is: $\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \, , \; \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

An approx. $(1-\alpha)\times 100\%$ CI

Confidence Interval for Population Proportion

Example: The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 (63.5%) of these sampled households to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportional of homes with indoor radon levels above 4 pCi/L.

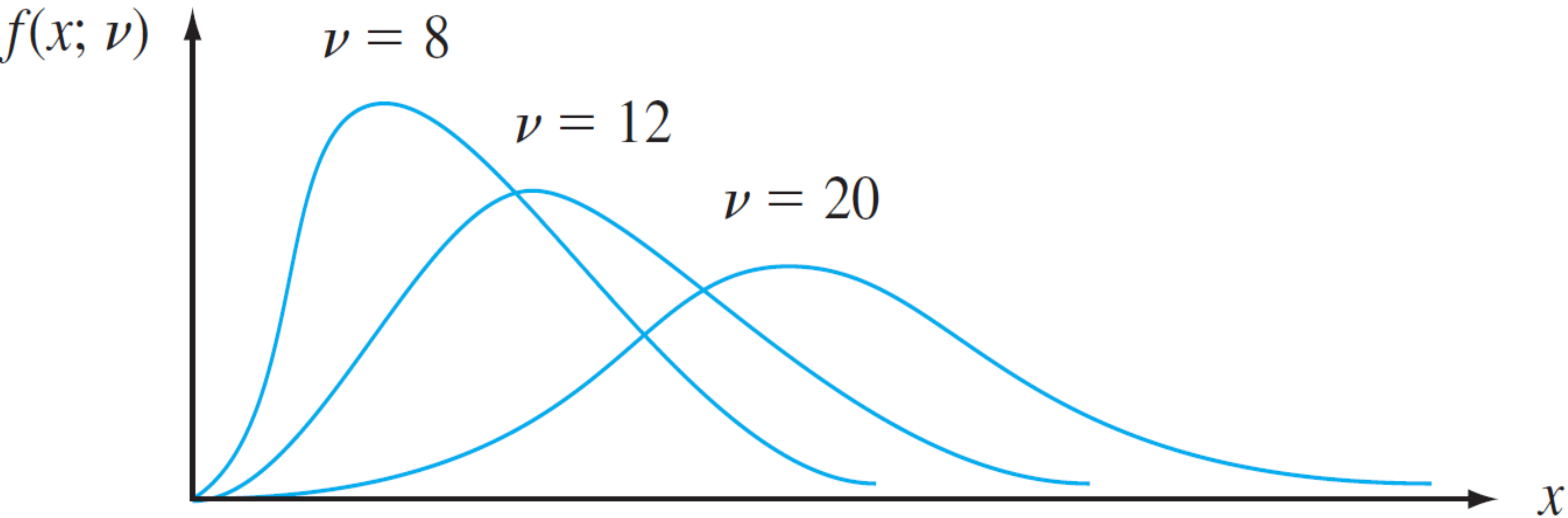
The Chi-Squared Distribution

Definition: Let *v* be a positive integer. The random variable *X* has a **chi-squared distribution** with parameter *v* if the pdf of *X*

$$f(x; \nu) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The parameter is called the **number of degrees of freedom** (df) of *X*. The symbol χ^2 is often used in place of "chi-squared."

The Chi-Squared Distribution



Confidence Intervals for Variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from a **normal distribution** with parameters μ and σ^2. Then the r.v.

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

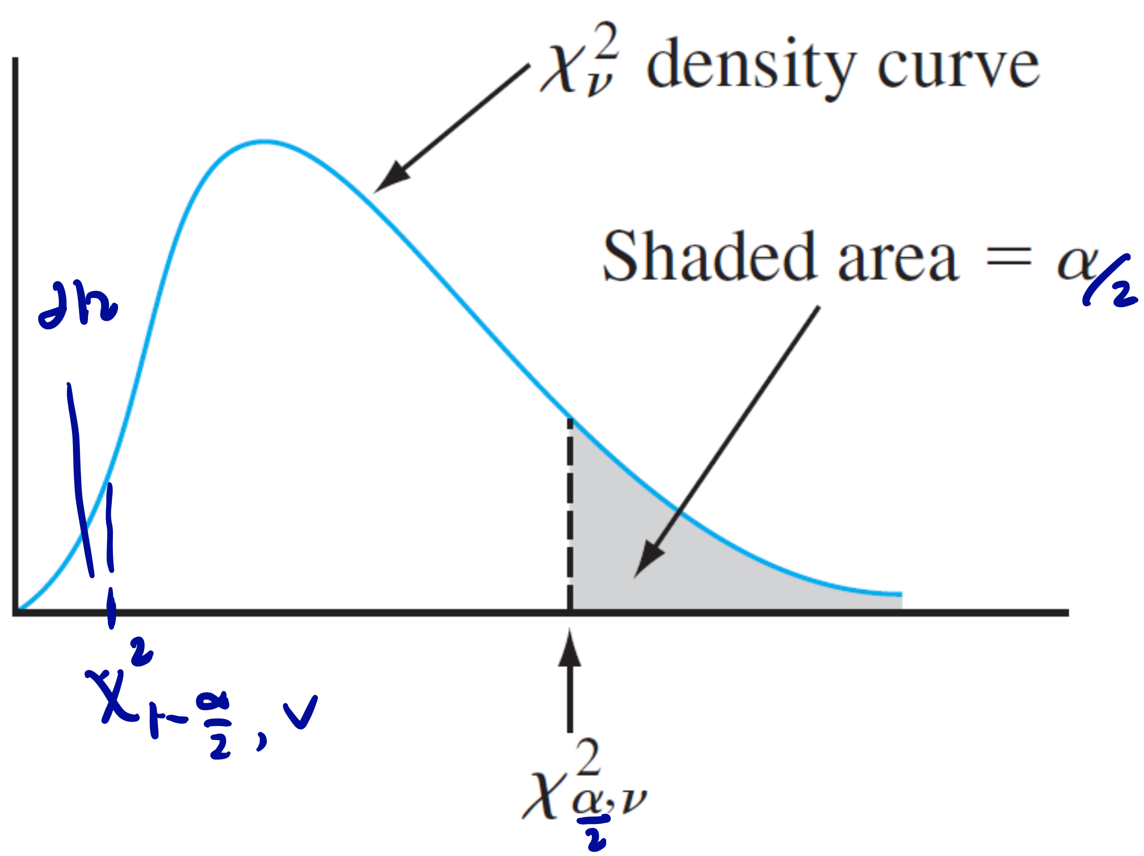
has a chi-squared (__χ^2__) probability distribution with $n - 1$ df.

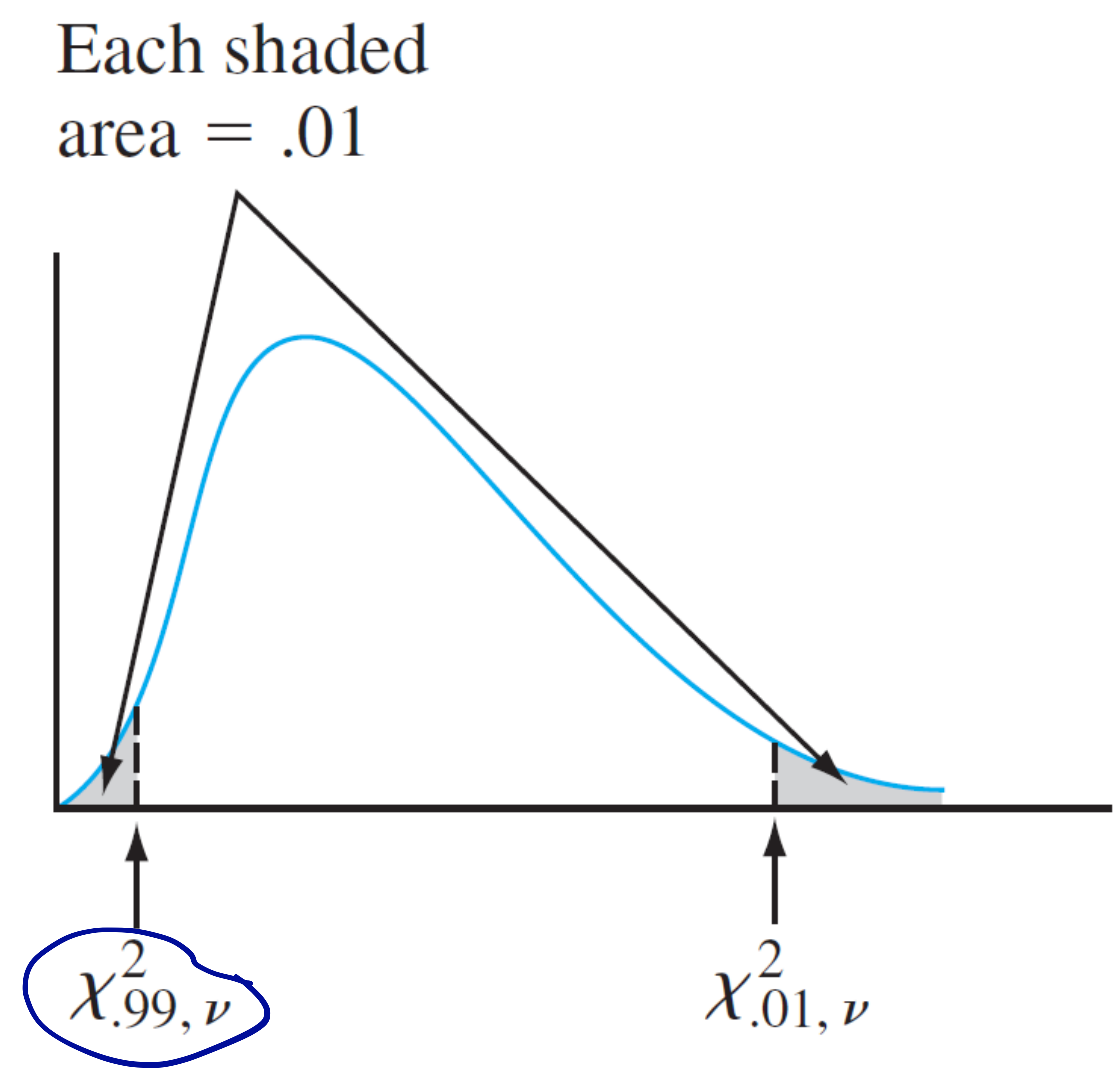
(In this class, we don't consider the case where the data is not normally distributed.)

CL : $(1-\alpha)\times 100\%$

Confidence Intervals for Variance

The chi-squared distribution is *not symmetric*, so these tables contain values of $\underline{\ \chi^2\ }$ both for α near 0 and 1.





Confidence Intervals for Variance

As a consequence:

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

$$\vdots$$

Or, equivalently: The $(1-\alpha)\times 100\%$ CI is

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

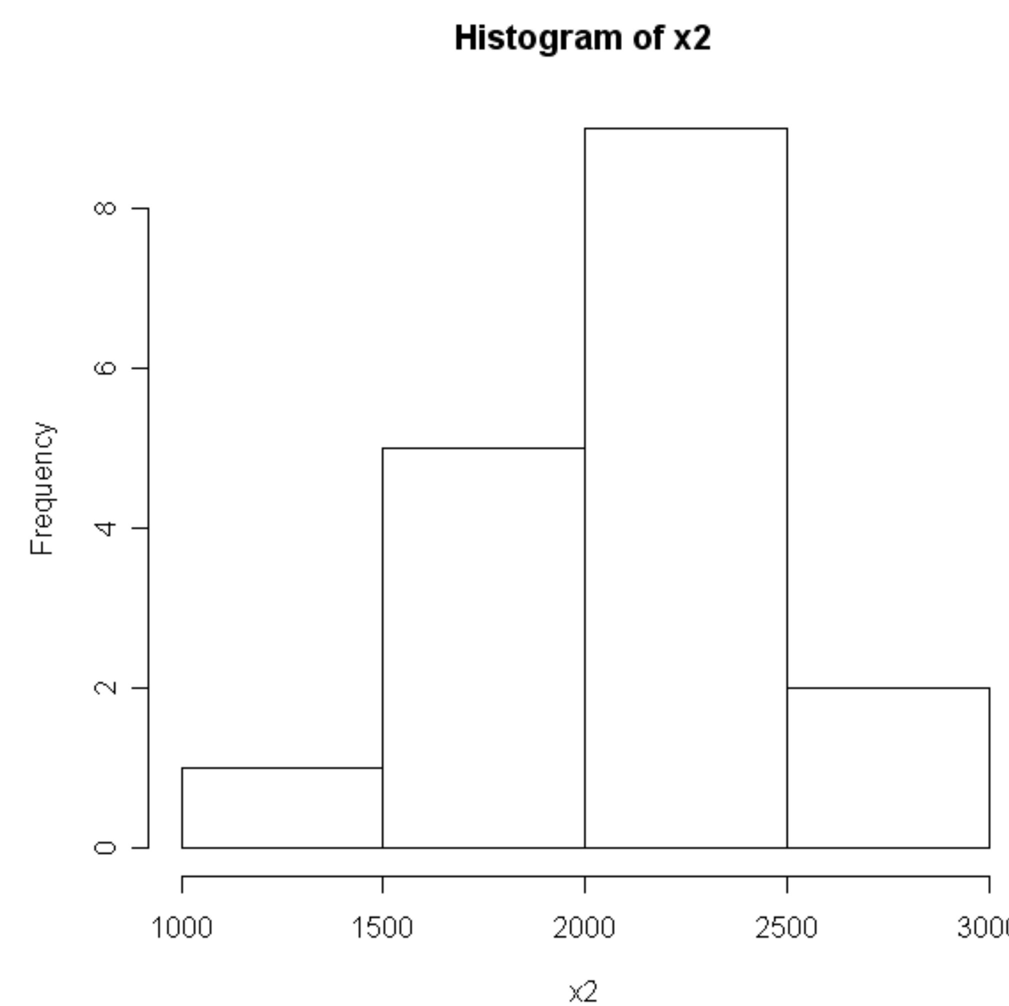
Thus we have a confidence interval for the variance.Taking square roots gives a CI for the standard deviation.

Confidence Intervals for Variance

The data on breakdown voltage of electrically stressed circuits are shown below.

The breakdown voltage is approximately normally distributed. $s^2 = 137{,}324.3$ and

n = 17



(76,171, 318,079) ← 95% CI for σ^2