7

Statistical Intervals (One sample)

(Chs 8.1-8.3)

Confidence Intervals

The CLT tells us that as the sample size *n* increases, the sample mean $\overline{X}$ is close to normally distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$.

Standardizing $\overline{X}$ by first subtracting its expected value and then dividing by its standard deviation yields the standard normal variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

How big does our sample need to be if the underlying population is normally distributed?

Confidence Intervals

Because the area under the standard normal curve between –1.96 and 1.96 is .95, we know:

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

This is equivalent to:

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

Confidence Intervals

The interval

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

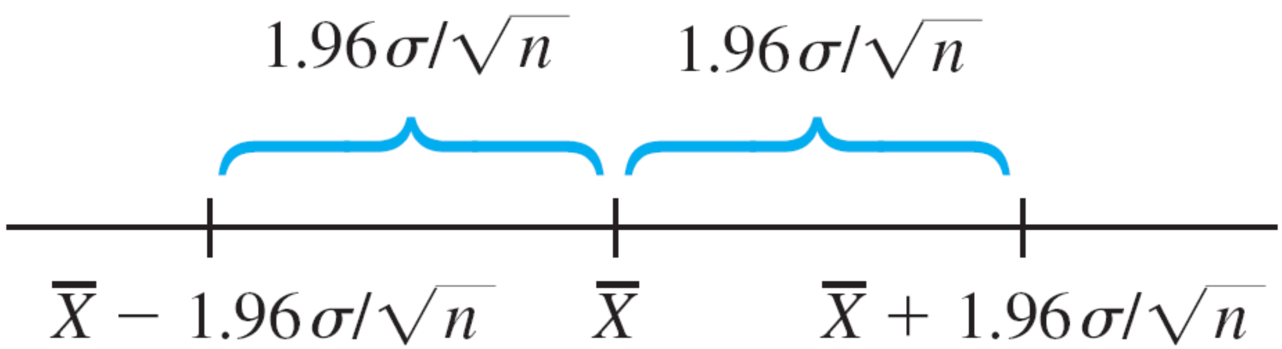
Is called the **95% confidence interval for the mean.**

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

Confidence Intervals

The CI interval is centered at the sample mean $\overline{X}$ and extends 1.96 $\sigma/\sqrt{n}$ to each side of $\overline{X}$.

The interval's width is 2 ⋅ (1.96) ⋅$\sigma/\sqrt{n}$, which is not random; only the location of the interval (its midpoint $\overline{X}$) is random.



Confidence Intervals

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

$$\left(\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

A concise expression for the interval is

$$\overline{x} \pm 1.96 \cdot \sigma/\sqrt{n}$$

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

Interpreting a Confidence Level

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

What does that mean??

Interpreting a Confidence Level

- 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.

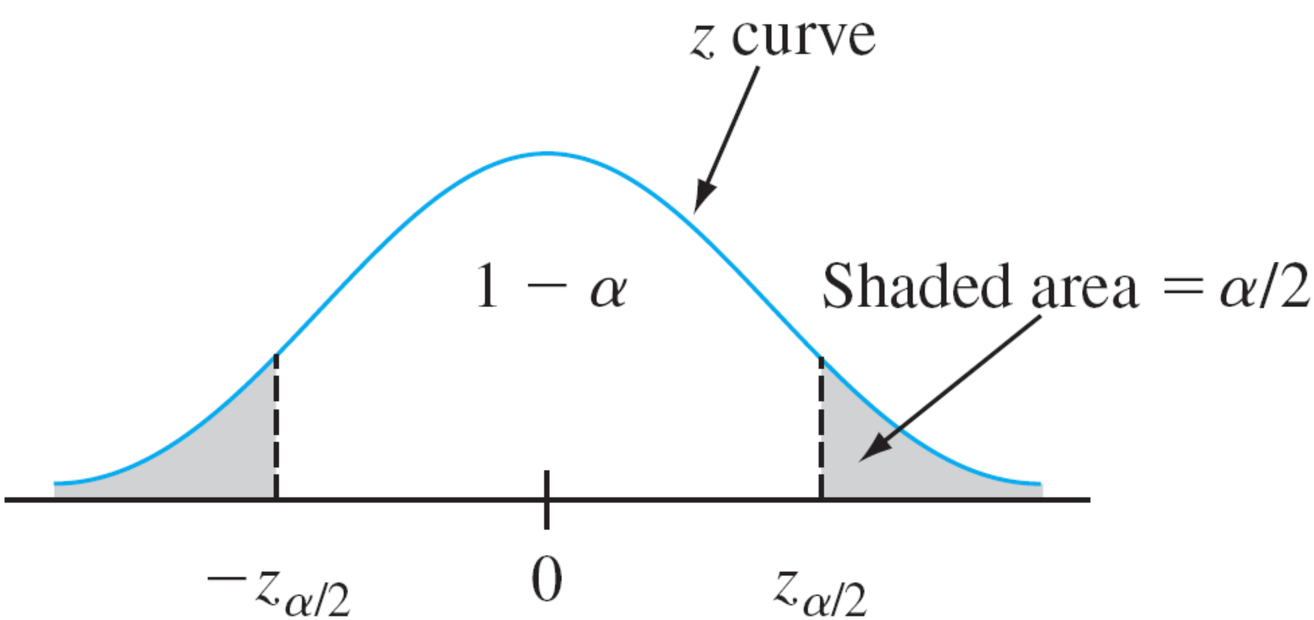
- The confidence level is *not a statement about any particular interval* instead it pertains to what would happen if a very large number of like intervals were to be constructed using the same CI formula.

Interpreting a Confidence Level

Demonstration through simulations

Other Levels of Confidence

Probability of $1 - \alpha$ is achieved by using $z_{\alpha/2}$ in place of $z_{.025} = 1.96$



$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha \text{ where } Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Other Levels of Confidence

A **100(1 – α)% confidence interval** for the mean μ when the value of σ is known is given by

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

or, equivalently, by $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

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.

Let’s calculate a confidence interval for true average hole diameter using a confidence level of 90%.

What about the 99% confidence interval?

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

Sample size computation

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

<u>Example</u>: A response time is normally distributed with standard deviation of 25 milliseconds. A new system has been installed, and we wish to estimate the true average response time μ for the new environment.

Assuming that response times are still normally distributed with $\sigma = 25$, what sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10?

Unknown variance

A difficulty in using our previous equation for confidence intervals is that it uses the value of σ, which will rarely be known.

Unknown variance

A difficulty in using our previous equation for confidence intervals is that it uses the value of σ, which will rarely be known.

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:

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

Unknown variance

A difficulty in using our previous equation for confidence intervals is that it uses the value of σ, which will rarely be known.

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:

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

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

$$(\bar{X} - \mu)/(S/\sqrt{n})$$

Unknown mean and variance

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

In the new standardized variable, both $\overline{X}$ *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.)

A Large-Sample Interval for μ

<u>**Proposition**</u>

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

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has approximately a <u>standard normal distribution</u>. This implies that

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

is a large-sample confidence interval for μ with confidence level approximately $100(1 - \alpha)\%$.

This formula is valid regardless of the population distribution for sufficiently large *n*.

	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

	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

A Small-Sample Interval for μ

- 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.*

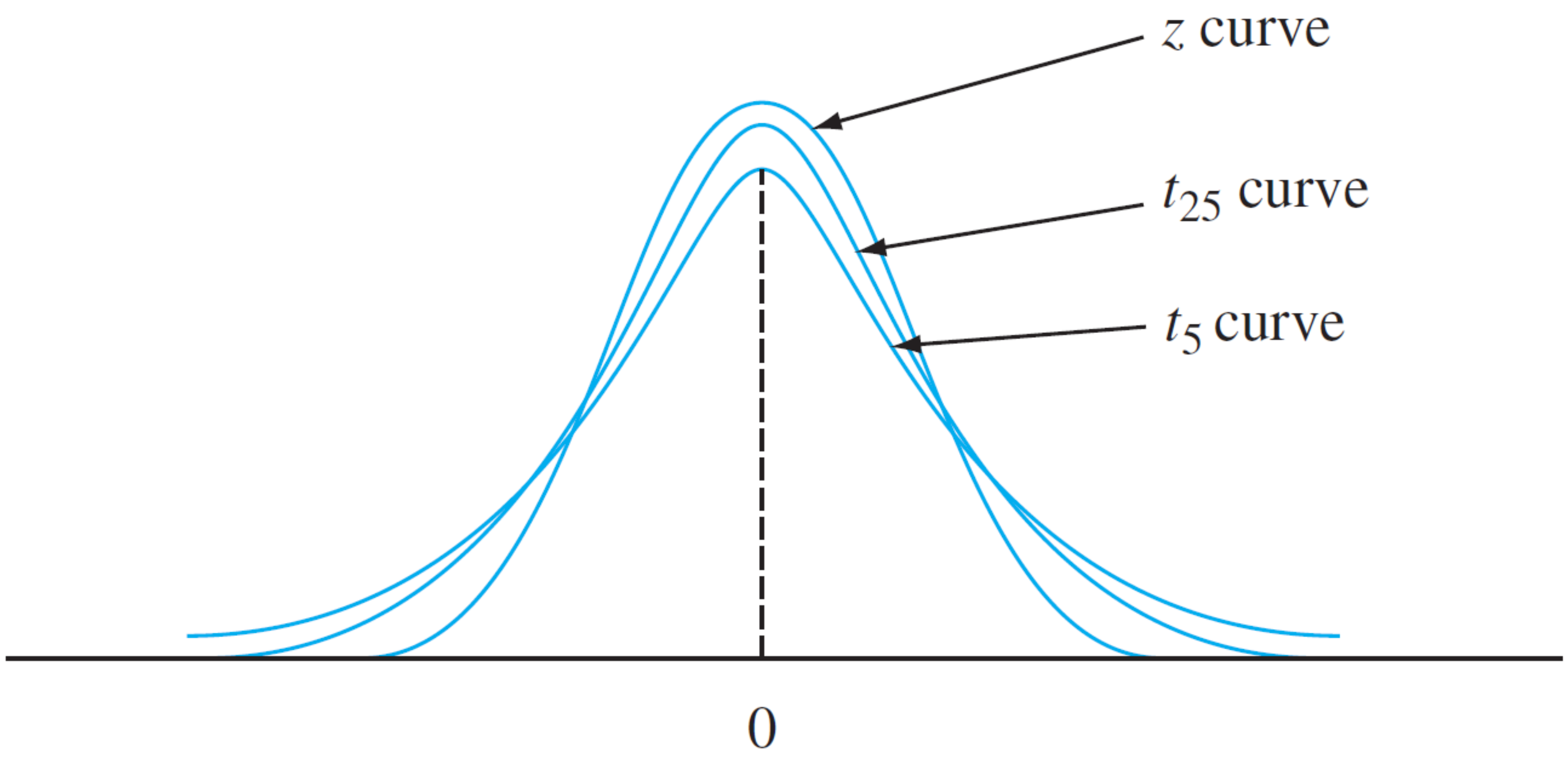
When $\overline{X}$ is the mean of a random sample of size *n* from a **normal distribution** with mean μ, the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

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

Properties of *t* Distributions

Figure below illustrates some members of the t-family



Properties of t Distributions

Properties of *t* Distributions

Let t_ν denote the t distribution with ν df.

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

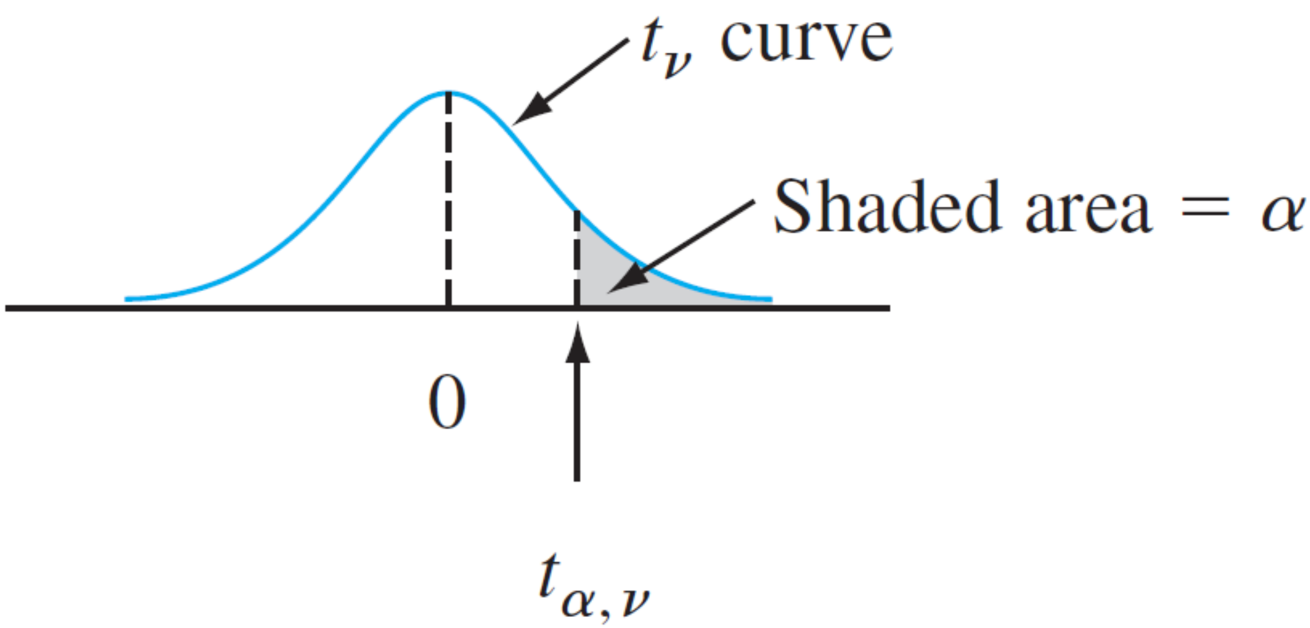
2. Each t_ν curve is more spread out than the standard normal (z) curve.

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

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

Properties of t Distributions

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 α; $t_{\alpha,\nu}$ is called a ***t* critical value**.



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

Tables of t Distributions

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

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

The *t* Confidence Interval

Let $\overline{X}$ and *s* be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ.

Then a 100(1 – α)% *t*-confidence interval for the mean μ is

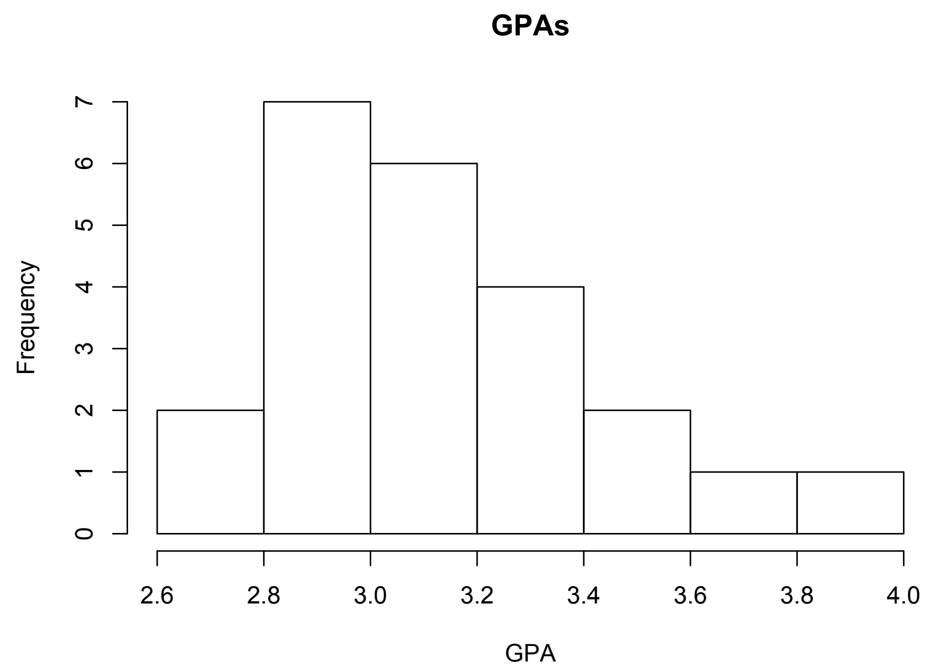
$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

or, more compactly: $\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}.$

Example

cont'd

GPA measurements for 23 students have a histogram that looks like this:



The sample mean is 3.146. The sample standard deviation is 0.308. Calculate a 90% CI for the mean GPA.

Confidence Intervals for μ

	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**

Confidence Intervals for μ

	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

When the t-distribution doesn't apply

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

- Make a specific assumption about the form of the population distribution and derive a CI based on that assumption.
- Use other methods (such as bootstrapping) to make reasonable confidence intervals.

A Confidence Interval for a Population Proportion

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 regarded as a **Binomial rv** with mean np and

$$\sigma_X = \sqrt{np(1 - p)}$$

A Confidence Interval for *p*

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 regarded as a **Binomial rv** with mean *np* and

$$\sigma_X = \sqrt{np(1 - p)}$$

If both *np* ≥ 10 and *n(1-p)* ≥ 10, *X* has approximately a normal distribution.

A Confidence Interval for *p*

The estimator of *p* is $\hat{p} = X / n$ (the fraction of successes).

$\hat{p}$ has approximately a normal distribution, and

$$\sigma_{\hat{p}} = \sqrt{p(1 - p)/n}$$

Standardizing $\hat{p}$ by subtracting *p* and dividing by $\sigma_{\hat{p}}$ then implies that

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

And the CI is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$$

A Confidence Interval for *p*

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.

CIs for the 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.

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \overline{X})^2}{\sigma^2}$$

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.)

The Chi-Squared Distribution

Definition

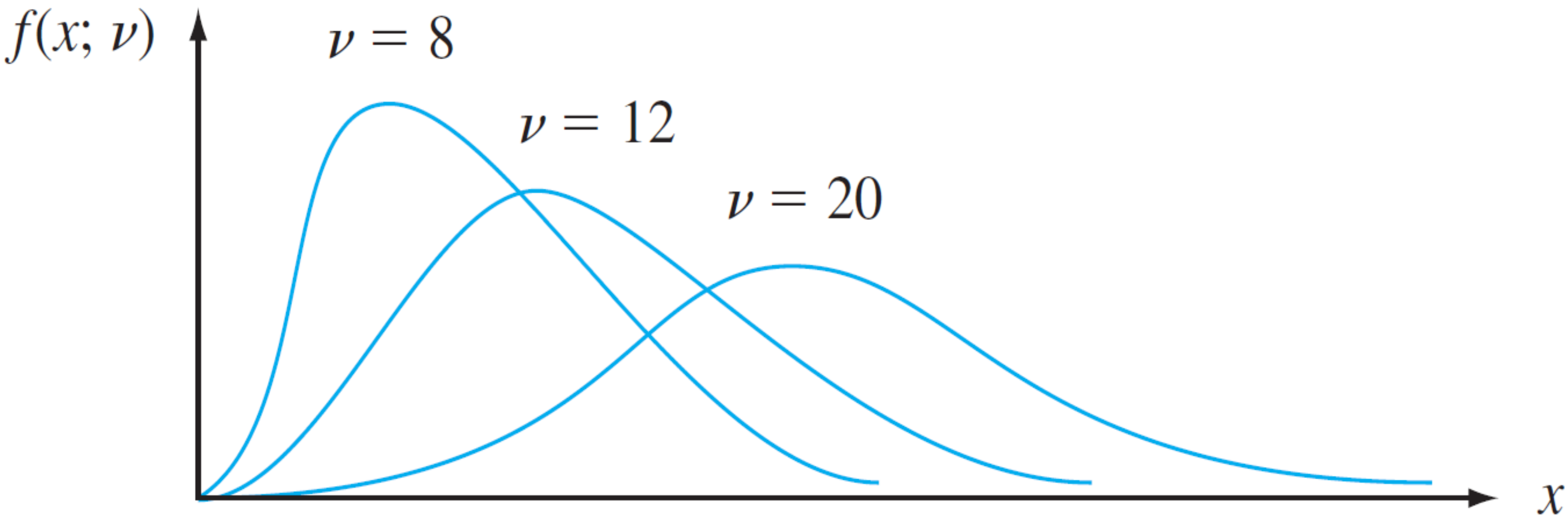
Let *v* be a positive integer. The random variable *X* has a **chi-squared distribution** with parameter *v* if the pdf of *X*

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

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."

CIs for the Variance

The graphs of several Chi-square probability density functions are



CIs for the Variance

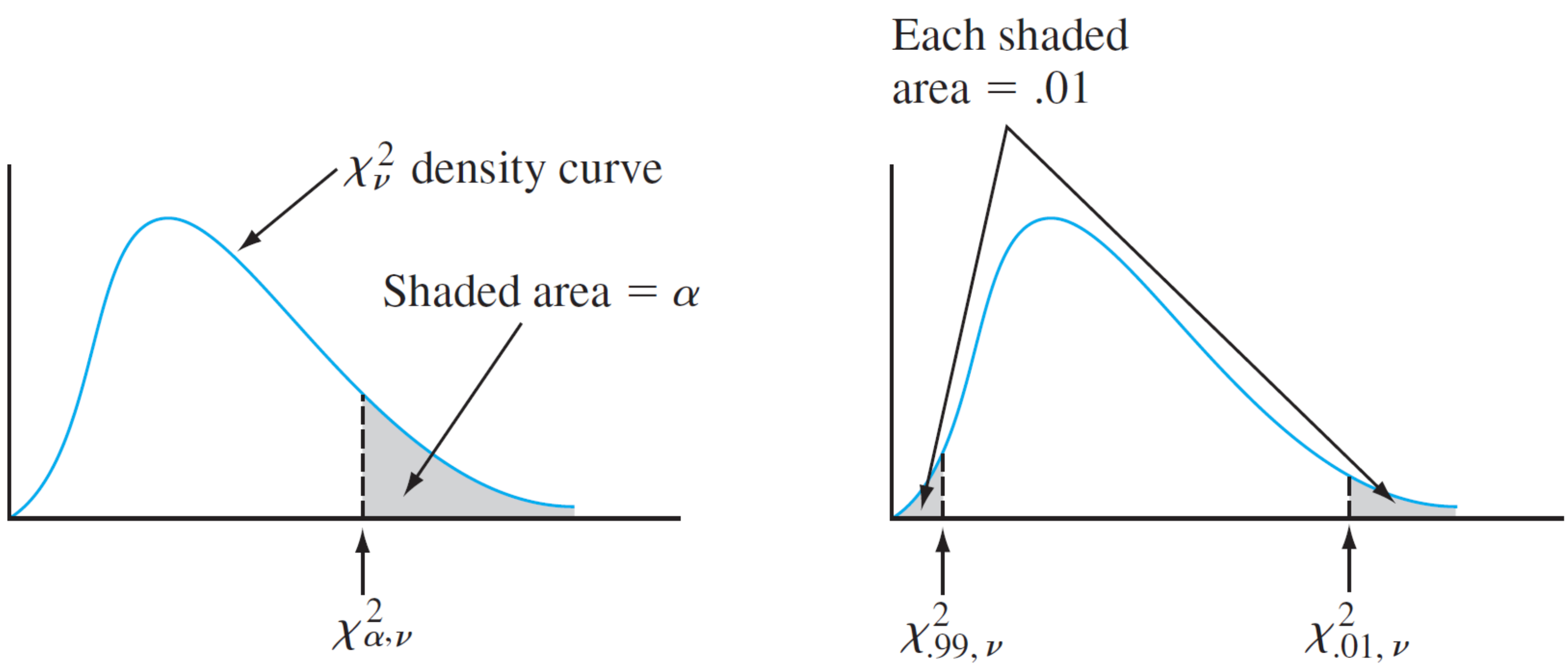
Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with parameters μ and σ^2. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \overline{X})^2}{\sigma^2}$$

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

CIs for the Variance

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



CIs for the Variance

As a consequence

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

Or equivalently

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

Thus we have a confidence interval for the variance σ^2.

Taking square roots gives a CI for the standard deviation σ.

Example

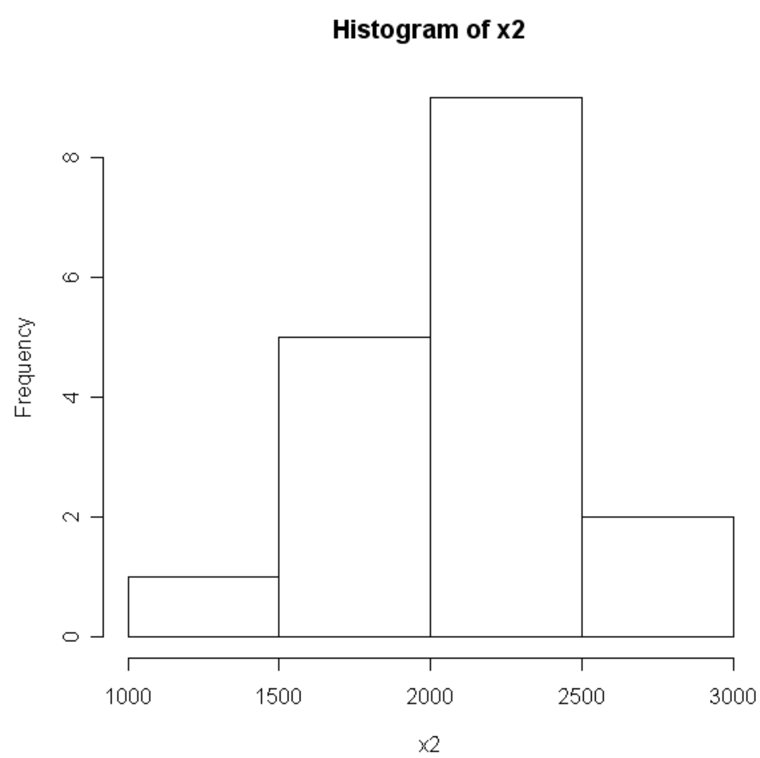
The data on breakdown voltage of electrically stressed circuits are:

1470	1510	1690	1740	1900	2000	2030	2100	2190
2200	2290	2380	2390	2480	2500	2580	2700	

breakdown voltage is approximately normally distributed.

s^2 = 137,324.3

n = 17



Confidence Intervals in R