

7

Statistical Intervals

Chapter 7 Stat 4570/5570

Material from Devore's book (Ed 8), and Cengage

Confidence Intervals

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

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

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

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

Basic Properties of Confidence Intervals

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

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

This is equivalent to:

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

which can be interpreted as the probability that the interval

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

includes the true mean μ is 95%.

Basic Properties of Confidence Intervals

The **interval**

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

is thus called the **95% confidence interval for the mean**.

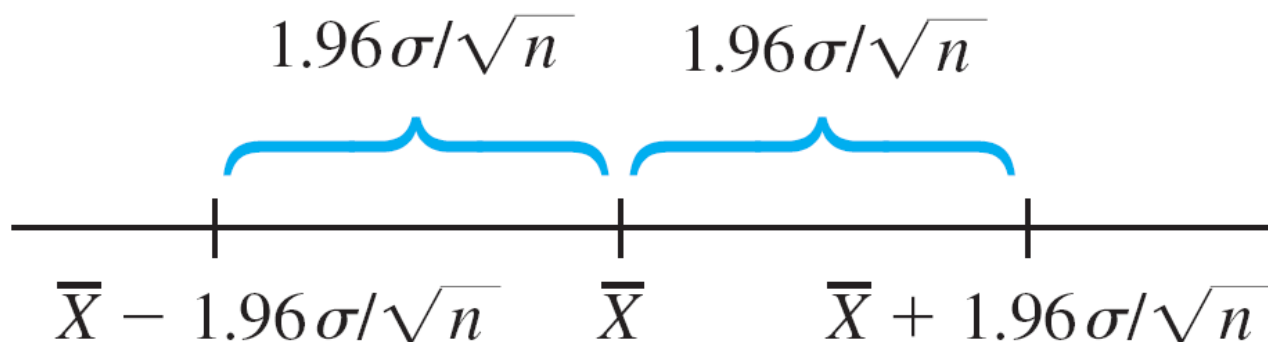
This interval varies from sample to sample, as the sample mean varies.

So the interval itself is a random interval.

Basic Properties of Confidence Intervals

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

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



Basic Properties of Confidence Intervals

For a given sample, the CI can be expressed either as

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right) \text{ is a 95\% CI for } \mu$$

or as

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ with 95\% confidence}$$

A concise expression for the interval is

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

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

Interpreting a Confidence Level

We started with an event (that the random interval captures the true value of μ) whose probability was .95

It is tempting to say that μ *lies* within this fixed interval with probability 0.95.

μ is a constant (unfortunately unknown to us). It is therefore *incorrect* to write the statement

$$P(\mu \text{ lies in } (a, b)) = 0.95$$

-- since μ either is in (a, b) or isn't.

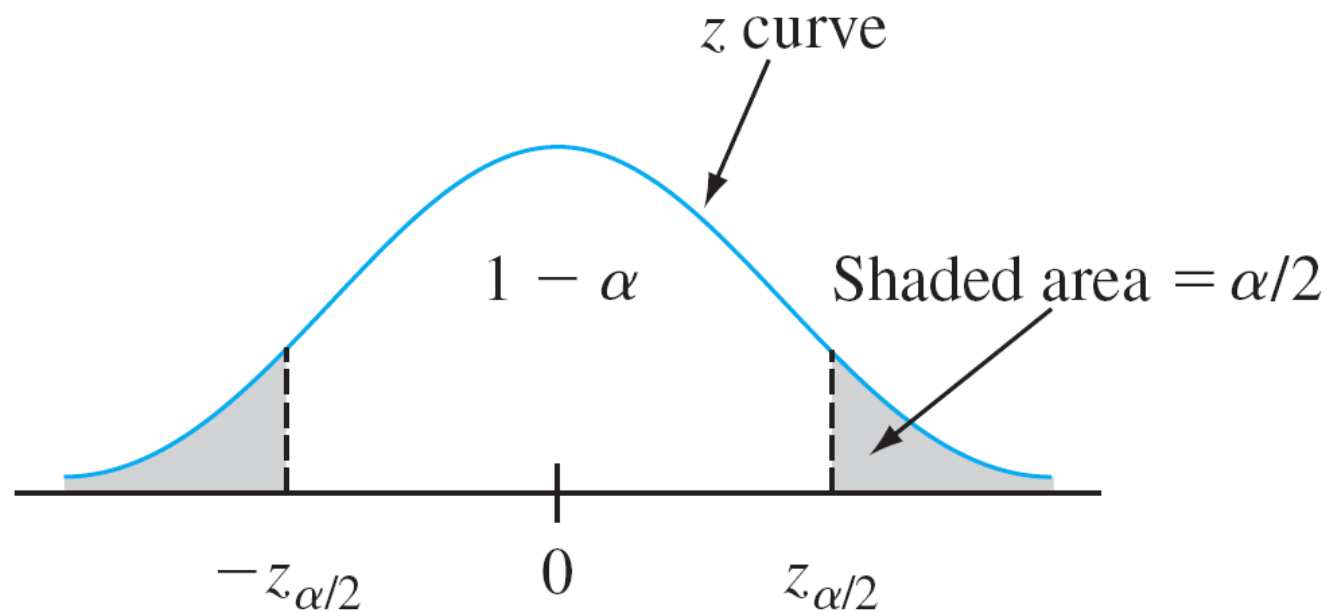
Basically, μ is not random (it's a constant), so it can't have a probability associated with its behavior.

Interpreting a Confidence Level

- Instead, a correct interpretation of “95% confidence” relies on the long-run relative frequency interpretation of probability.
- To say that an event A has probability .95 is to say that if the same experiment is performed over and over again, in the long run A will occur 95% of the time.
- So the right interpretation is to say that 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.

Other Levels of Confidence

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



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

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

The formula for the CI can also be expressed in words as
Point estimate \pm (*z critical value*) \cdot (*standard error*).

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 is the width of the interval?

What about the 99% confidence interval?

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

Sample size computation

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

Example: A response time is Normally distributed with standard deviation 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 mean and variance

We now know that a CI for the mean μ of a normal distribution and a large-sample CI for μ for any distribution with a confidence level of $100(1 - \alpha)\%$ is:

$$\bar{x} \pm z_{\alpha/2} \cdot \sigma / \sqrt{n}$$

A practical difficulty is the value of σ , which will rarely be known. Instead we work with the standardized random variable

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

Where the sample standard deviation s has replaced σ .

Unknown mean and variance

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

In the new standardized variable, both \bar{X} and S vary in value from one sample to another.

Thus the distribution of this new variable should be wider than the Normal to reflect the extra uncertainty. This is indeed true when n is small.

However, for large n the substitution of S for σ adds little extra variability, so this variable also has approximately a standard normal distribution.

A Large-Sample Interval for μ

Proposition

If n is sufficiently large ($n > 40$), the standardized random variable

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

has approximately a standard normal distribution. This implies that

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

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 .)

A Large-Sample Interval for μ

Generally speaking, $n > 40$ will be sufficient to justify the use of this interval.

This is somewhat more conservative than the rule of thumb for the CLT because of the additional variability introduced by using S in place of σ .

Small sample intervals for the mean

- The CI for μ presented in earlier section is valid provided that n is large
 - Rule of thumb: $n > 40$
 - The resulting interval can be used whatever the nature of the population distribution.
- The CLT cannot be invoked, however, when n is small
 - Need to do something else when $n < 40$
- When $n < 40$ and the underlying distribution is unknown, we have to
 - make a specific assumption about the form of the population distribution
 - derive a CI based on that assumption.
- For example: develop a CI for μ when the population is described by a gamma distribution, another interval for the case of a Weibull distribution, and so on.

***t* Distribution**

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

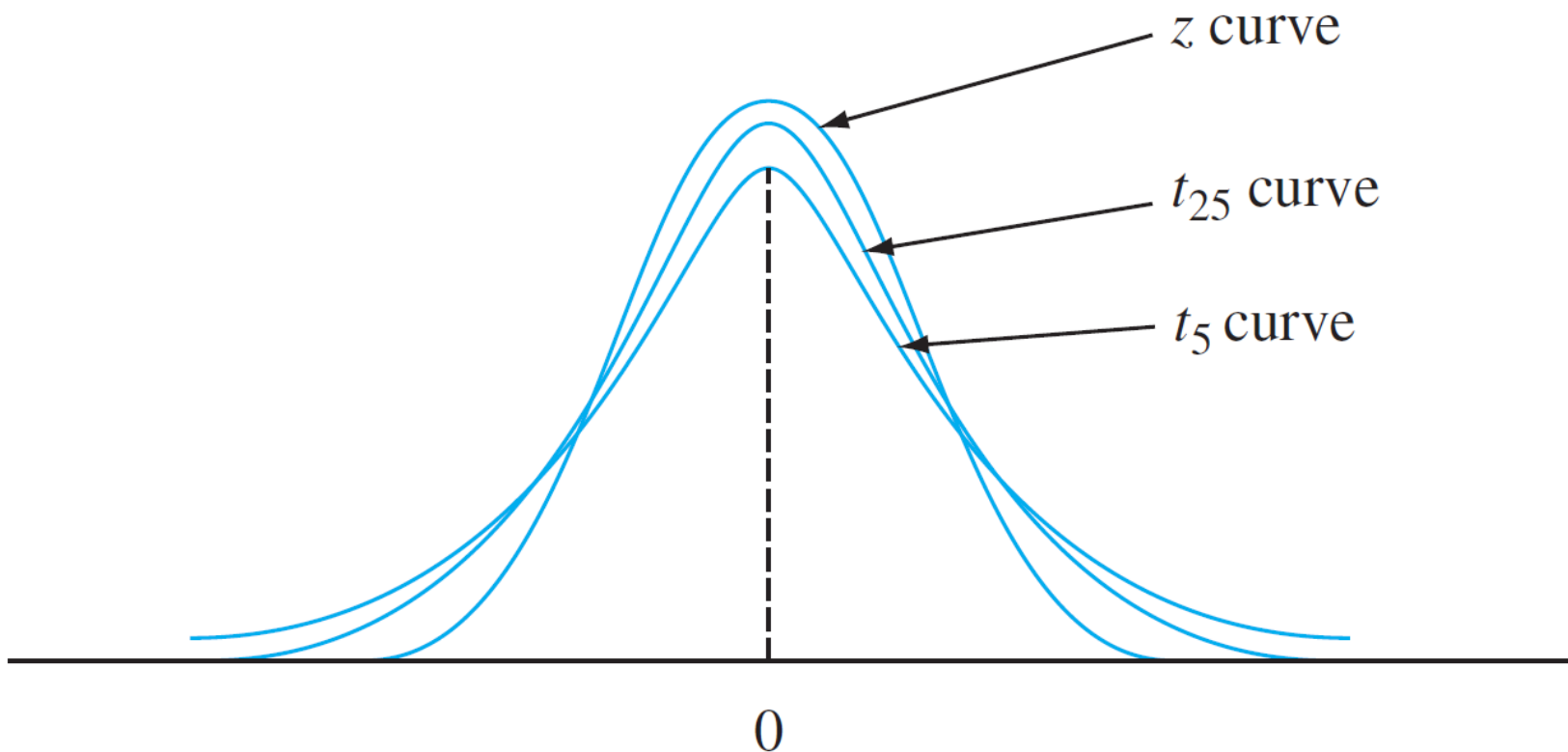
When \bar{X} is the mean of a random sample of size n from a **normal distribution** with mean μ , the random variable

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

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

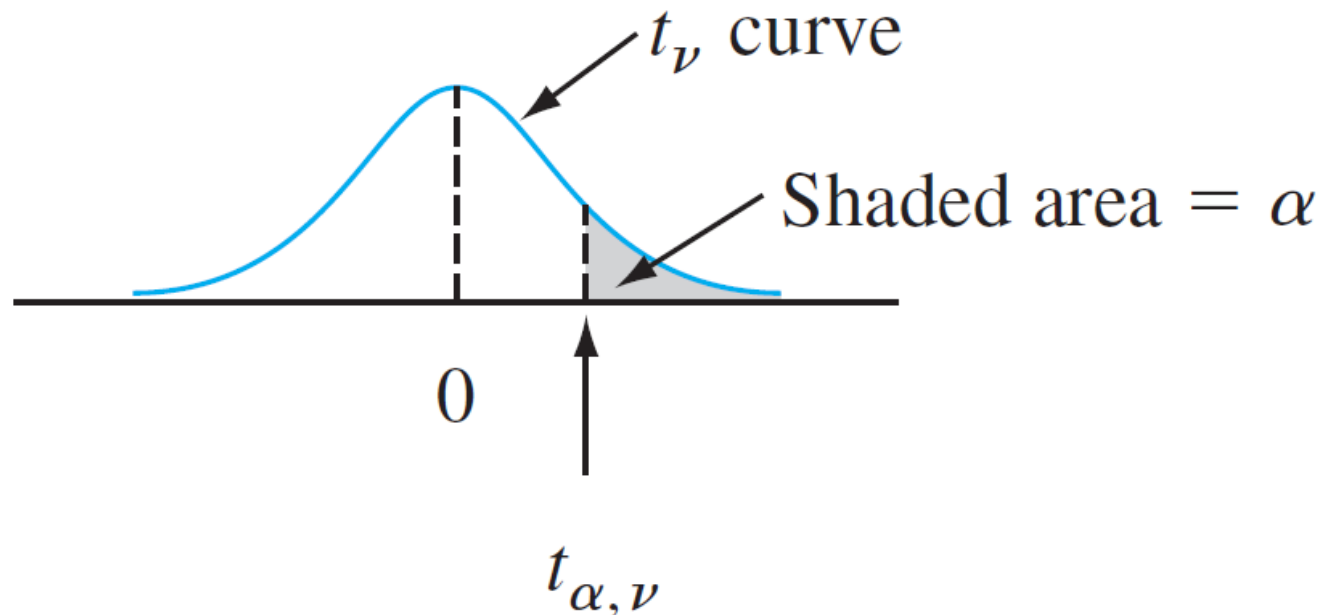
Properties of t Distributions

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

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 One-Sample 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 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 $\bar{x} \pm t_{\alpha/2, n-1} \cdot s/\sqrt{n}$.

So when the true variance is not known and the sample size is small ($n \leq 30$) and the underlying population is believed to be normally distributed, then we use the t -distribution CI.

Example

cont' d

A dataset on the modulus of material rupture (psi):

6807.99	7637.06	6663.28	6165.03	6991.41	6992.23
6981.46	7569.75	7437.88	6872.39	7663.18	6032.28
6906.04	6617.17	6984.12	7093.71	7659.50	7378.61
7295.54	6702.76	7440.17	8053.26	8284.75	7347.95
7422.69	7886.87	6316.67	7713.65	7503.33	7674.99

There are 30 observations.

The sample mean is 7203.191

The sample standard deviation is 543.5400.

Comparison of t vs z in R

Intervals Based on Nonnormal Population Distributions

The one-sample t CI for μ is robust to small or even moderate departures from normality unless n is very small.

By this we mean that if a critical value for 95% confidence, for example, is used in calculating the interval, the actual confidence level will be reasonably close to the nominal 95% level.

If, however, n is small and the population distribution is non-normal, then the actual confidence level may be considerably different from the one you think you are using when you obtain a particular critical value from the t table.

Summary of Confidence Intervals so far...

- If population variance σ^2 is known and the underlying population is known (or assumed to be) normally distributed **or** if $n > 30$ then we use a confidence interval based on the normal dist. (i.e. z-scores).
- If population variance σ^2 is unknown and if $n > 40$ then we use a CI based on z-scores with $\sigma \approx s$ (there is no assumption made here about the distribution).
- If population variance σ^2 is unknown and the underlying population is known (or assumed to be) normally distributed **and** if $n \leq 30$ then we use a CI based on the *t*-dist. with $\sigma \approx s$ and if $n > 30$ use a CI based on z-scores with $\sigma \approx s$.

A Confidence Interval for a Population Proportion

Let p denote the proportion of “successes” in a population, where *success* identifies an individual or object that has a specified property (e.g., individuals who graduated from college, computers that do not need warranty service, etc.).

A random sample of n individuals is to be selected, and X is the number of successes in the sample.

X can be thought of as a sum of all X_i 's, where 1 is added for every *success* that occurs and a 0 for every *failure*, so $X_1 + \dots + X_n = X$.

Thus, X can be regarded as a **Binomial rv** with mean np and

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

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

A Confidence Interval for a Population Proportion

The natural estimator of p is $\hat{p} = X/n$, fraction of successes. Since \hat{p} is the sample mean, $(X_1 + \dots + X_n)/n$ \hat{p} has approximately a normal distribution. As we know that, $E(\hat{p}) = p$ (unbiasedness) and $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$

The standard deviation $\sigma_{\hat{p}}$ involves the unknown parameter p . 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 $\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n}$

Confidence Intervals for the Variance of a Normal Population

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

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

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

We know that the chi-squared distribution is a continuous probability distribution with a single parameter ν , called the number of degrees of freedom, with possible values $1, 2, 3, \dots$.

Confidence Intervals for the Variance of a Normal Population

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

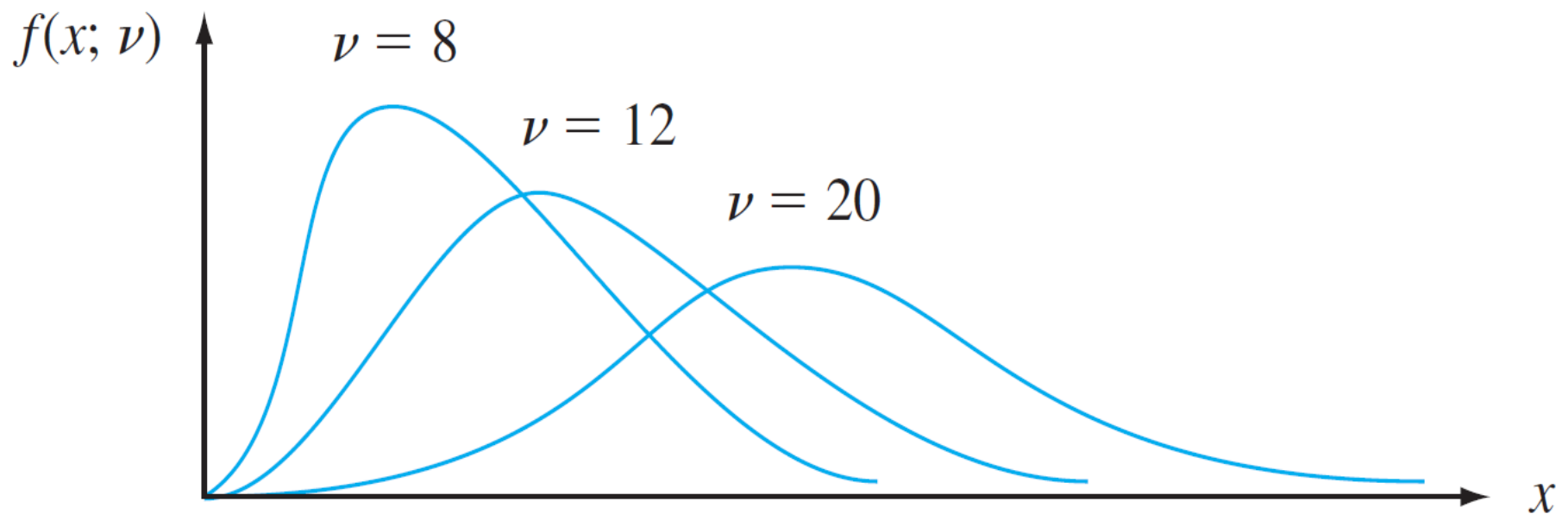
$$\frac{(n - 1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2}$$

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

We know that the chi-squared distribution is a continuous probability distribution with a single parameter ν , called the number of degrees of freedom, with possible values $1, 2, 3, \dots$.

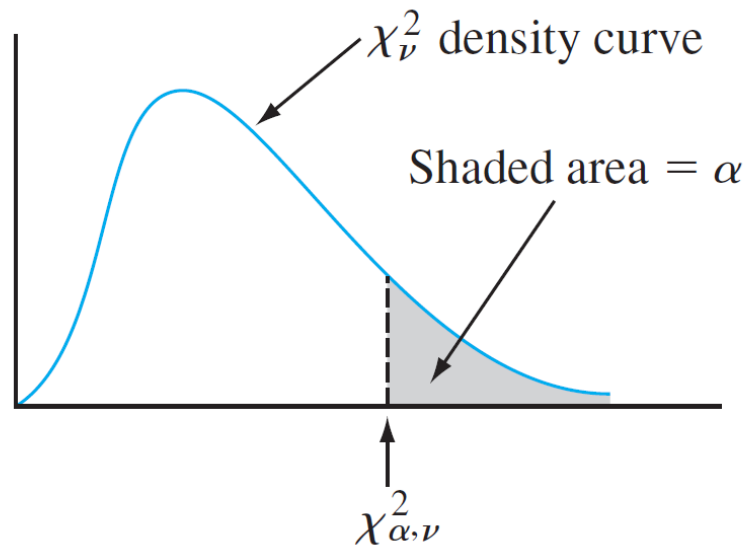
Confidence Intervals for the Variance of a Normal Population

The graphs of several Chi-square probability density functions are

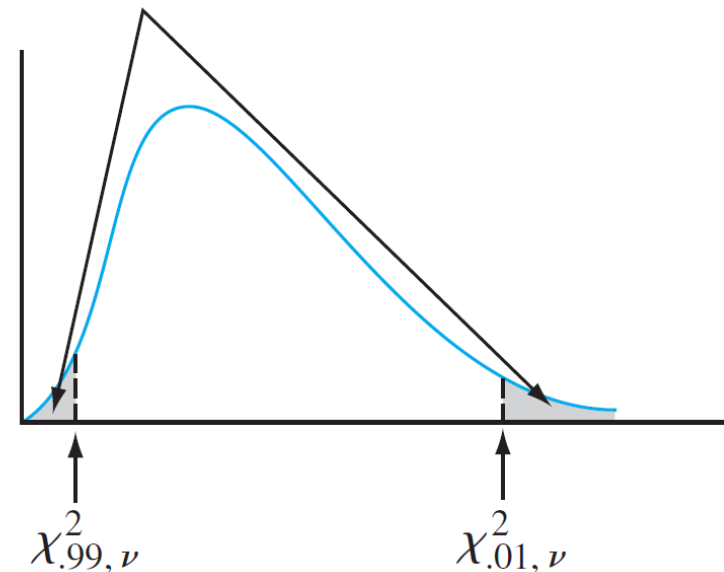


Confidence Intervals for the Variance of a Normal Population

The chi-squared distribution is *not symmetric*, so Appendix Table A.7 contains values of $\chi_{\alpha, \nu}^2$ both for α near 0 and 1



Each shaded area = .01



Confidence Intervals for the Variance of a Normal Population

As a consequence

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

Or equivalently

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

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

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

Confidence Intervals for the Variance of a Normal Population

A **$100(1 - \alpha)\%$ confidence interval for the variance σ^2 of a normal population** has lower limit

$$(n - 1)s^2 / \chi_{\alpha/2, n-1}^2$$

and upper limit

$$(n - 1)s^2 / \chi_{1-\alpha/2, n-1}^2$$

A **confidence interval for σ** has lower and upper limits that are the square roots of the corresponding limits in the interval for σ^2 .

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.

$$s^2 = 137,324.3$$

$$n = 17$$

