Confidence intervals

CE 311S

CONFIDENCE INTERVAL BASICS

A confidence interval gives a plausible range for a population parameter.

The following statements each express a confidence interval:

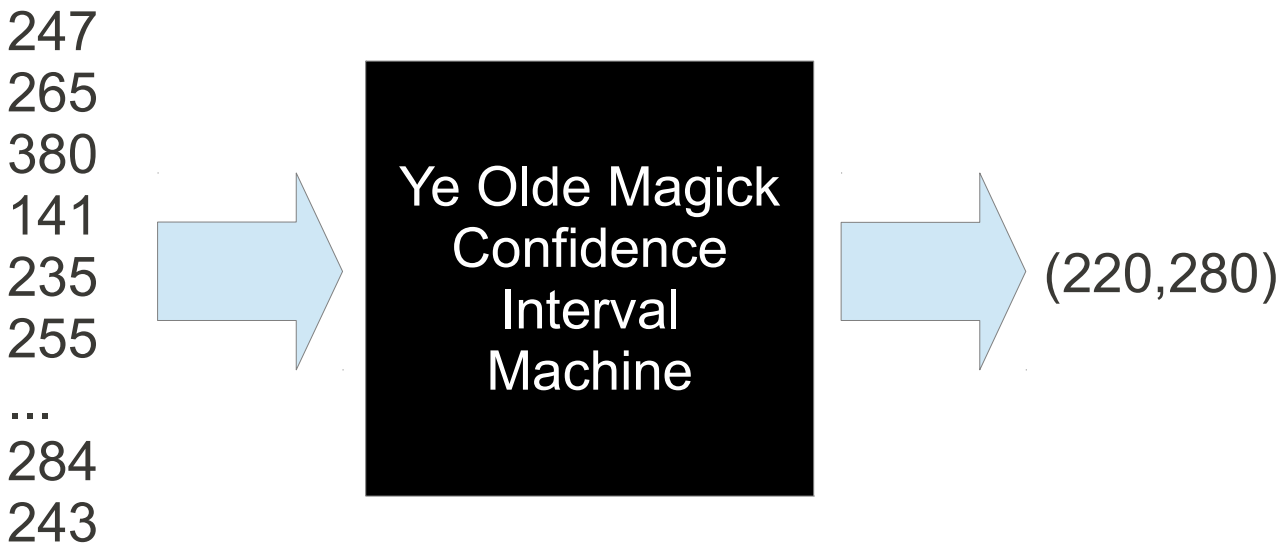
- The percentage of UT students living on campus is between 35% and 40% (with 90% confidence)
- The amount an average student spends on food each month is $\$250 \pm 30$ (95% CI)

A few issues

- Why an interval? Isn't it easier to just report a single value?
- How should we report an interval?

In this class, we'll use interval notation to describe a confidence interval, e.g., (220, 280) for the amount an average student spends on food each month

For now, let's assume that we have a "black box" procedure for calculating a confidence interval.



In a few slides we'll give formulas for these intervals, but let's spend more time on what they actually mean.

How to interpret an interval

"(220, 280) is a 95% confidence interval for the average amount students spend on food."

It is very tempting, but very **WRONG**, to say "there is a 95% probability that the average amount students spend on food is between \$220 and \$280."

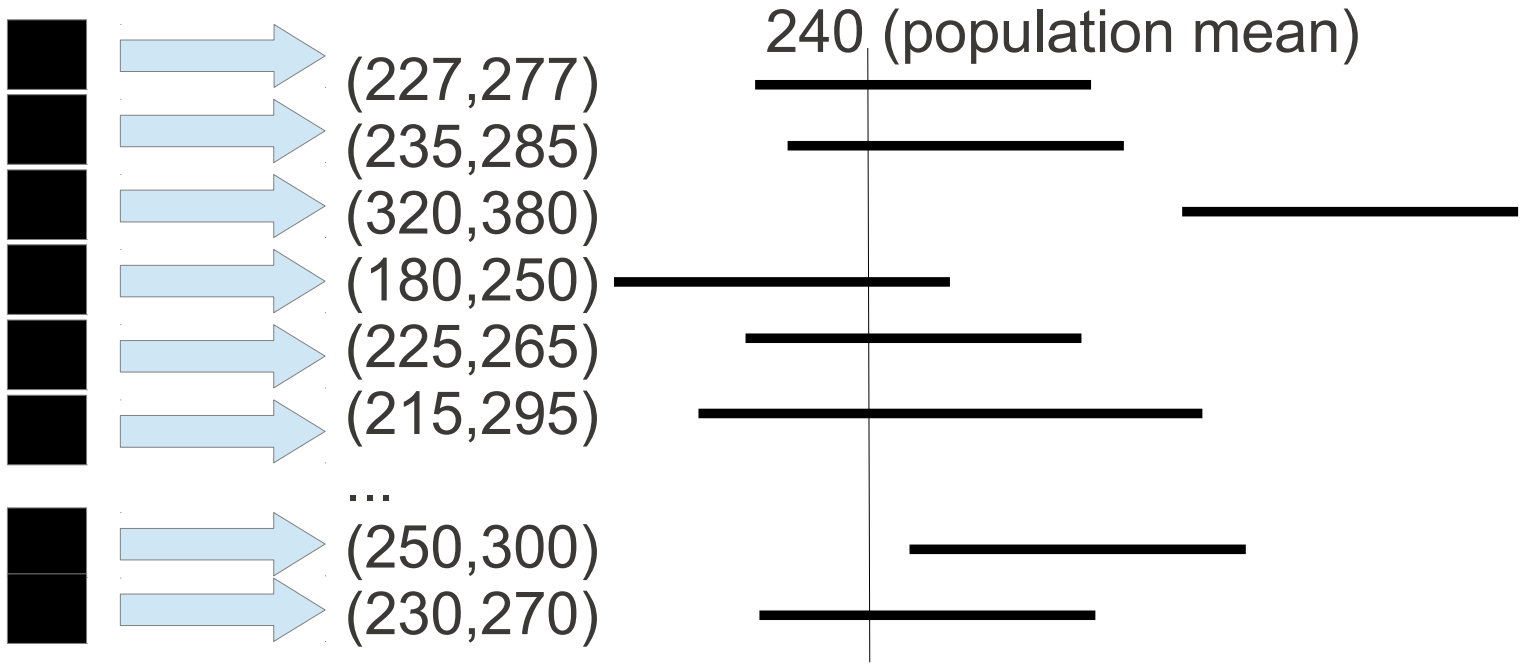
The average amount students spend on food is a population parameter which is **constant**, so it does not make sense to talk about the probability that it lies in an interval — either it does or it doesn't.

Think about it this way: if we redid the survey, the true average amount students spend would not change. However, we would end up with a different interval based on our results.

The correct interpretation is the following:

"The interval (220, 280) contains the average amount students spend on food with 95% probability."

Let's say that the true population average is 240. We want to design our black box so that, if conducted a very large number of surveys, 95% of the resulting confidence intervals would contain 240 (the true mean).



In reality, we don't usually have the money to conduct all of these studies. This example is not meant to illustrate the way we would actually go about things; but to show what we mean by "95% confidence"

The 95% CI is (220, 280). If we **increased** the level of confidence (say, to 99%), would the interval get larger or smaller?

What is the 100% CI? Why don't we use that?

There is a tension between **width** and **confidence**. A high-confidence interval is wide (and less specific). A narrow interval is easier to use for making decisions, but you have less confidence that it's actually correct. Typically 95% confidence is a good balance, but some applications need a higher or lower level.

INTERACTIVE INTERVALS

For each of the following items, write down a 90% confidence interval based on your intuition.

1. Population of the United States on January 1, 2000
2. The year of Napoleon's birth
3. Length of the Mississippi-Missouri river in miles
4. Maximum takeoff weight of a Boeing 747 (in pounds)
5. Seconds for a radio signal to travel from the earth to the moon
6. Latitude of London
7. Minutes for a space shuttle to orbit the earth
8. Length between towers of the Golden Gate Bridge (in feet)
9. Number of signers of the Declaration of Independence
10. Number of bones in an adult human body

DERIVING A CONFIDENCE INTERVAL

I collect spending data from 50 students; the sample mean $\overline{X}$ is \$250, and the sample standard deviation is \$108. How can I develop a 95% confidence interval?

By the central limit theorem, $\overline{X}$ has a normal distribution with mean μ (the population mean) and standard deviation $\sigma/\sqrt{n}$ (where σ is the population standard deviation).

Therefore, $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$.

Usually we want a symmetrical interval; from the normal tables, the probability Z is between -1.96 and $+1.96$ is 95%.

Therefore, $P(-1.96 < Z < 1.96) = P(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96)$ or, after some algebra,

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

So, if we substitute our survey results ($\overline{X} = 250$, $\sigma \approx 108$, $n = 50$) we obtain the interval (220, 280)

One technicality: we don't actually know the population standard deviation σ; we only have the sample standard deviation s. We'll come back to this issue later; for now, just assume it's close enough.

Example

You poll 100 students for the number of times they eat out each week. Your results have a sample mean of 4 and a standard deviation of 2. What is a 90% confidence interval on the population mean?

You may find some of the following Z values handy:

$$\Phi(-1.96) = 0.025 \quad \Phi(-1.64) = 0.050 \quad \Phi(-1.28) = 0.10$$
$$\Phi(1.28) = 0.90 \quad \Phi(1.64) = 0.950 \quad \Phi(1.96) = 0.975$$

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

What can we do to make this interval narrower?

In the previous example, the width of the 90% confidence interval was 0.66. What should the sample size be to reduce its width to 0.5?

The width is $2 \times 1.64 \times 2/\sqrt{n}$. If $n \geq 172$, the width will be less than 0.5

INTERVAL ESTIMATION

To recap the process we used to derive the α-confidence interval for the population mean μ, based on the sample mean $\overline{X}$:

- Use the fact that the sample mean is related to the population mean through the central limit theorem: $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.
- Write a probability statement for a standard normal distribution, for the given confidence level $1-\alpha$.
- Rearrange this statement to get an interval around $\overline{X}$

We can use this pattern to form other kinds of intervals as well.

General interval estimation process

The following process works whenever we want to find an α-confidence interval for some population parameter θ, based on a statistic Q:

- Determine what distribution Q has in terms of θ and your random sample.
- Write a probability statement for Q, for the given confidence level $1 - \alpha$.
- Rearrange this statement to get an interval around θ

If the interval takes the form $[\hat{\Theta}_l, \hat{\Theta}_r h$, then we say that $\hat{\Theta}_l$ is the low estimator, and $\hat{\Theta}_h$ the high estimator.

FANCIER CONFIDENCE INTERVALS

Other types of confidence intervals

- Formulas for confidence intervals
- More on interpreting confidence intervals
- What if we don't know the population standard deviation?
  - Is the sample large?
  - Is the underlying population normally-distributed?

We said a formula for the 95% confidence interval was

$$(\overline{X} - 1.96\sigma/\sqrt{n} < \mu < \overline{X} + 1.96\sigma/\sqrt{n})$$

Where did 1.96 come from? Is there an easy way to write a formula based on the desired level of confidence?

We can write a general form using *z-critical values*.

z_α is the value for which $P(Z > z_\alpha) = \alpha$

That is, the area under the curve to the *right* of z_α is α.

What is the difference between Φ and z_α?

Some common z-critical values are:

α	z_α	Percentile
0.1	1.28	90
0.05	1.645	95
0.025	1.96	97.5
0.01	2.33	99
0.005	2.58	99.5
0.001	3.08	99.9
0.0005	3.27	99.95

Using these, the confidence interval formula can be rewritten

$$(\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n})$$

where $1 - \alpha$ is the confidence level.

Why $z_{\alpha/2}$ instead of z_{α}?

The sample size needed for a confidence interval to have width w is

$$n = \left(2z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

SOME MORE MATH

One problem with the formula

$$(\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n})$$

is that we don't know σ, but only s.

However, if n is large, $s \approx \sigma$, so $(\overline{X} - \mu)/(s/\sqrt{n})$ will be approximately normally distributed.

For this rule of thumb to hold, we need $n > 40$. (Why not 30?)

WHAT IF WE DON'T HAVE A LARGE SAMPLE?

If $n < 40$, then we can't trust that $(\overline{x} - \mu)/(s/\sqrt{n})$ is approximately normal, so the above analysis fails.

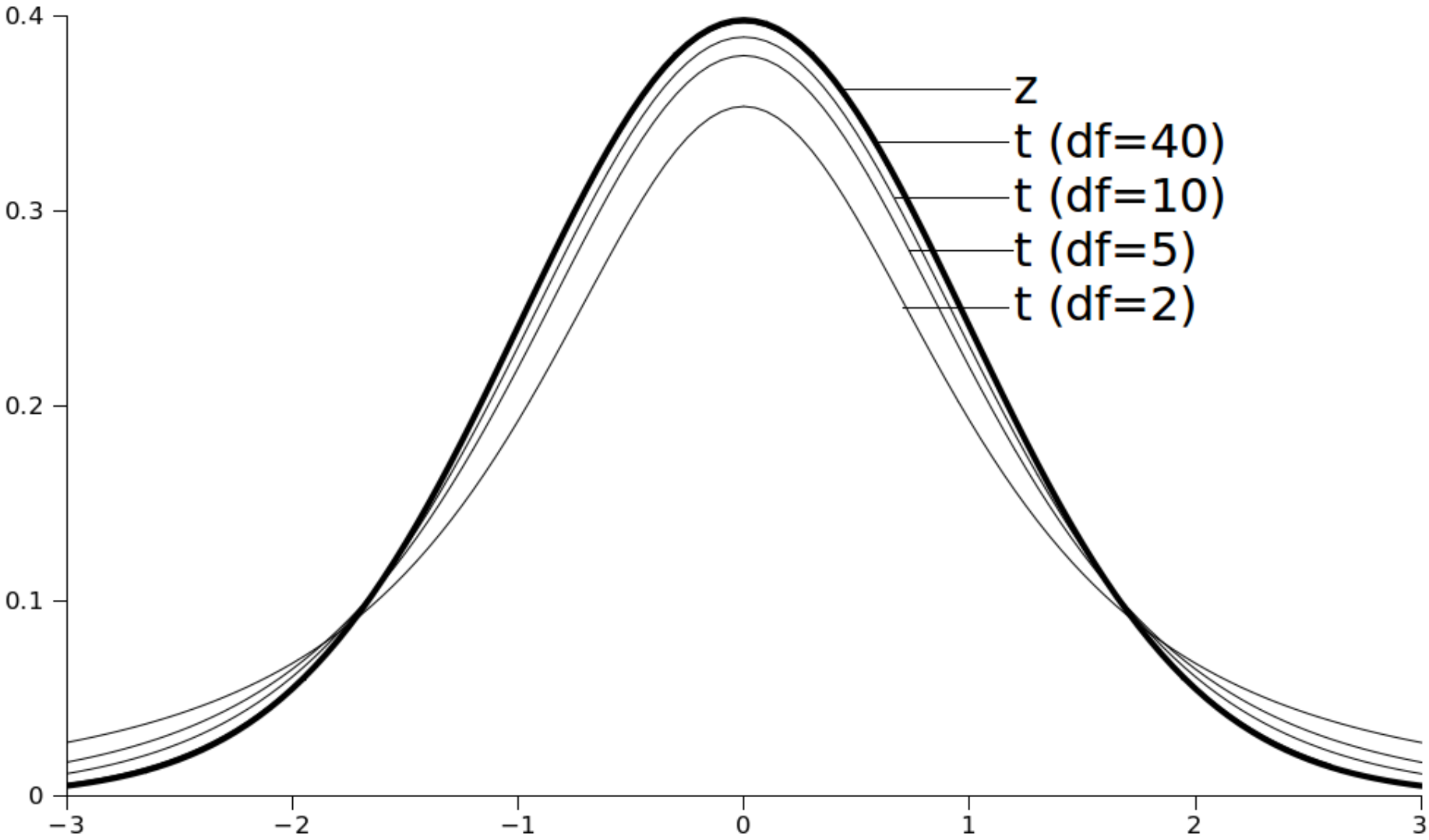
The one exception is when the underlying population is normally distributed. In this case,$(\overline{x} - \mu)/(s/\sqrt{n})$ has a t-distribution.

Remember that $(\overline{x} - \mu)/(\sigma/\sqrt{n})$ has a normal distribution; by using s instead of σ, we have the t-distribution.

Think of the t distribution as a "spread out" version of the normal distribution, with the additional spread reflecting uncertainty in the sample standard deviation.

The t distribution has one extra parameter: the "degrees of freedom" (df), which are equal to $n - 1$.

As $n \to \infty$, the t distribution approaches the normal distribution. (This is why it's OK to use s instead of σ when $n > 40$.)


0.4
0.3
0.2
0.1
0
−3
−2
−1
0
1
2
3
z
t (df=40)
t (df=10)
t (df=5)
t (df=2)

The PDF and CDF of the t-distribution are more complicated than they're worth; the t values can be found in a table (see Canvas). $t_{\alpha,\nu}$ denotes the t-critical value: the probability that the t distribution with ν degrees of freedom is greater than $t_{\alpha,\nu}$ is α.

The formula for a confidence interval in this case is

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

where the confidence level is $1 - \alpha$.

Example

You count the number of cars on I-35 in downtown for nine days, obtaining a sample mean of 50,000 vehicles and a sample standard deviation of 10,000 vehicles. Find a 95% confidence interval on the population mean, assuming the number of cars on I-35 follows a normal distribution.

Example

You have collected the following data on the cost of a studio apartment in Austin:

500 550 575 580 590

Assuming that this cost is normally distributed in the population, what is a 99% confidence interval on the mean cost?

Some selected t-values:

	α					
ν	0.10	0.05	0.025	0.01	0.005	0.001
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208

ONE-SIDED INTERVALS

How would I construct a one-sided confidence interval?

I am designing a structure which involves a number of connectors. I want a 95% confidence interval for the mean strength of these connectors.

If the strength is *lower* than I think, there is a big problem. But if the strength is *higher* than I think, there is no problem.

Therefore, I want a *one-sided* 95% confidence interval of the form (K, ∞). How can I derive this?

$$P(Z < z_{.05}) = 0.95 \Rightarrow P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < 1.645\right) = 0.95$$

So the desired interval is $\left(\overline{x} - z_\alpha \frac{s}{\sqrt{n}}, \infty\right)$

Example

I test 100 of these connectors, and find that the sample mean and sample standard deviation are 1500 psi and 200 psi, respectively.

The 95% *one-sided* interval is
$(1500 - 1.645 \times (200/\sqrt{100}), \infty) = (1467, \infty)$, and this interval contains the true mean strength with 95% probability.

Example

You are designing a drainage system for Austin, and want a one-sided 90% confidence interval on the average monthly rainfall. Based on a sample of 49 months, I find a sample mean and sample standard deviation of 5 inches and 3 inches, respectively. What is this interval?

α	z_α	Percentile
0.1	1.28	90
0.05	1.645	95
0.025	1.96	97.5
0.01	2.33	99
0.005	2.58	99.5
0.001	3.08	99.9
0.0005	3.27	99.95

STANDARD DEVIATION

So far, we've only spoken about confidence intervals for the mean value.

It's possible to do the same thing for the standard deviation. To see how, it is helpful to retrace how we derived the confidence interval for the mean.

Here is how we found the confidence interval for the mean

1. Decided to use the sample mean $\overline{x}$ as an estimator for the population mean μ
2. From the Central Limit Theorem, we know what kind of distribution $\overline{X}$ has (normal) along with its parameters (mean, standard deviation)
3. We can write probability expressions using this distribution, and solve for an interval involving μ.

Is there something like the central limit theorem which tells us the distribution of the sample standard deviation?

It turns out that if the underlying population is normally distributed, the sample standard deviation satisfies the following relationship:

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with variance σ^2. Then the scaled random variable

$$\frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared distribution with $n-1$ degrees of freedom.

Remember the chi-squared distribution with ν degrees of freedom? It was a special case of the gamma distribution, with pdf

$$\begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

There is a table for chi-squared values as well (Table A.7), which gives χ^2-critical values.

If we have ν degrees of freedom, the area to the right of $\chi^2_{\alpha,\nu}$ is equal to α.

So, we can derive a confidence interval in the following manner:

1. Choose a desired level of confidence.
2. Pick two critical χ^2 values so that the area between them is the desired level of confidence.
3. Do algebra magic to find an interval for σ.

Example

Based on a sample of 20 students, the mean and standard deviation of the number of parking tickets in a semester are 15 and 3. What is a 95% confidence interval for the population standard deviation?

For a symmetric confidence interval, choose the values $\chi^2_{0.975,19} = 8.906$ and $\chi^2_{0.025,19} = 32.852$

Then $Pr(8.906 < (n-1)S^2/\sigma^2 < 32.852) = 0.95$

Or, after some magic,

$$Pr\left(\frac{(n-1)S^2}{32.852} < \sigma^2 < \frac{(n-1)S^2}{8.906}\right) = 0.95$$

Substituting $n = 20$ and $S = 3$, we obtain the interval (5.21, 19.2) for the variance, or (2.28, 4.38) for the standard deviation.

In general, the confidence interval for the variance is given by the formula

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

For a confidence interval on the standard deviation, take the square root of both endpoints.

OTHER TYPES OF INTERVALS

There are two other types of intervals that are commonly used:

- **Prediction Intervals**: What is an interval for a single observation from the population?
- **Tolerance Intervals**: What is an interval that will contain a certain fraction of the population?

PREDICTION INTERVALS

Examples where prediction intervals might be useful are questions such as:

"With 95% confidence, what range would describe the number of speeding tickets that a single student receives?"

"With 90% confidence, what is the upper limit on the rainfall in Austin next month?"

Here is the logic used to derive a prediction interval:

- We have a random sample $X_1, \ldots, X_n$ and want to predict the value of X_{n+1}.
- A reasonable guess for X_{n+1} is $\overline{X}$, because the expected prediction error is

$$E[\overline{X} - X_{n+1}] = \mu - \mu = 0$$

- Furthermore, the variance in the prediction error is

$$V[\overline{X} - X_{n+1}] = \sigma^2/n + \sigma^2 = \sigma^2\,(1 + 1/n)$$

- Finally, $\overline{X}$ and X_{n+1} are both normally distributed, so $\overline{X} - X_{n+1}$ is as well.
- Since we only have s (and not σ), we need to use a t distribution instead.

Going through the algebra, we find that for a *prediction level* of $1 - \alpha$, the appropriate prediction interval is

$$(\overline{x} - t_{\alpha/2,n-1} \cdot s\sqrt{1 + 1/n}, \overline{x} + t_{\alpha/2,n-1} \cdot s\sqrt{1 + 1/n})$$

Example

Based on a sample of 20 students, the mean and standard deviation of the number of parking tickets in a semester are 15 and 3. What is a 95% prediction interval for the number of tickets a single student receives?

By comparison, the 95% confidence interval for the sample mean is (14.6, 16.4). Why the difference?

TOLERANCE INTERVALS

Examples of questions tolerance intervals answer include:

"With 95% certainty, what range of parking tickets would cover at least 90% of the students?"

"With 90% certainty, what is the upper limit of rainfall for 95% of the months?"

This interval is actually very easy to calculate, and is simply

$$(\overline{x} - C_{\alpha,k}s, \overline{x} + C_{\alpha,k}s)$$

where $C_{\alpha,k}$ is the tolerance critical value for a confidence level of α, where you want to capture at least k percent of the population.

Table A.6 (on Canvas) lists off tolerance critical values.

A potential source of confusion is between α and k. α is the level of confidence we want that the interval is correct. k is the proportion of the population that we want the interval to cover.

Example

With 95% confidence, what range of values encompasses the number of tickets that 90% of students receive? (Remember $n = 20$, $\overline{x} = 15$ and $s = 3$.)

Table A.6 gives $C_{0.95,0.90} = 2.310$ for $n = 20$.

So, the interval is (8.07, 21.93)