Objectives

**Estimating with confidence**

- Confidence intervals.
  - Sections 6.1 and 7.1 in IPS.
  - Page 174-180 OS3.

- Choosing the sample size

- \( t \) distributions. Further reading
  

- One-sample \( t \) confidence interval for a population mean
  - Page 219 OS3
Overview of Inference

- Sample ≠ population, and sample mean \( \bar{x} \) ≠ population mean \( \mu \). But we do not know the value of \( \mu \). If we want to draw any conclusions about \( \mu \) then we have to use \( \bar{x} \) to do so.

- Methods for drawing conclusions about a population from sample data are called **statistical inference**.

- There are two main types of inference:
  - **Confidence Intervals** - estimating the value of a population parameter, and
  - **Tests of Significance** - assessing evidence for a claim (hypothesis) about a population.
Motivation: confidence intervals 1

- $\bar{x} = 4.2$ is the average review that 174 customers gave. We are going to assume that these scores are representative (i.e. not biased towards those who simply love or hate the product).

- But there would have been thousands who have bought this product. What would the average be amongst the population who bought this product. Ideally, we would have the average amongst this far larger population, we call this $\mu$. 

![Image of a fitness tracker with reviews and rating information]
Motivation: confidence intervals 2

- I sample from the class these 5 heights 61, 63, 65, 66, 72.
- The sample mean is 65.4.
- The population standard deviation is known to be 4.3 inches.
- It is unlikely that the population mean (where the population are all undergraduates in A&M) is the same as the sample mean.
- We know that if the sample mean is normally distributed then there is a 95% chance it will be within 1.96 standard errors of the population mean.
- In this chapter we use this piece of information to construct confidence intervals for locating the mean.
Topics: Confidence intervals and Margins of error

Learning Targets:

Confidence intervals

- Understand what a confidence interval is.
- Know how to construct a confidence interval at different levels of confidence.
- Be able to check if a confidence interval is giving the stated level of confidence.

- Be able to assess how the stated level of confidence compares with the true level of confidence.
- Understand how the distribution of the original data influences the confidence interval.
- Be able to use the statcrunch app for confidence intervals.
Review: properties of the sample mean

The sample mean $\bar{x}$ is a unique number for any particular sample. If you had obtained a different sample (by chance) you almost certainly would have had a different value for your sample mean.

In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, $\mu$. 

![Diagram showing sample means and sampling distribution of $\bar{x}$]
Because the sampling distribution of $\bar{X}$ is narrower than the population distribution, the estimates $\bar{X}$ tend to be closer to the population parameter $\mu$ than individual observations are.

If the population is normally distributed $N(\mu, \sigma)$, the sampling distribution is $N(\mu, \sigma/\sqrt{n})$,
Example 1: Potassium

- We recall from Chapter 5, the example where the level of potassium in person’s blood sample is normally distributed with mean 3.8 and standard deviation 0.2.
- 4 blood samples are taken and the sample mean calculated.
- Using the results in Chapter 5, since the blood samples are normally distributed the sample mean will also be normally distributed. The standard error of the sample mean (based on 4) is 0.1.
- By normality of the sample mean, the sample mean (of 4 blood samples) will be within 1.96 standard errors of the true mean potassium level.
- But typically we do not know a person’s mean level. We only have their sample average. Can we locate the mean from the sample average?
In Chapter 4, we learnt that if a variable were normally distributed with mean \( \mu \) and standard deviation \( \sigma \), 95\% of the population will lie within 1.96 standard deviations of the mean. We write that 95\% of the population lie in the interval

\[
[\mu - 1.96 \times \sigma, \mu + 1.96 \times \sigma]
\]

Returning to the sample mean, suppose it has mean \( \mu \) and standard error \( \sigma/\sqrt{n} \) (Chapter 5) and is normally distributed. 95\% of the sample means will lie within 1.96 standard errors of the mean. We write this as 95\% of sample means lie in the interval:

\[
\left[ \mu - 1.96 \times \frac{\sigma}{\sqrt{n}}, \mu + 1.96 \times \frac{\sigma}{\sqrt{n}} \right]
\]

- This expression looks unwieldy. But remember \( \sigma \) and \( n \) are both numbers.
But typically, the mean is **unknown**, our objective is to locate the true mean based on the sample mean.

- To do this we turn the story around, if the sample mean lies in the interval
  \[ \left( \mu - 1.96 \times \frac{\sigma}{\sqrt{n}}, \mu + 1.96 \times \frac{\sigma}{\sqrt{n}} \right) \]

- This is the same as saying the mean \( \mu \) lies in the interval
  \[ \left[ \text{sample mean (average)} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \text{sample mean (average)} + 1.96 \times \frac{\sigma}{\sqrt{n}} \right] \]

- Thus 95% of the all confidence intervals will contain the true mean.

We call this a 95% confidence interval for the mean.

In the next few slides we construct 95% confidence intervals for the mean for normally distributed data and non-normally distributed data.
With 95% confidence mean $\mu$ lies in this interval.
Confidence intervals for Normally distributed data
Example 1: Normal data – sample size three

- Female heights are approximately normally distributed. The standard deviation of a human height is known to be 3.8 inches.
- Using the results in Chapter 5, if the sample size is 3 and the average is taken, then the standard error of the average is $3.8/\sqrt{3} = 2.19$.

\[
\bar{X} \sim N\left(\mu, \frac{3.8}{\sqrt{3}}\right)
\]

- This means that 95% of all sample means will lie in the interval

\[
[\mu - 1.96 \times 2.19, \mu + 1.96 \times 2.19]
\]

- We interview 3 people, their heights are 64, 68, 72 inches. The average is 68. Thus we believe with 95% confidence that the mean height lies in the interval

\[
[68 - 1.96 \times 2.19, 68 + 1.96 \times 2.19] = [63.7, 72.29]
\]
In the next two slides we illustrate what is meant by 95%

- The first plot is the density of heights.
- The second plot is the density of the sample mean (based on three).

The mean of both distributions have the same center (which is as expected). The sample mean is more likely to be “close” to the mean than an individual observation (variation is less).

Finally, we observe that the distribution of the sample mean is also normal.

The sample mean (say, 68) is one number from the green distribution.
In the plot on the right each horizontal line corresponds to one sample mean and its 95% confidence interval. The middle vertical line corresponds to the true population mean (which in practice is unknown).

**If the sample mean is normally distributed**, about 95 of the intervals would cross the middle line (all green horizontal lines).

In reality, we can only construct one 95% confidence interval. 95% refers to our confidence in the interval containing the mean.
To make the green line plot (in the previous page) in Statcrunch:

Go to Applets -> Confidence intervals -> for a mean
• From data (select Data set you want to use or use the default options)
• Initial confidence (0.95)
• Initial sample size (give size of sample used)
• Interval (select Z with pop std. dev.)
Some important points

- When the sample size is small, for example, just five observations, it is very hard to say whether it was drawn from a normal distribution. One can make a QQplot, but this is a plot with just 5 points.

- It is difficult to judge the distribution from these 5 points. However, the CLT tells us that the sample will be close to normal even if the original data is not. This will be of great help (see the non-Gaussian discussion in the next few slides).
Other confidence levels.

- If we want more confidence in the interval, we have to **increase** the percentage level, which, in turn, will increase the length of the interval.

- To have 99% confidence in the interval. We need to obtain a symmetric interval centered about the mean, where 99% of the sample means will lie inside. Using the z-tables this corresponds to $z = 2.56$ standard errors from the mean.

- For the height example, where the sample mean is 68:
  - This is the 95% confidence interval for the mean height
    \[
    [68 - 1.96 \times 2.19, 68 + 1.96 \times 2.19] = [63.7, 72.29]
    \]
  - This is the 99% confidence interval for the mean height
    \[
    [68 - 2.56 \times 2.19, 68 + 2.56 \times 2.19] = [62.3, 73.6]
    \]
Suppose weights are normally distributed with unknown mean but known standard deviation 30 pounds. 10 people are interviewed. The average weight of these 10 people was 150 pounds. What is the 95% confidence interval for the mean weight of people.

(A) [91.2,208.8]
(B) [131.4,168.6]

http://www.easypolls.net/poll.html?p=59d3bce0e4b0ede9e93921ec
The Confidence interval for the mean weight of a student (on the previous slide) is [131.4, 168.6].

This tells us the mean weight can lie anywhere within this interval.

Just because the majority of this interval is greater than 140 pounds, we cannot make statements like “the mean weight is more likely to be more than 140 pounds than less than 140 pounds”.

Such statements are incorrect. Do not use them and do not be fooled in the exam if it is used. A confidence interval simply gives a range of possible means.
Hypokalemia is diagnosed when the blood potassium level is below 3.5mEq/dl. The potassium in a blood sample varies from sample to sample and follows a normal distribution with unknown mean but standard deviation 0.2. A patient’s potassium is measured over 4 days. The sample over 4 days is 3, 3.5, 3.9, 4.4, its sample mean is 3.7.

**Question:** Construct a 95% confidence interval for the mean potassium and discuss whether the patient is likely to be diagnosed with Hypokalemia.

**Answer:** The standard error for the sample mean is $0.2/\sqrt{4} = 0.1$. Thus the 95% confidence interval for the mean potassium level is $[3.7 \pm 1.96 \times 0.1] = [3.504, 3.894]$. This means with 95% confidence we believe the mean lies in this interval.

Since 3.5 or less does not lie in this interval, it suggests that the patient does not have low potassium. There is a precise way of answer this specific problem which we discuss in Chapter 7 (called statistical testing).
Confidence intervals for Non-normal data
Example 2: Skewed data – sample size 3

- In the previous example we looked at height data, which tends to be normal. In this example we consider **NON-normal data and confidence intervals**.
- As an example, we use the salaries from NBL.
- We (randomly) sample three salaries and take the average.
- The average is estimating the mean is 4.3 and standard error s.e = $\frac{5.5}{\sqrt{3}} = 3.17$.
- We construct a 99% confidence interval to locate the mean:

$$\left[ \bar{X} - 2.56 \times \frac{5.5}{\sqrt{3}}, \bar{X} + 2.56 \times \frac{5.5}{\sqrt{3}} \right]$$

The confidence interval is constructed under the assumption that the sample mean is normal. In the next slide, we investigate the validity of this statement.
Example: Non-normal data – sample size 3

- We use the salaries from NBL.

This is a histogram of NBL salaries. The green plot is histogram of the sample mean of three salaries. Observe the population means (the green line and triangle) of both plots is the same.

The green plot has less spread (s.e. = $5.5/\sqrt{3} = 3.17$) than the top plot ($\sigma = 5.5$) but it is still skewed!
Suppose the we do not know the histogram on the previous page but do observe three salaries 0.5, 0.8, 1.5 million dollars. The sample mean is 0.93 million dollars. The 95% confidence interval for the mean salary is

\[ [0.93 - 2.57 \times 3.17, 0.93 + 2.57 \times 3.17] = [-7.2, 9.1] \]

This interval makes no sense. There cannot be negative average baseball salaries. So the mean cannot be negative, which we remove

\[ [0, 9.1] \]

But there is another fundamental problem (due the lack of normality of the sample mean), which we discuss in the next slide.
We state the interval has 99% confidence in the interval, yet we have less. It is the equivalent to saying one is 99% sure that something is true, when in reality they are less sure. The stated confidence level is misleading.

The mismatch between the stated confidence level and the true confidence level is because the sample mean is not normally distributed. The real distribution is the green plot two slides ago (observe that it is right skewed).

Each horizontal line corresponds to a 99% confidence interval for the mean salary. The black vertical line corresponds to the mean salary. The green horizontal line contain the mean the red ones don’t. In this example we see that only 95 out of a 100 go through the mean.

<table>
<thead>
<tr>
<th>CI Level</th>
<th>Containing μ</th>
<th>Total</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>95</td>
<td>100</td>
<td>0.95</td>
</tr>
</tbody>
</table>

We state the interval has **99%** confidence in the interval, yet we have less. It is the equivalent to saying one is 99% sure that something is true, when in reality they are less sure.

The stated confidence level is **misleading**.

The mismatch between the stated confidence level and the true confidence level is because the sample mean is not normally distributed. The real distribution is the green plot two slides ago (observe that it is right skewed).
The delay in departures of 5 flights are noted. These delay times are 5.5, 6, 9, 22.5, 55 minutes (only delay times were recorded). The sample mean of this data set is 19.6 and standard deviation is 21 minutes. Construct a 99% confidence interval for the mean delay time. For this question use the normal distribution.

(A) [-4.5,43.7]  
(B) [0,43.7]  
(C) [-34.3,73.5]  

http://www.easypolls.net/poll.html?p=59d3bd22e4b0ede9e93921ed
Question Time

- A histogram and QQplot of delay times is given below

- Comment on the reliability of the CI in the previous question.

(A) The delay time data is right skewed.

(B) The sample mean is right skewed and we do not have 99% confidence in the interval.

(C) The sample mean is normally distributed and we have 99% confidence in the interval.

(D) [A] and [B]  

(E) [A] and [C]

- [http://www.easypolls.net/poll.html?p=59d3bd59e4b0ede9e93921ee](http://www.easypolls.net/poll.html?p=59d3bd59e4b0ede9e93921ee)
Example 3: Skewed data – sample size 50

- When the sample size is small, we need to be careful about the confidence statements we make, as they may not match the correct confidence level (and mislead the reader).
- This is because if the data is not normal, then for small sample sizes the average will not be normal (see Chapter 5).
- However, as the sample size grows the average using that sample will get closer to normal.
- We return to the salary data but draw a sample of size 50 and construct the 95% confidence interval for mean.
Non-normal data – sample size 50

The top plot is the histogram of NBL salaries.

The green plot is the histogram of the sample mean based 50 randomly sampled salaries.

The green plot has less spread (s.e. = $5.5/\sqrt{50} = 0.77$) than the top plot ($\sigma = 5.5$).

There is not an obvious skew in the distribution of the average. We can check by making a QQplot.
The QQplot of the averages shows that in the tails of the distributions there is still some deviation from normality. But is a lot “more normal” than averages of samples of size 3.

We also see from the plot of 100 99% confidence interval (drawn using this distribution) that only one does not cross the mean. Thus it seems the stated level of confidence (99%) matches quite well the true confidence of the interval.
I sample from the class these 5 heights 61, 63, 65, 66, 72.

The sample mean is 65.4.

The population standard deviation is known to be 4.3 inches.

The height data is not normal (we know this because it is a mix of two male/female subpopulations), but the sampling distributions app used in chapter 5 shows that the distribution of the sample mean was close to normal.

In reality we will not observe the entire population to run this app. But using just the observations and making a QQplot, suggests the sample mean based on 5 will not differ too much from normality.
Non-normal Data: Heights in class

- We now construct a 95% confidence interval for the population mean

\[
\left[ 65.4 - 1.96 \times \frac{4.3}{\sqrt{5}}, 65.4 + 1.96 \times \frac{4.3}{\sqrt{5}} \right] = [61.6, 69.2]
\]

- Using the sampling distribution applet in Chapter 5, the sample mean of the heights based on 5 is quite close to normal. Therefore we can be close to 95% confident the mean lies within this interval.
Confidence intervals: Calculation practice

- You want to rent an unfurnished one-bedroom apartment in Dallas. The average (sample mean) monthly rent over 10 randomly sampled apartments is 980 dollars. Assume that monthly rents follow a normal distribution with standard deviation 280 dollars.

- **Question:** Construct a 95% confidence interval for the mean monthly rent of a one-bedroom apartment.

- **Answer:** The standard error for the sample mean is $280/\sqrt{10} = 88.54$. The 95% CI is $[980 \pm 1.96 \times 88.54] = [806, 1153]$. With 95% confidence we believe the mean price of one-bedroom apartments in Dallas lies in this interval.

- Remember that the CI is referring to mean price of apartments and not the actual price of apartments.
Question Time

Does the 95% confidence interval [806,1153] mean that 95% of all rents prices should lie in this interval?

(A) No, this is a confidence interval for the mean not the apartment price. The mean cost lies somewhere in this interval (we do not know where).

(B) Yes, 95% of all rents will lie within this interval.

http://www.easypolls.net/poll.html?p=59d3bdb3e4b0ede9e93921ef
Question Time

A realtor wants to know if the mean price of one bedroom apartments in Dallas is more than 1100 dollars a month. Based on the 95% CI for the mean \([980 \pm 1.96 \times 88.54] = [806,1153]\), what can you say?

- **(A)** The 95% confidence interval for the mean is [806,1153] dollars. The mean can be anywhere in this interval. As this interval contains both values above and below 1100 dollars, we do not know. We do not have enough data to answer her question.
- **(B)** The majority of the confidence interval [806,1153] is less than 1100, which suggests the mean is less than 1100 dollars.

**Hint** The majority of this interval is less than $1100 simply because the sample mean 980 is less than 1100. If the same mean were greater than $1100, then the majority of the interval would be mean more than $1100.
Different levels of confidence

- As we mentioned a few slides earlier, there is no need to restrict ourselves to 95% confidence intervals.
- The level of confidence we use really depends on how much confidence we want. For example, you would expect a 99% confidence interval to be more likely to contain the mean than a 95% confidence interval.
- To construct a 99% confidence interval we use exactly the same prescription as that used to construct a 95% confidence interval, the only thing that changes is 1.96 goes to 2.57 (if you look up -2.57 in the z-tables you will see this corresponds to 0.5%, so 99% of the time the sample mean will lie within 2.57 standard errors from the mean).
  - A 99% CI for the mean one-bedroom apartment price is \([980 \pm 2.57 \times 88.54]\). Length of interval is \(2 \times 2.57 \times 88.54\)
  - A 90% CI for the mean one-bedroom apartment price is \([980 \pm 1.64 \times 88.54]\). Length of interval is \(2 \times 1.64 \times 88.54\)
Sample size and length of the CI

- Let us return to the apartment example. We recall that the 95% confidence interval for the mean price is $[980 \pm 1.96 \times 88.54] = [806,1153]$. The length of this interval is $2 \times 1.96 \times 88.54 = 347$.

- **Question:** Suppose I take a sample of 100 apartments in Dallas, the sample mean based on this sample is $1000$, construct the 95% CI for the mean.

- **Answer:** The standard error is $280/\sqrt{100} = 28$ (much smaller than when the sample size is 10), and the CI is $[1000 \pm 1.96 \times 28]$. The length of this interval is $2 \times 1.96 \times 28 = 109$.

- What we observe is:
  - The length of the interval **does not depend** on the sample mean, it simply locates the center of the interval.
  - The length depends on (i) the confidence level (ii) the standard deviation and (iii) the sample size.
  - The length of the interval gets smaller as the sample size increases.
Question Time

- Is a 100% confidence interval for the mean informative?
  - Yes
  - No

 http://www.easypolls.net/poll.html?p=59d3be45e4b0ede9e93921f1

- What influences the length of a confidence interval?
  - (A) The sample mean
  - (B) The sample size
  - (C) The confidence level
  - (D) A and B
  - (E) B and C

 http://www.easypolls.net/poll.html?p=59d3bebae4b0ede9e93921f3
The CLT in action: Poll results

The first plot is a histogram for the just one family.

The second plot is a histogram for the average amongst two families.

The standard deviation of the first plot is 1.5 and the standard deviation of the second plot is 1.

Observe that the average over two looks more bell shaped.
Using Margin of Error to calculate sample size
Learning Targets:

- Margin of Error:
  - Understand what a margin of error is.
  - Understand what factors contribute to the size of the margin of error.
  - Understand why the margin of error is important when designing an experiment.
  - Know how to calculate a sample size based on a given margin of error.
What is a good length for a confidence interval

- You read in a newspaper that
  
  *The proportion of the public that supports same-sex marriage is somewhere between 55% ± 15%.*

- Where did this interval come from?
  - It is impossible to interview the entire nation. Instead a survey was done. The proportion in the survey who supported same-sex marriage was 55%, the uncertainty (standard error) for this estimator evaluated and the (probably 95%) confidence interval for the population proportion is [55-15,55+15]% = [40%,70%].
  - This is an extremely large interval, it is so wide, that it is really not that informative about the opinion of the public.
  - As we will see on a later slide, the reason it is so wide is that the sample size is *too small*. A larger sample size is required to make the interval narrower and better locate the national proportion.
Motivating the margin of error

- Typically, before data is collected, we need to decide how large a sample to collect.
- This is usually done by deciding how much `above and below’ the estimator seems reasonable. For example, $[55-3,55+3]\% = [52,58]\%$ is more information.
- The 3% in the interval is known as a the **margin of error**.
- Based on any given margin of error we can determine the sample size (see formula on next page).
Margin of Error

- Margin of error is the name given for the plus and minus part in the confidence interval.

- The 95% confidence interval for the population mean is 
  \[ \text{sample mean} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} \], **the margin of error is** \( 1.96 \times \frac{\sigma}{\sqrt{n}} \).

- For example, in the previous example the margin of error for the CI based on 10 apartments is \( 1.96 \times 88.54 \).
- The margin of error for the CI based on 100 apartments is \( 1.96 \times 28 \).

- The margin of error is a measure of reliability. For a given confidence level, the smaller the margin error the more precisely we can pinpoint the true mean.

- Suppose we want the margin or error to be equal to some value, then we can find the sample size such that we obtain that margin of error. Solve for \( n \) the equation \( \text{MoE} = 1.96 \times \frac{\sigma}{\sqrt{n}} \) (the Margin of Error and the standard deviation \( \sigma \) are given): \( n = \left(1.96 \times \frac{\sigma}{\text{MoE}}\right)^2 \)

- This equation is given in the cheat sheet.
MoE: Calculation practice

- An company is collecting review for a pedometer watch. It will construct a 80% confidence interval for the mean rating of its pedometer watch.
- It is known that the standard deviation for the pedometer watch is about 1.2.
- The company would like the margin of error to be 0.2. How many people should they contact to obtain this margin of error?
- We know that the margin of error for an 80% CI is

\[
\text{MoE} = 1.28 \times \frac{\sigma}{\sqrt{n}}
\]

- \textbf{Answer:} Solving the above \(1.28 \times 1.2/\sqrt{n} = 0.2,\)
  \(\sqrt{n}=(1.28 \times 1.2/0.2) =7.68.\) Solving this gives \(n = 59.\)
- This means we need to question 59 people, such that the 80% confidence interval has margin of error 0.2.
MoE: When the standard deviation is unknown?

- In the previous example we assumed the standard deviation was known. In general before we collect the data, we will not have much information about the standard deviation. It will be unknown.

- However, we can make an educated guess on the range of values that the standard deviation takes.

- For example, the standard deviation for human heights is probably between 2-5 inches.

- Based on this information we can find the sample size whose Margin of Error is at most a certain length.
MoE: Calculation practice with unknown $\sigma$

- **Question** How large a sample size do we require such that the margin of error for a 95% confidence interval for the mean of human heights is maximum 0.25 inch, given that we do know that $\sigma$ lies somewhere between 2-5 inches.

- **Answer** We know that the formula is $n = (1.96 \times \sigma/0.25)^2$. We need to choose the standard deviation to place in the formula.
  - If we use $\sigma=2$, then the sample size we should choose is $n=(1.96 \times 2/0.25)^2 = 246$.
  - If we use $\sigma=5$, then the sample size we should choose is $n=(1.96 \times 5/0.25)^2 = 1537$.
  - For standard deviations between 2 and 5, the sample size will be between 246 – 1537.

- But we do not know $\sigma$ so which sample size to choose?
Suppose we choose $\sigma_1$ to determine the sample size

$$n = \left( \frac{1.96 \times \sigma_1}{0.25} \right)^2$$

But the real $\sigma$ was $\sigma_{\text{true}}$ then the real margin of error is

$$MoE = 1.96 \times \frac{\sigma_{\text{true}}}{\sqrt{n}} = 1.96 \times \frac{\sigma_{\text{true}}}{\left( \frac{1.96\times\sigma_1}{0.25} \right)}$$

$$= 0.25 \frac{\sigma_{\text{true}}}{\sigma_1}$$

We can see from the above this will be less or equal to 0.25 only if

$$\sigma_{\text{true}} < \sigma_1$$
This means always using what we believe is the **maximum standard deviation** in the calculation of margin of error i.e.,

\[ n = \left( \frac{1.96 \times \sigma_{MAX}}{\text{MoE}} \right)^2 \]

To be sure that the MoE is **maximum 0.25**, we need to use the largest sample size of \( n=1537 \).
A health agency wants to construct a 95% confidence interval for the mean height of a 5 year old. The standard deviation of a 5 year old is known to be somewhere between 1 to 3 inches. What is the minimum sample size required to ensure the margin of error is less than 0.5 inches?

(A) 16   (B) 44   (C) 62   (D) 139   (E) 156

http://www.easypolls.net/poll.html?p=59d3bf1fe4b0ede9e93921f7
Intervals and Margin of Error

- Suppose we are given the 95% confidence interval for the mean to be [10, 20].
  - The **sample mean** is in the middle of the interval which is 15.
  - The interval can be written as [15-5, 15+5].
  - The margin of error is 5.

- Suppose are given the 80% confidence interval [100, 140].
  - The sample mean is in the middle of the interval which is 120.
  - The interval can be written as [120-20, 120+20].
  - The margin of error is 20.
  - If the sample size is 5, we find \( \sigma \) by solving
    \[
    20 = 1.28 \times \frac{\sigma}{\sqrt{5}}
    \]
    \[
    \sigma = 20 \times \frac{\sqrt{5}}{1.28} = 35
    \]
**MoE: Calculation practice**

- **Question:** A 95% confidence interval for the mean length of parrots beaks is \([4,10] = [7-3,7+3]\) inches. The sample size is 20. By what factor should the sample size increase such that the margin of error is reduced to 1?

**Answer:** The margin of error (using formula) is

\[
3 = 1.96 \times \frac{\sigma}{\sqrt{20}}
\]

- You can solve for \(\sigma\) and then calculate the sample size using the formula.
- Or do the following: We want to decrease the MoE, such that MoE = 1. New margin of error is one third of the old one:

\[
\text{new } MoE = 1 = \frac{3}{3} = \frac{1}{3} \times \text{old } MoE = \frac{1}{3} \times 1.96 \times \frac{\sigma}{\sqrt{20}}
\]
Maths reminder

\[
\frac{1}{3\sqrt{20}} = \frac{1}{\sqrt{9}\sqrt{20}} = \frac{1}{\sqrt{9} \times 20} = \frac{1}{\sqrt{180}}
\]

Using the maths above:

\[
\text{new MoE} = 1 = 1.96 \times \frac{\sigma}{\sqrt{180}}
\]

We need to increase the sample size from 20 to 180 (interview 180 people!), this is a 9 fold increase to reduce the margin of error from 3 to 1.

This is a substantial increase in the sample size.
**MoE: Calculation practice (tricky)**

- **Question:** A 95% confidence interval for the mean length of parrots beaks is \([4,10] = [7-3,7+3]\) inches. It is based on a sample of size \(n\). By what factor should the sample size increase such that the margin of error is 1?

- **Answer:** This looks like an impossible question because we don’t have any obvious information on the standard deviation or sample size. But we can break the problem into steps:
  
  - The margin of error is \(3 = 1.96 \times \sigma/\sqrt{n}\).
  
  - We cannot solve for \(\sigma\) because \(n\) is not given. So we have to use the same trick given on the previous slide. To reduce by a margin of error by a third:

\[
\text{new MoE} = 1 = \frac{3}{3} = \frac{\text{old MoE}}{3} = \frac{1}{3} \times 1.96 \times \frac{\sigma}{\sqrt{n}}
\]
Again we use some maths (to take $3$ into the square root):

\[
\frac{1}{3\sqrt{n}} = \frac{1}{\sqrt{9\sqrt{n}}} = \frac{1}{\sqrt{9n}}
\]

Using the maths above:

\[
\text{new MoE} = 1 = \frac{3}{3} = \frac{1}{3} \times 1.96 \times \frac{\sigma}{\sqrt{9n}}
\]

We need to increase the sample size from $n$ to $9n$ this is a 9 fold increase to reduce the margin of error from 3 to 1.

This is exactly what we did previously when $n = 20$. 
Example If a sample size of 20 give the 95% confidence interval for the mean to be [2,10], how large a sample size is required to reduce the margin of error to 1/2 (0.5)?

Solution Since the confidence interval is [2,10] = [6-4,6+4], thus

\[\text{MoE} = 4 = 1.96 \times \frac{\sigma}{\sqrt{20}}\]

new \(\text{MoE} = 0.5 = \frac{4}{8} = \frac{4}{8} \times 1.96 \times \frac{\sigma}{\sqrt{20}} = 1.96 \times \frac{\sigma}{\sqrt{8^2 \times 20}}\)

This means increasing sample size from 20 to 1280.

We see that to decrease the margin of error from 4 to \(\frac{1}{2}\) (by an eighth) we need to increase the sample size by factor 64!
Confidence intervals and the t-distribution
Learning targets:
- Understand that the t-distribution is only used because typically the population standard deviation is rarely ever known. Instead it needs to be estimated from the data.
- Use the t-distribution to construct confidence intervals.

Conditions for using the t-distribution.
- Observations are a SRS
- If sample size is small observations are close to normal.
So far we have assumed that the standard deviation is known, even though the mean is unknown.

In some situations, this is realistic. For example, in the (previous) potassium example, it seems reasonable to suppose that the amount of variation for everyone is the same (and after years of data collection this is known). But everyone has their own personal mean level, which is unknown.

However, in most situations, the population standard deviation unknown.
Estimating the standard deviation

- Given the data: 68, 68.5, 68.9 and 69.4 the sample mean is 68.7, how to `get' the standard deviation to construct a confidence interval?

- We do not know the standard deviation, but we can estimate it using the formula

\[ s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \]

- For our example it is

\[ s = \sqrt{\frac{1}{3} ([-0.7]^2 + [-0.2]^2 + [0.2]^2 + [0.7]^2)} = 0.59 \]

- Once we have estimated the standard deviation we replace the unknown true standard deviation in the z-transform with the estimated standard deviation:

\[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{\bar{X} - \mu}{s/\sqrt{n}} \]

\[ \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \rightarrow \bar{X} \pm 1.96 \frac{s}{\sqrt{n}} \]
Using the z-transform with the estimated standard deviation

\[
\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{\bar{X} - \mu}{s/\sqrt{n}}  \quad \quad \quad \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \Rightarrow \bar{X} \pm 1.96 \frac{s}{\sqrt{n}}
\]

- After this we *could* conduct the analysis just as before. However, we will show in the next few slides (with the aid of Statcrunch) that this strategy leads to misleading confidence levels. The real level of confidence will be less than the stated level (equivalent to saying one is 95% confident a statement is true, when in reality they are only 85% confident).
To illustrate the problem of estimating the standard deviation and carrying on as usual we consider some examples:

- We consider heights which are *normally distributed* with mean 67 and standard deviation 3.8.
- This is a thought experiment. We will draw (get heights) from this distribution, but we shall pretend we do not know the mean or standard deviation.
- We will construct a 95% confidence interval for the mean height based on the sample mean. We estimate the standard deviation using the data.
- We consider the case the sample size is \( n = 3 \).
- We consider the case the sample size is \( n = 50 \).
- Everything is as in the perfect example given at the start of this chapter. The only *difference* is that we estimate the standard deviation from the data (previously the standard deviation was given).
Example 1: Normal data – sample size 3 using normal dist.

The original data is normal.

Three heights are drawn from this distribution. For one typical sample the sample mean is 69.9 and **sample standard deviation is 1.73**. This sample standard deviation clearly **underestimates the true standard deviation of 3.8**.

This is the density of the sample mean. The means are aligned but the spread is less than the spread in the original data.
For the data given on the previous slide. If we ignore the fact that we estimate the standard deviation, the regular confidence interval is
\[
\left[69.9 - 1.96 \times \frac{1.73}{\sqrt{3}}, 69.6 + 1.96 \times \frac{1.73}{\sqrt{3}}\right]
\]

1.73 is the estimated standard deviation. The true standard deviation is 3.4. It is clear that the above interval is narrower than it should be.

To investigate the impact it has on the true level of confidence. We draw several samples of size 3, estimate the mean, standard deviation and construct the 95% CI for the mean.
Each horizontal line is a CI. The vertical line is the true mean 67. The read lines do not intersect with the mean. We see that only 84% CI cross the mean. **84% is a lot smaller than then stated confidence level. This is not a 95% CI for the mean.**
But the data is normal. Therefore the sample mean is normal.

This means it **cannot** be an issue of the central limit theorem not kicking in.

There is another problem.

This problem is that in addition to estimating the mean using the sample mean (this is why we have constructed a confidence interval), we are **also** estimating the standard deviation.

The estimated standard deviation is causing an additional amount of “uncertainty” that so far we have not accounted for.

This is why the true level of confidence does not match the stated level.
We now draw a sample of size 50 from a normal distribution.

For this example, the estimated standard deviation 4.07 is far closer to the true 3.8. This in general is true for large sample sizes.

The distribution of the sample mean is also normal, but far narrower than the original data.

For the example given on the right the 95% CI for where the mean lies is

\[
\left[ 68.0 - 1.96 \times \frac{4.07}{\sqrt{50}}, 68.0 + 1.96 \times \frac{4.07}{\sqrt{50}} \right]
\]
The 95% confidence interval for the previous example is

\[
\left[ 68.0 - 1.96 \times \frac{4.07}{\sqrt{50}}, 68.0 + 1.96 \times \frac{4.07}{\sqrt{50}} \right]
\]

4.07 is close to the true standard deviation 3.8. Thus the length of the interval has not been hugely effected when estimating the standard deviation from the data.

Looking at the number of times the mean is contained within the 95% confidence interval (on the right) we see that 97 intervals intersect the population mean. This tells us that the confidence interval is close to the stated 95% level of confidence.

Same normal distribution (no need to use CLT here), the only difference is the sample size. Why the difference???
- We observe that as we increase the sample size, the level of confidence seems to match the true level of confidence.
- This has **nothing to do with the CLT kicking in**. The data is coming from a normal distributions, so its sample mean is normal.
- The sample standard deviation is becoming a better estimate of the true sample standard deviation.
Take home message from the thought experiment

- Simply replacing the true standard deviation with the estimated standard deviation seems to have severe consequences on the actual confidence we have in the interval.
  - It is like saying “trust me I am sure the mean is in there”. When it is not.
- When the sample size was small there tends to be an underestimation in the standard error, resulting in the stated 95% confidence interval having a lower (unknown) level of confidence.
- To see why consider the z-transforms of the sample mean with known and estimated standard deviations:
  - \( \frac{(\text{sample mean} - \mu)}{\sigma/\sqrt{n}} \)
  - \( \frac{(\text{sample mean} - \mu)}{s/\sqrt{n}} \)
- In the next few slides we show that when we estimate the standard deviation the z-transform no longer has a standard normal distribution, but has instead a t-distribution (we should call it a t-transform).
That nice Mr. Gosset

- Just over 100 years ago, W.S. Gosset was a biometrician who worked for Guiness Brewery in Dublin, Ireland.
- His hobby was statistics.
- Gosset realized that his inferences with small sample data seemed to be incorrect too often – his true confidence level was less than stated. We observed the same in the simulations.
- He worked out the proper method that took into account substituting $s$ for $\sigma$.
- But he had to publish under a pseudonym: Student (probably because Gosset was a sweet and modest person).
Just like the sample mean is random with a distribution, so is the sample standard deviation.

Here we take a sample of size 10 from a normal distribution and calculate its sample mean and variance.

In Statcrunch: Applets -> Sampling distributions
- Select the population
- Select the Statistic(s)
  - First choose Mean
  - Second choose Std. Dev.
- Compute
  - Choose Sample size.
  - Press 1000 times (several times).
Sample means and standard deviations

This is the original distribution with mean 64.5 and standard deviation 2.5

Here is a “typical” sample of size 10. The sample mean and standard deviation of this sample is 64.7 and 2.9.

Repeating the above several times. This is the histogram of the sample means.

Repeating the above several times this is the histogram for the standard deviation. Observe it is positive and slightly skewed.
Sample sizes and estimated standard deviation

- The sampling distribution of the sample standard deviation (n=5)
- The sample distribution of the sample standard deviation (n=30)

Observe that as the sample size increases the estimator of the sample standard deviation becomes less variable (1.70 reduces to 0.65). Large amount of variability in the sample standard deviation influences the confidence interval.
An extreme example. We observe two male heights 68 and 70 inches.

The sample mean and sample standard deviation are

\[
\text{sample mean } = \bar{x} = \frac{1}{2} (68 + 70) = 69
\]

\[
\text{sample standard deviation } = s = \sqrt{\frac{1}{1} \left[ (68 - 69)^2 + (70 - 69)^2 \right]} = 1.41
\]

Because 69 is simply an estimate of the mean, we need to construct a confidence interval about 69, for where we believe the true, population mean lies.
1.41 measures the average spread of 68 and 70, but it is a terrible estimate for the standard deviation of all heights. This means we cannot have 95% confidence in this interval when the sample standard deviation is so bad.

\[
\left[ 69 - 1.96 \times \frac{1.41}{\sqrt{2}}, 69 + 1.96 \times \frac{1.41}{\sqrt{2}} \right] = [67, 71]
\]

To correct for the lousy standard deviation estimate, we widen the interval. The correct 95% confidence interval for locating the population mean is

\[
\left[ 69 - 12.71 \times \frac{1.41}{\sqrt{2}}, 69 + 12.71 \times \frac{1.41}{\sqrt{2}} \right] = [56.6, 81.3]
\]
Student’s $t$ distributions

- When $\sigma$ is estimated from the sample standard deviation $s$, the sampling distribution for $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ will depend on the sample size.

The sample distribution of

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is a $t$ distribution with $n - 1$ degrees of freedom.

- The degrees of freedom ($df$) is a measure of how well $s$ estimates $\sigma$. The larger the degrees of freedom, the better $\sigma$ is estimated.

- This means we need a new set of tables!

- Further reading:
Distribution of the t-transform

We see a few slides back in the computer simulation that when \( n \) is very large, \( s \) is a very good estimate of \( \sigma \), and the corresponding \( t \) distributions are very close to the normal distribution.

The \( t \) distributions become wider (thicker tailed) for smaller sample sizes, reflecting that \( s \) can be smaller than \( \sigma \), so the corresponding \( t \)-transform is more likely to take extreme values than the \( z \)-transform.
Impact on confidence intervals

Suppose we want to construct the 95% confidence interval for the mean. The standard deviation is unknown, so as well as estimating the mean we also estimate the standard deviation from the sample. The 95% confidence interval is:

\[
\left[ \bar{X} - t_{n-1}(2.5) \times \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1}(2.5) \times \frac{s}{\sqrt{n}} \right]
\]

The blue area is proportion and for the 95% corresponds to 2.5%
Examples

Sample size n=3.
The 95% CI for the mean is

\[\bar{X} - 4.3 \times \frac{s}{\sqrt{3}}, \bar{X} + 4.3 \times \frac{s}{\sqrt{3}}\]

MoE = \(4.3 \times \frac{s}{\sqrt{3}}\)

Sample size n = 10.
The 95% CI for the mean is

\[\bar{X} - 2.26 \times \frac{s}{\sqrt{10}}, \bar{X} + 2.26 \times \frac{s}{\sqrt{10}}\]

MoE = \(2.26 \times \frac{s}{\sqrt{10}}\)

You observe that these confidence intervals are wider than the confidence intervals using a normal distribution. This is to compensate for the estimation of the standard deviation from the data.
As the sample size grows the degrees of freedom grow. This means going down the table to obtain the confidence levels.

For a very large sample size (n=1000), using either the t-distribution or the normal distribution give almost the same result. This is because when the sample size is 1000, the estimate of the standard deviation is likely to be very close to the population standard deviation.

This also explains what we have observed previously. When the sample size is 50, we do not need to compensate much for estimating the standard deviation.

The confidence level is given at the bottom of the table.
Example 3: Normal data – sample size 3, using t-dist

We return to the previous example, where the sample size is three, the sample mean is 4.3 and sample standard deviation 4.3. The correct 95% is

$$
\left[ 69.9 - 4.3 \times \frac{1.73}{\sqrt{3}}, 69.6 + 4.3 \times \frac{1.73}{\sqrt{3}} \right]
$$

See whether we really do have 95% confidence in this interval. We do a thought experiment and construct 100 confidence intervals using the t-distribution (with 2 df). We observe that 95 of them intersect with the vertical line. This tells us we really do have 95% confidence that the population mean lies in the interval we have constructed.
t-values

- t-values are like z-values (but are based on the t-distribution). We will use them frequently in Chapter 7.
- We practice using them here.
- Unlike the normal tables. The values inside the t-table are the t-values and not probabilities.
- We focus in the t-distribution with 10df.
  - 1.812 means the area to the right of 1.812 is 5%.
  - By symmetry of the t-distribution, the area to the left of -1.812 is 5%.
  - The area between -1.812 to 1.812 is 90%.
  - The area to the right of 2.228 is 2.5%.
  - By symmetry of the t-distribution, the area to the left of -2.228 is 2.5%.
  - The area between -2.228 and 2.228 is 95%
- We can use the t-tables to obtain bounds for probabilities.
Again we focus on the t-distribution with 10df.

The t-value is 2, what is the area to the **right** of 2?

- Since 2 is between 1.812 (area to right is 5%) and 2.228 (area to the right is 2.5%). The area to the right of 2 is between 2.5-5%

The t-value is 2, what is the area to the **left** of 2?

- Since 2 is between 1.812 (area to right is 5%) and 2.228 (area to the right is 2.5%). The area to the left of 2 is between 95-97.5%

The t-value is -2, what is the area to the **left** of -2?

- The area to the left of -2 is the same as the area to the right of 2.
- Using the above, this means the area to the left of -2 is between 2.5-5%
Question Time

- The t-value for a t-distribution with 12 degrees of freedom is 1.25. What is the area to the right of 1.25?
  - (A) less than 10%
  - (B) between 10-15%
  - (C) between 85-90%
  - [Link](http://www.easypolls.net/poll.html?p=59dbe70ae4b0601923a70256)

- The t-value for a t-distribution with 12 degrees of freedom is -0.8. What is the area to the left of -0.8?
  - (A) The number is not on the table
  - (B) Less than 20%
  - (C) between 20-25%
  - (D) between 75-80%
  - [Link](http://www.easypolls.net/poll.html?p=59dbe73be4b0601923a70258)
Let us return to the example of prices of apartments in Dallas. 10 apartments are randomly sampled. The **sample mean and the sample standard deviation** based on this sample is 980 dollars and 250 dollars (**both are estimators** based on a sample of size ten). Construct a 95% confidence interval for the mean.

(A) The 95% confidence interval for the mean price is \([980 \pm 2.262 \times 79]=\left[801,1159\right]\).

(B) The 95% confidence interval for the mean price is \([980 \pm 2.262 \times 250]\).

(C) The 95% confidence interval for the mean price is \([980 \pm 2.228 \times 79]\).

http://www.easypolls.net/poll.html?p=59dbe779e4b0601923a7025b
The t-distribution does not correct for non-normal data

The t-distribution is **used only** to correct for estimating the standard deviation from the data. It cannot correct for lack of normality of the sample mean.

**Example:** We draw a sample of size three from skewed distribution. The distribution of the averages is clearly skewed. For each sample the 95% CI for the mean is evaluated (on the right) using the t-distribution. Observe that only 88 contain the mean (less confidence than we have stated). The t-distribution cannot correct for the lack of normality of the sample mean.
Example: Large sample sizes and CIs

- We return to the NBL salary data. Where in reality **both** the mean and standard deviation are estimated from the data:

  ![Summary statistics table]

  - We use a t-distribution with 816 degrees of freedom.

  ![Probability and t* values table]

  - With 95% confidence we believe the mean salary lies in the interval

    \[
    [4.3 \pm 1.96 \times 0.19] = [3.92, 4.67]
    \]
To see if we really do have the stated level of confidence in this interval we use the Statcrunch app to see what happens over several replications.

Despite the salaries being hugely skewed, when the sample size is 817 the sample mean is close to normal and we see from the applet that the confidence intervals have close to 95% confidence (compare 95 with 96).
The flight delay times of planes leaving an airport in California are monitored (on time flights or early departures are not included, hence no negative times). The delay in departures of 6 flights are noted. These delay times are 5.5, 10.5, 13, 22.5, 45, 55 minutes. The sample mean of this data set is 25.25 and sample standard deviation is 20.2 minutes. Construct a 99% confidence interval for the mean delay time (use a t-distribution with 5df).

(A) \([25.25 \pm 4.032 \times 20.2]\)  \(\text{(B)} \ [0, 25.25 + 4.032 \times 8.24]\)  \(\text{(C)} \ [25.25 \pm 3.65 \times 20.2]\)

(D) \([25.25 \pm 3.65 \times 8.24]\)  \(\text{(E)} \ [0, 20.2 + 4.032 \times 25.25].\)

http://www.easypolls.net/poll.html?p=59dbe7d8e4b0601923a7025d
Question Time

Based on the plots below, comment on the reliability of the confidence interval constructed in the previous question.

(A) The delay time data is right skewed.
(B) The sample mean is right skewed and we do not have 99% confidence in the interval.
(C) The sample mean is normally distributed and we have 99% confidence in the interval.
(D) [A] and [B]  (E) [A] and [C]

http://www.easypolls.net/poll.html?p=59dbe8fce4b0601923a70260
Statcrunch: Confidence intervals

- Load data into Statcrunch.
- T Stats -> One Sample -> With Data
- Highlight column.
- Check Confidence interval box and state level (default is 0.95 = 95%)
Example 1: Amazon product scores

What the data looks like

Histogram of data, which is also given by Amazon.

I tend to only buy products which get over 4 stars. Can I be sure that this tracker would get an average of over 4 stars if all customers bought it?
Summary Statistics

- The sample mean is 4.2
- The sample standard deviation is 1.37.
- The standard error is \( \frac{1.37}{\sqrt{174}} = 0.104 \)

The Statcrunch output:

- The 95% confidence interval for where the true mean score rating of the tracker is \( [4.25 \pm 1.973 \times 0.103] = [4.04, 4.57] \)
- The entire interval is above 4, suggesting that if all Amazon customers bought the tracker, the mean would be more than 4 stars. But..
We really do need to check if we have 95% confidence in this interval. The average of 4.2 is a number from the green histogram.

As we can see from the QQplot of the green histogram, it is close to normal.

Thus we really do have pretty much close to 95% confidence in the interval.

Of course we need to be a little careful when analyzing Amazon scores. The people who tend to volunteer scores may be biased. Often you grade a product if it is really bad or so good you want everyone to know.
Amazon example 2

- Recall the watch
- 4.8 is the sample mean based on 58 reviews.

4.8 is the sample mean it is one number from from the green histogram on the right.

In the analysis we are assuming this histogram is normal. *Even though it looks a bit left skewed.*
The 95% confidence interval for the mean is [4.70, 4.99]

One sample T confidence interval:
\[ \mu : \text{Mean of variable} \]

95% confidence interval results:
<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample Mean</th>
<th>Std. Err.</th>
<th>DF</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>LintelekSm</td>
<td>4.8448276</td>
<td>0.072970521</td>
<td>57</td>
<td>4.6987066</td>
<td>4.9909485</td>
</tr>
</tbody>
</table>

This is constructed under the assumption that the sample mean is normal. But we recall from the green histogram on the previous page that it does not look so normal. This is confirmed by looking at its QQplot.

![Sample means QQplot](image)

The sample mean does not look very normal.

The implication of this is…
we state that with 95% confidence interval for the mean is [4.70, 4.99]

In reality, we see from the plot below (based on a though experiment using several replications) we only have about 89% confidence in the interval.

We are less than 95% confident that the mean is contained in [4.70, 4.99].

But since this interval is so much above 4.0. It seems unlikely the true mean is 4.0 or less.
Example: Red Wine 1

It has been suggested that drinking red wine in moderation may protect against heart attacks. This is because red wine contains polyphenols which act on blood cholesterol. To see if moderate red wine consumption increases the average blood level of polyphenols, a group of nine randomly selected healthy men were assigned to drink half a bottle of red wine daily for two weeks. The percent change in their blood polyphenol levels are presented here:

0.7, 3.5, 4.0, 4.9, 5.5, 7.0, 7.4, 8.1, 8.4

Sample average = 5.50
Sample standard deviation $s = 2.517$
Degrees of freedom $df = n - 1 = 8$

Summary statistics:

<table>
<thead>
<tr>
<th>Column</th>
<th>n</th>
<th>Mean</th>
<th>Variance</th>
<th>Std. dev.</th>
<th>Std. err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyphenol(9)</td>
<td>9</td>
<td>5.5</td>
<td>6.335</td>
<td>2.5169426</td>
<td>0.83898086</td>
</tr>
</tbody>
</table>
What is the 95% confidence interval for the mean percent change?

First, we determine what \( t^* \) is. The degrees of freedom are \( df = n - 1 = 8 \) and \( C = 95\% \).

\[
8 \quad 0.706 \quad 0.889 \quad 1.108 \quad 1.397 \quad 1.860 \quad 2.306 \quad 2.449 \quad 2.896 \quad 3.355 \quad 3.833 \quad 4.501 \quad 5.041
\]

The margin of error \( m \) is:

\[
m = t^* \times \frac{s}{\sqrt{n}} = 2.306 \times \frac{2.517}{\sqrt{9}} \approx 1.93.
\]

So the 95% confidence interval is \( 5.50 \pm 1.93 \), or 3.57 to 7.43.

We can say “With 95% confidence, the mean of percent increase is between 3.57% and 7.43%.” The corresponding Statcrinch output is below:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample Mean</th>
<th>Std. Err.</th>
<th>DF</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyphenol(9)</td>
<td>5.5</td>
<td>0.83898086</td>
<td>8</td>
<td>3.5653067</td>
<td>7.4346933</td>
</tr>
</tbody>
</table>
Example continued: checking the reliability of the analysis

In order to check if we really have close to the stated 95% confidence interval for the mean we run the statcrunch applet on the polyphenol data set (using sample size = 9 and the t-distribution).

We see that 95 out of the 100 (pseudo confidence intervals) contain the mean. Thus we can say we are 95% confident the mean change in polyphenol level lies in the interval [4.04, 4.57].
Example: Red wine 2

Let us return to the same study, but this time we increase the sample size to 15 male volunteers. The data is

\[0.7, 3.5, 4, 4.9, 5.5, 7, 7.4, 8.1, 8.4, 3.2, 0.8, 4.3, -0.2, -0.6, 7.5\]

The sample mean in this case is 4.3 and the sample standard deviation is 3.06.

The df = 14.

Since the sample size has increased, it is likely that the sample standard deviation is closer to the true standard deviation, thus the t-value is now 2.145.

Thus with 95% confidence we believe the true mean change in polyphenol level lies in

\[
4.3 \pm 2.145 \times \frac{3.06}{\sqrt{15}} = [2.6, 6]
\]
To understand whether a 95% confidence interval constructed from the data (using the $t$-distribution) is really a 95% confidence interval, 1000 confidence intervals were constructed. The results are summarized in the applet below. Based on the applet, which statement(s) are correct?

(A) The $t$-distribution is correcting for the lack of normality in the data.

(B) We really do have 95% confidence in this interval.

(C) We seem to have 80.2% confidence in this interval.

(D) [A] and [C]  

(E) [B] and [C].

http://www.easypolls.net/poll.html?p=59dbe955e4b0601923a70262
Statcrunch: Confidence intervals

- Statcrunch will construct the confidence interval for you. Therefore it is important to connect the calculations with the statistical output.
- Stat -> T Stats -> One Sample -> With Data
- You will get the following drop down menu.

The box on the right is the output (it is superimposed on the window used to generate the output). Observe that L.Limit – U. limit gives the confidence interval [2.6, 6] calculated on the previous slide. DF = 14, matches with the degrees of freedom.
Interpretation of confidence, again

- The confidence level $C$ is the proportion of all possible random samples (of size $n$) that will give results leading to a correct conclusion, for a specific method.

- In other words, if many random samples were obtained and confidence intervals were constructed from their data with $C = 95\%$ then 95\% of the intervals would contain the true parameter value.

- In the same way, if an investigator always uses $C = 95\%$ then 95\% of the confidence intervals he constructs will contain the parameter value being estimated.

- But they never knows which ones do!

- Changing the method (such as changing the value of $t^*$) will change the confidence level.

- *Once computed*, any individual confidence interval either will or will not contain the true population parameter value. It is not random.

- It is *not correct* to say $C$ is the probability that the true value falls in the particular interval you have computed.
Accompanying problems associated with this Chapter

- Quiz 6
- Quiz 7
- Quiz 7a
- Quiz 8
- Homework 4 (part of it)