



Confidence intervals and the t-distribution

Topic: Unknown standard deviation 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.

An unknown: the standard deviation

- ❑ 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 potassium level example (in Chapter 7), the data has been collected over years. And it was seen that the amount of variation of potassium samples for an individual is about the same for **all individuals** but the mean level depends on the individual
- ❑ However, **in most** situations, the population standard deviation unknown.

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

Estimating the standard deviation

- Given the data: 68, 68.5, 68.9 and 69.4 the sample mean is 68.7, how to 'estimate' the standard deviation to construct a confidence interval?
- We do not know the standard deviation, but we can **estimate** it using the formula (you do not have to do it)

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

$$68.7 = \frac{1}{4} [68 + 68.5 + 68.9 + 69.4]$$
$$= \bar{X}$$

- For our example it is

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

$$\left[\bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}} \right] = \left[\bar{X} \pm 1.96 \times \frac{S}{\sqrt{n}} \right]$$


→ Is this a 95% CI for the mean?

Estimating the standard deviation

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

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

z - transform


$$\frac{\bar{X} - \mu}{\underbrace{\sigma / \sqrt{n}}_{s.e.}} \Rightarrow \frac{\bar{X} - \mu}{s / \sqrt{n}} = \text{is normal?}$$

$\sim N(0, 1)$

Using the z-transform with the estimated standard deviation

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

- ❑ We *could* conduct the analysis just as before.
- ❑ However, we will show in the next few slides that this strategy leads to misleading confidence levels.
- ❑ The real level of confidence will be less than the claimed level.

- To illustrate the problem of estimating the standard deviation and carrying on as usual we consider a specific example:

□ We consider the population of heights which are **normally distributed** with mean 67 and standard deviation 3.8. $\Rightarrow \sigma = 3.8$ $\mu = 67$

- This is a thought experiment. We will draw (sample) 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.

$$\left[\bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}} \right]$$

- We separately consider the two cases:
 - The sample size is $n = 3$.
 - The sample size is $n = 50$.
- The height data is normal. The only **difference** between what we are doing now and what we did previously is that we estimate the standard deviation from the data (previously the standard deviation was given).
- **What we are doing here has nothing to do with the CLT. Do get confused with this.**

Example 1: Normal data – sample size 3 using normal dist.

Review: The original data is **normal**.

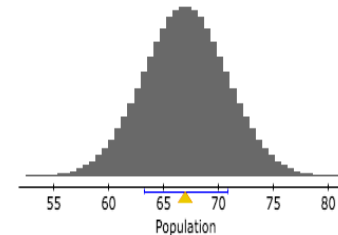
Three heights are drawn from this distribution. For one typical sample the **sample mean is 69.9** and **sample standard deviation is 1.73**.

This sample standard deviation clearly **underestimates the true standard deviation of 3.8**.

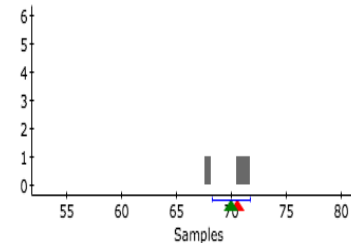
This is the density of the sample mean. The means are aligned but the spread is less than the spread in the original data.

Sampling Distributions

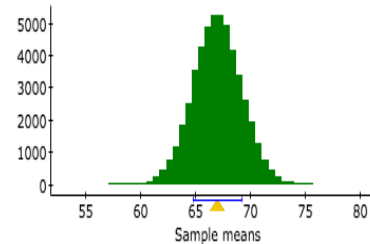
1 time 5 times 1000 times Reset Info



Population	
Mean	67 = μ
Median	67
Std. dev.	3.8 = σ



Samples	
Sample size	3
Mean	69.9587 = \bar{x}
Median	70.5702
Std. dev.	1.7323 = s



Sample means	
# of Samples	50001
Mean	67.0141
Median	67.0097
Std. dev.	2.1974

$$= \frac{3.8}{\sqrt{3}}$$

For the data given on the previous slide. If we ignore the that we estimate the standard deviation, the regular confidence interval is

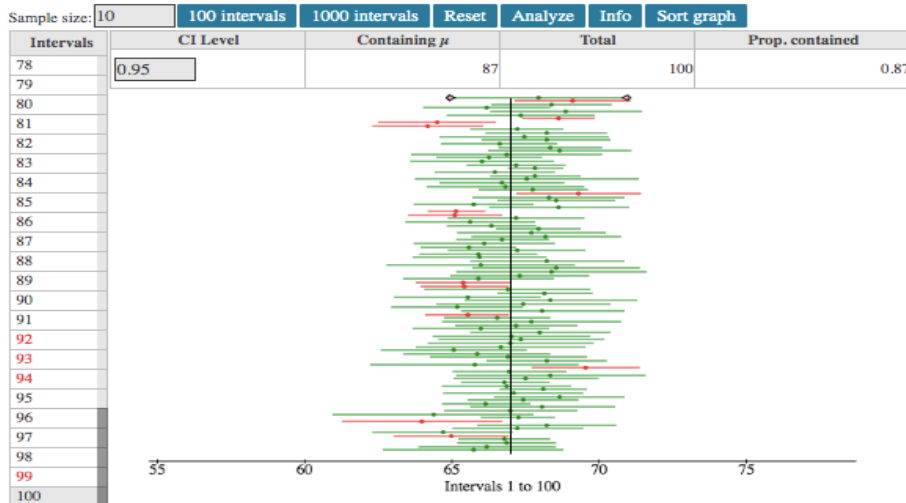
$$\left[69.9 - 1.96 \times \frac{1.73}{\sqrt{3}}, 69.6 + 1.96 \times \frac{1.73}{\sqrt{3}} \right]$$

1.73 estimated from data.

✓ $\left[69.9 - 1.96 \times \frac{3.4}{\sqrt{3}}, 69.6 + 1.96 \times \frac{3.4}{\sqrt{3}} \right]$

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.

Confidence intervals for a mean: Normal population ($\mu=67, \sigma=3.8$) Type=Z (with s)



In the 100 intervals on the left only 87 contain the population mean. We do not have 95% confidence in the interval.

- ❑ But the data is normal. Therefore the sample mean is normal.
- ❑ This means it **cannot** be an issue of the central limit theorem not holding.
- ❑ There is another problem.
- ❑ This issue is that in addition to estimating the mean we are **also** estimating the standard deviation. We have not accounted for the uncertainty in the sample standard deviation.
- ❑ Using the normal distribution when we estimate the standard deviation gives the wrong results!

Make a guess

- ❑ What do you think happens when we increase the sample size.
- ❑ Will be estimate of the population standard deviation be as bad?

Example 2: Normal data – sample size 50 using normal

We now draw a sample of size 50 from a normal distribution.

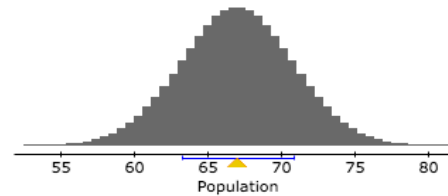
For the example given on the right the 95% CI for where the mean lies is

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

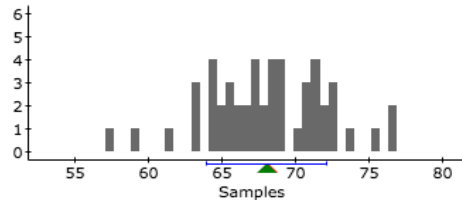
$$\left[68 - 1.96 \times \frac{3.9}{\sqrt{50}}, 69 + 1.96 \times \frac{3.9}{\sqrt{50}} \right] \checkmark$$

Sampling Distributions

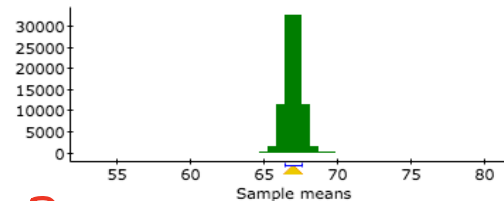
1 time 5 times 1000 times Reset Info



Population	
Mean	67
Median	67
Std. dev.	3.8



Samples	
Sample size	50
Mean	68.0429 = \bar{x}
Median	68.1314
Std. dev.	4.0784 = s



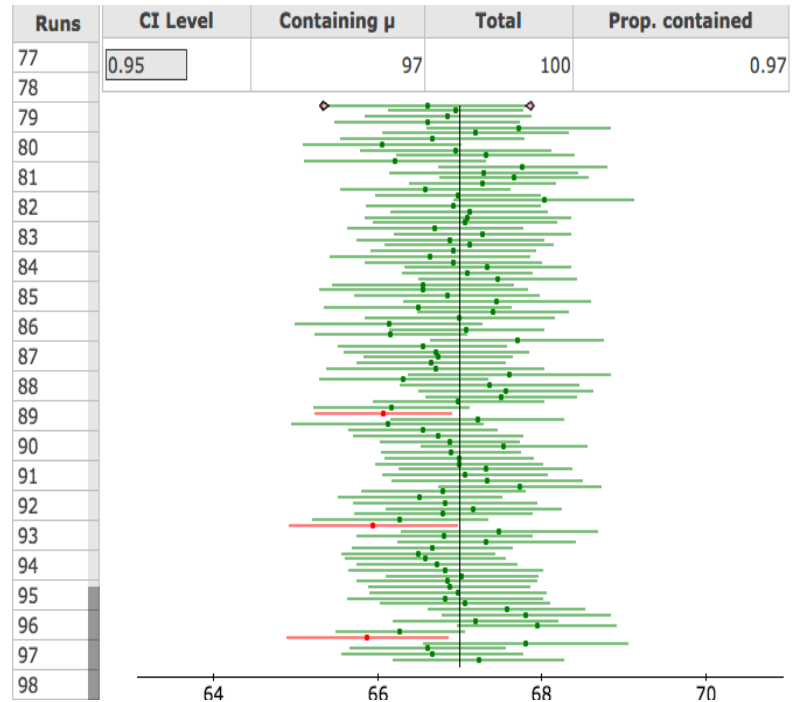
Sample means	
# of Samples	91000
Mean	67.0008
Median	67.0009
Std. dev.	0.5382

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 population mean is contained within 97 of the confidence intervals.

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. The sample mean is normal.
- ❑ The reason is that the sample standard deviation is becoming a **better estimate of the true sample standard deviation.**
- ❑ This is a principle in statistics, the larger the sample size the better the estimator tends to be.

Take home message from the thought experiment

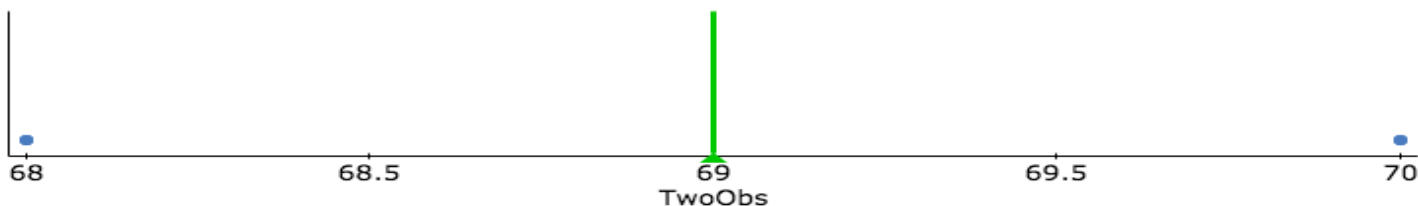
- Simply replacing the true standard deviation with the estimated standard deviation seems to have severe consequences on the actual confidence we have in the interval.
 - It is like saying “trust me I am sure the mean is in there”.
When it is not.
- When the sample size is small there tends to be an underestimation in the standard error, resulting in the stated 95% confidence interval having a lower level of confidence.

Estimating the mean with two observations

- ❑ An extreme example. We observe two male heights 68 and 70 inches.
- ❑ The sample mean and sample standard deviation are

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

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



Because 69 is simply a 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.

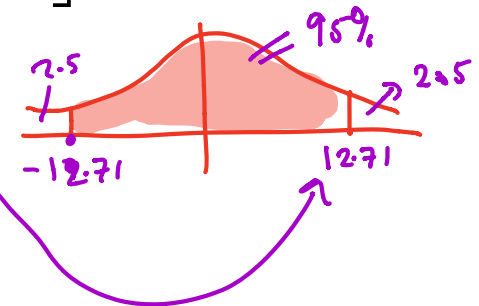
- The **incorrect interval** is

$$\left[\bar{x} - 1.96 \times \frac{s}{\sqrt{2}}, \bar{x} + 1.96 \times \frac{s}{\sqrt{2}} \right] = [67, 71]$$

- To correct for the bad standard deviation estimate. We widen the interval. The **correct** 95% confidence interval for locating the population mean is

$$\left[69 - 12.71 \times \frac{1.41}{\sqrt{2}}, 69 + 12.71 \times \frac{1.41}{\sqrt{2}} \right] = [56.6, 81.3]$$

df	Upper-tail pr					
	.25	.20	.15	.10	.05	.025
1	1.000	1.376	1.963	3.078	6.314	12.71
2	0.816	1.061	1.386	1.886	2.920	4.303
	50%	60%	70%	80%	90%	95%
	Confidence					



$n = 2$
 $(2 - 1)$

Student's t distributions

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

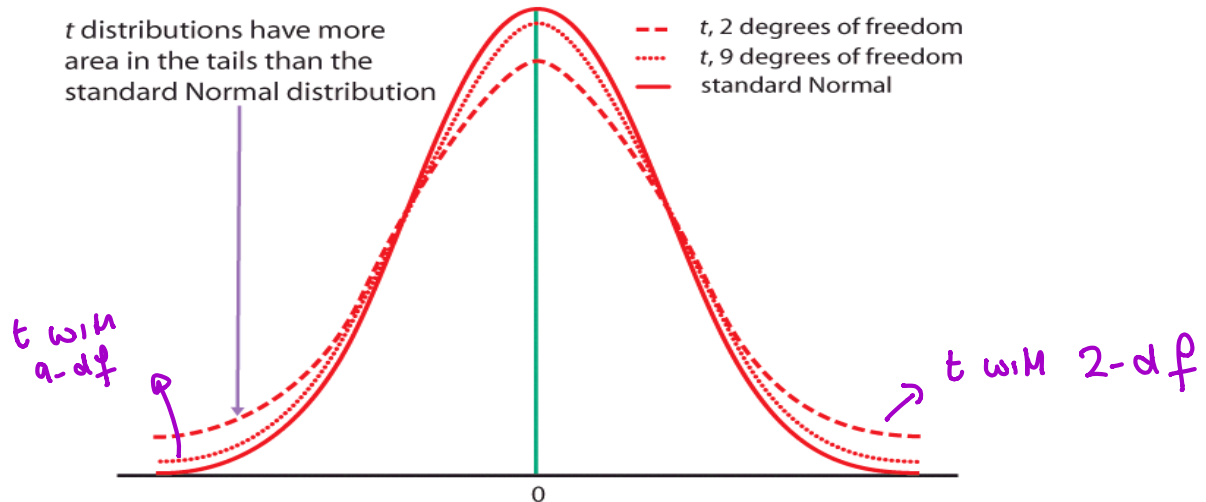
is a t distribution with $n - 1$ degrees of freedom.

- The degrees of freedom (df) is a measure of how well s estimates σ . *The larger the degrees of freedom, the better σ is estimated.*
- We use the t -tables to obtain these “critical” values.*

Distribution of the t-transform

The t distributions is wide (has thicker tailed) for smaller sample sizes, reflecting that s can be smaller than σ .

The thick tails ensure that the 80%, 95% confidence intervals are wider than those of a standard normal distribution (so are better for capturing the population mean).



Impact on confidence intervals

Suppose we want to construct the 95% confidence interval for the mean.

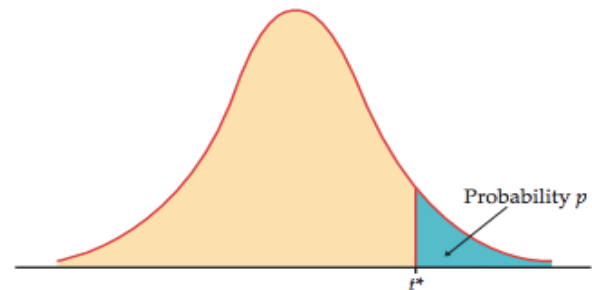
The standard deviation is unknown, so as well as estimating the mean we also estimate the standard

deviation from the sample. The 95% confidence interval is:

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

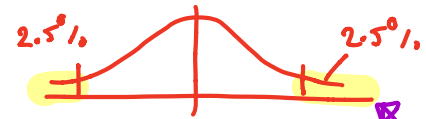
The blue area is proportion and for the 95% corresponds to 2.5%

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^* .



Tables

Examples: using the tables



Sample size $n=3$. ($df = 2$)

The 95% CI for the mean is

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

$$\text{MoE} = 4.3 \times \frac{s}{\sqrt{3}}$$

Sample size $n = 10$. ($df = 9$)

The 95% CI for the mean is

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

$$\text{MoE} = 2.26 \times \frac{s}{\sqrt{10}}$$

t distribution critical values

df	Upper-tail pr					
	.25	.20	.15	.10	.05	.025
1	1.000	1.376	1.963	3.078	6.314	12.71
2	0.816	1.061	1.386	1.886	2.920	4.303
	50%	60%	70%	80%	90%	95%

Confidence

t distribution critical values

df	Upper-tail <i>t</i>					
	.25	.20	.15	.10	.05	.025
1	1.000	1.376	1.963	3.078	6.314	12.71
2	0.816	1.061	1.386	1.886	2.920	4.303
3	0.765	0.978	1.250	1.638	2.353	3.182
4	0.741	0.941	1.190	1.533	2.132	2.776
5	0.727	0.920	1.156	1.476	2.015	2.571
6	0.718	0.906	1.134	1.440	1.943	2.447
7	0.711	0.896	1.119	1.415	1.895	2.365
8	0.706	0.889	1.108	1.397	1.860	2.306
9	0.703	0.883	1.100	1.383	1.833	2.262
	50%	60%	70%	80%	90%	95%

Confidence

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.

df	Upper-tail p					
	.25	.20	.15	.10	.05	.025 = 2.5%
1	1.000	1.376	1.963	3.078	6.314	12.71
2	0.816	1.061	1.386	1.886	2.920	4.303
3	0.765	0.978	1.250	1.638	2.353	3.182
4	0.741	0.941	1.190	1.533	2.132	2.776
5	0.727	0.920	1.156	1.476	2.015	2.571
6	0.718	0.906	1.134	1.440	1.943	2.447
7	0.711	0.896	1.119	1.415	1.895	2.365
8	0.706	0.889	1.108	1.397	1.860	2.306
9	0.703	0.883	1.100	1.383	1.833	2.262
10	0.700	0.879	1.093	1.372	1.812	2.228
11	0.697	0.876	1.088	1.363	1.796	2.201
12	0.695	0.873	1.083	1.356	1.782	2.179
13	0.694	0.870	1.079	1.350	1.771	2.160
14	0.692	0.868	1.076	1.345	1.761	2.145
15	0.691	0.866	1.074	1.341	1.753	2.131
16	0.690	0.865	1.071	1.337	1.746	2.120
17	0.689	0.863	1.069	1.333	1.740	2.110
18	0.688	0.862	1.067	1.330	1.734	2.101
19	0.688	0.861	1.066	1.328	1.729	2.093
20	0.687	0.860	1.064	1.325	1.725	2.086
21	0.686	0.859	1.063	1.323	1.721	2.080
22	0.686	0.858	1.061	1.321	1.717	2.074
23	0.685	0.858	1.060	1.319	1.714	2.069
24	0.685	0.857	1.059	1.318	1.711	2.064
25	0.684	0.856	1.058	1.316	1.708	2.060
26	0.684	0.856	1.058	1.315	1.706	2.056
27	0.684	0.855	1.057	1.314	1.703	2.052
28	0.683	0.855	1.056	1.313	1.701	2.048
29	0.683	0.854	1.055	1.311	1.699	2.045
30	0.683	0.854	1.055	1.310	1.697	2.042
40	0.681	0.851	1.050	1.303	1.684	2.021
50	0.679	0.849	1.047	1.299	1.676	2.009
60	0.679	0.848	1.045	1.296	1.671	2.000
80	0.678	0.846	1.043	1.292	1.664	1.990
100	0.677	0.845	1.042	1.290	1.660	1.984
1000	0.675	0.842	1.037	1.282	1.646	1.962
z*	0.674	0.841	1.036	1.282	1.645	1.960
	50%	60%	70%	80%	90%	95%

The confidence level is given at the bottom of the table.

Example 3: Normal data – sample size 3, t-distribution

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% confidence for the mean is

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

By replacing the normal distribution with the t-distribution we really do have 95% confidence that the interval contains the mean.

Data:

- Simulate
 - Distribution:
 - Mean:
 - Std. dev.:
- Using data table
 - Values in:
 - Where:

Initial confidence level:

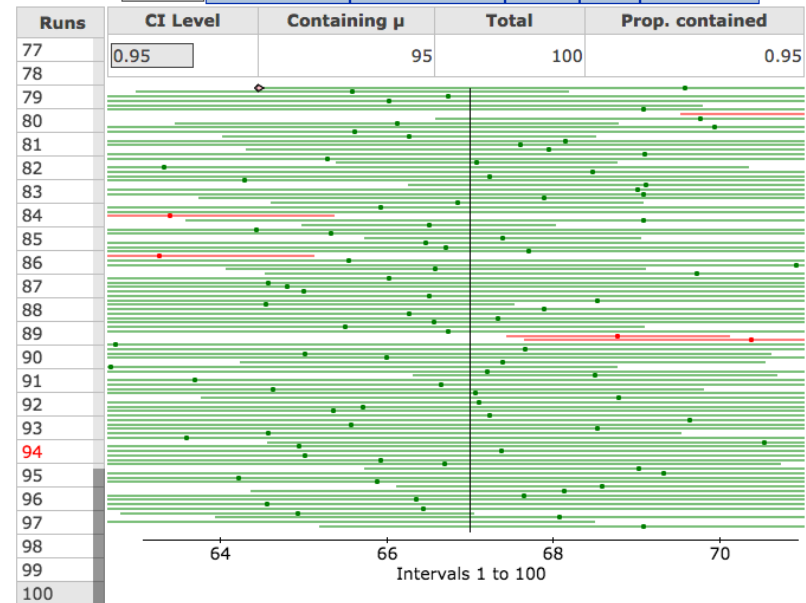
Initial sample size:

Interval:

- T
- Z (with pop. std. dev.)
- Z (with sample std. dev.)

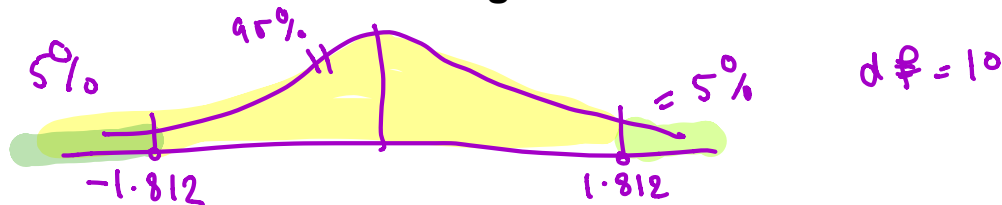
Confidence intervals for the mean using normal values with mean(μ)=67 and std.

Sample size:



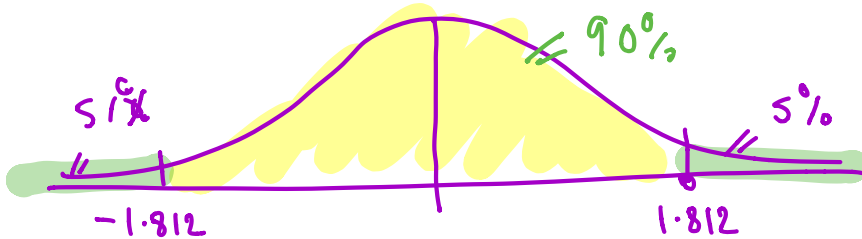
t-values

- t-values are like z-values (but are based on the t-distribution). 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. $(n=11)$
 - 1.812 means the area to the **right** of 1.812 is 5%.

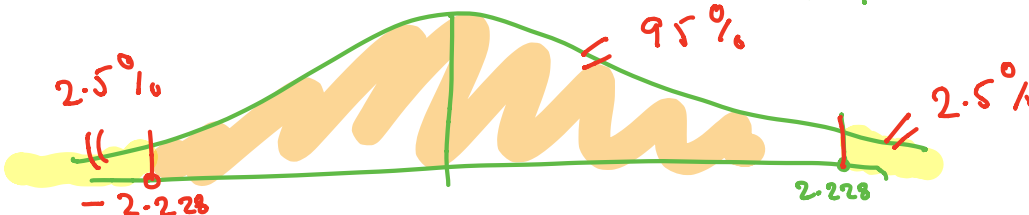


- By symmetry of the t-distribution, the area to the **left** of -1.812 is 5%.

- The area between -1.812 to 1.812 is 90% .



- The area to the right of 2.228 is 2.5% . ($df = 10$)

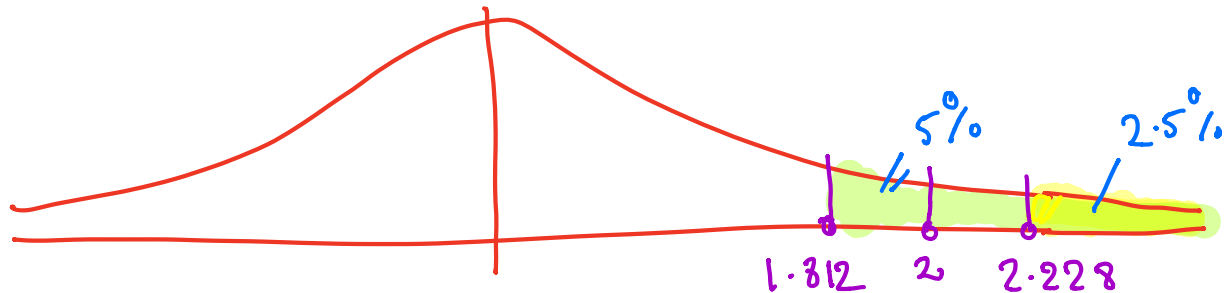


- By symmetry of the t-distribution, the area to the **left** of -2.228 is 2.5% .

- The area between -2.228 and 2.228 is 95%

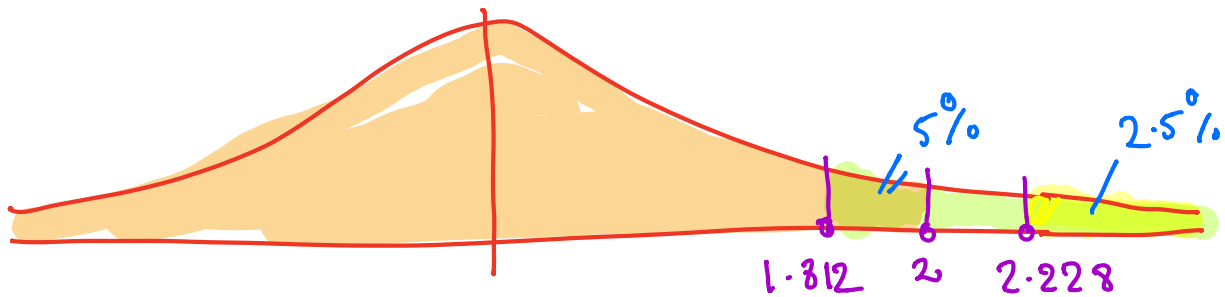
- The area below the distribution can also give probabilities.
- Again we focus on the t-distribution with 10df.
- The t-value is 2, what is the area to the **right** of 2?
 - Since 2 is between 1.812 (area to right is 5%) and 2.228 (area to the right is 2.5%). The area to the right of 2 is between 2.5-5%

df = 10



Because 2 is between 1.812 and 2.228, then the area to the right of 2 is between 2.5% and 5%.

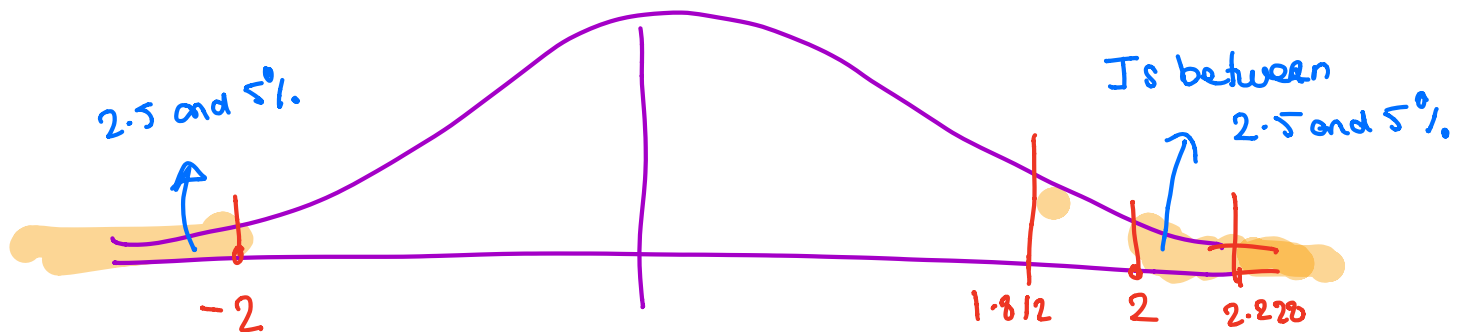
- The t-value is 2, what is the area to the **left of 2**?
- Since 2 is between 1.812 (area to right is 5%) and 2.228 (area to the right is 2.5%). The area to the left of 2 is between 95-97.5%



Because 2 is between 1.812 and 2.228.
The area to the left of 2 is greater than the area to the left of 1.812 but less than the area to the left of 2.228. So the area is between 95 to 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 right of 2.

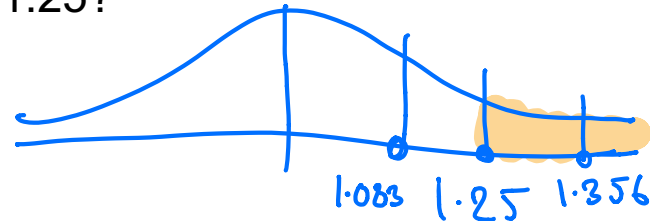
- Using the previous slide, the area to the left of -2 is between 2.5-5%



Question Time

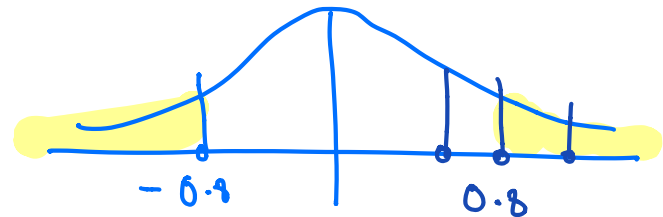
- The t-value for a t-distribution with 12 degrees of freedom is 1.25. What is the area to **the right** of 1.25?

- (A) less than 10%
- (B) between 10-15%
- (C) between 85-90%



- 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%



Question Time

$n=16$ We return to the example of prices of apartments in Dallas. 10 apartments are randomly sampled. The **sample mean and the sample standard deviation** $\bar{x} = 980$ based on this sample is 980 dollars and 250 dollars $s = 250$ (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 $[980 \pm 2.262 \times 79] = [801, 1159]$.

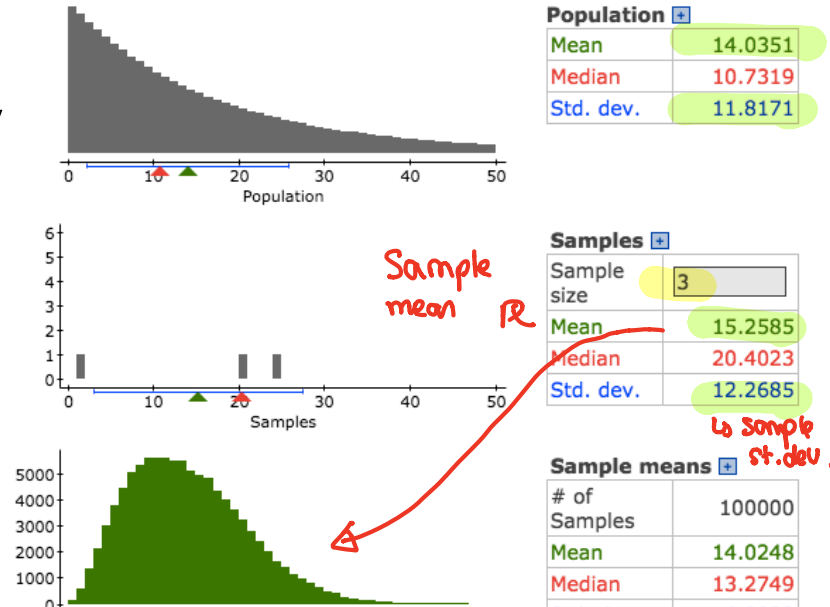
□ (B) The 95% confidence interval for the mean price is $[980 \pm 2.262 \times 250]$.

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

$$\begin{aligned} & \left[\bar{x} \pm t_{q(0.025)} \times \frac{s}{\sqrt{10}} \right] \\ & = \left[980 \pm 2.262 \times \frac{250}{\sqrt{10}} \right] \end{aligned}$$

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.



True standard error of $\bar{x} = \frac{11.8}{\sqrt{3}} \approx 6.9$

Estimated standard error = $\frac{12.26}{\sqrt{3}} = 7.06$

4.3

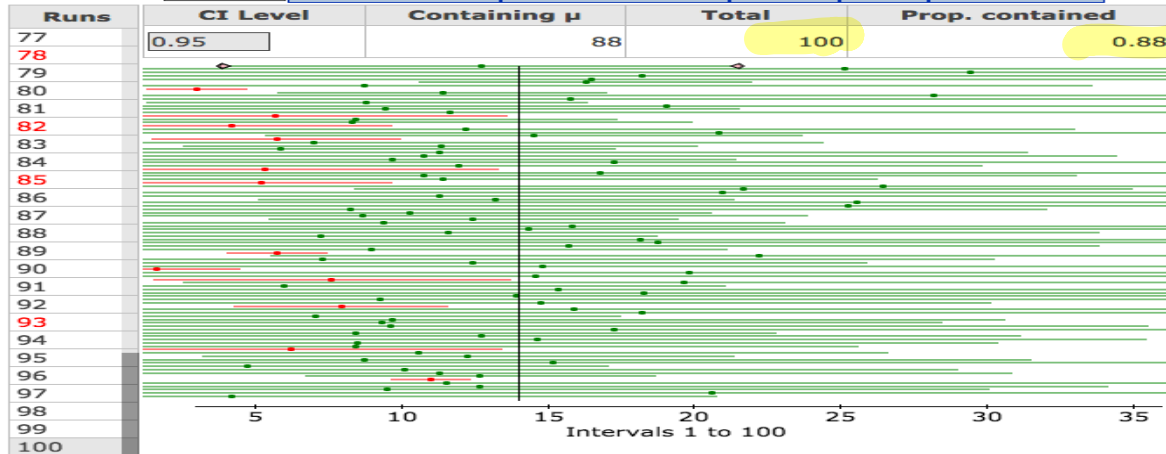
$$\left[\bar{x} \pm t_{2}^{(95\%)} \times \frac{s}{\sqrt{3}} \right]$$

Example: We draw a sample of size three from the skewed distribution on the previous page . For each sample the 95% CI for the mean is evaluated using the t-distribution (each confidence interval is plotted below). 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.**

Confidence intervals for the mean using right skewed values with mean(μ)=14 and

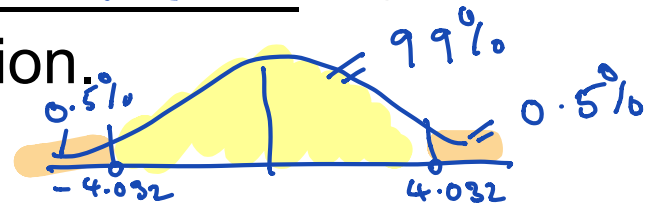
Sample size:

[100 intervals](#) [1000 intervals](#) [Reset](#) [Info](#) [Sort graph](#)



Complete the blanks

- ❑ We use the t-distribution instead of the normal distribution when we estimate the standard deviation from the data
- ❑ Using a t-distribution instead of a normal distribution leads to a wider confidence interval.
- ❑ The t-distribution does not correct for the lack of normality of the original data.
- ❑ As the sample size grows the estimated standard deviation tends to get closer to the population standard deviation.



Question Time

- 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).

$$\bar{x} = 25.25$$

$$s = 20.2$$

$$n = 6$$

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

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

$$df = 5$$

$$99\% \rightarrow 0.5\%$$

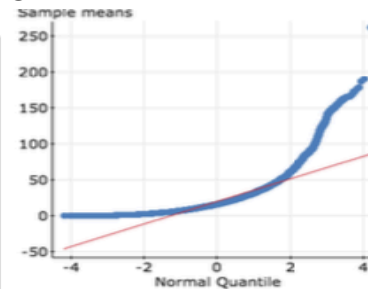
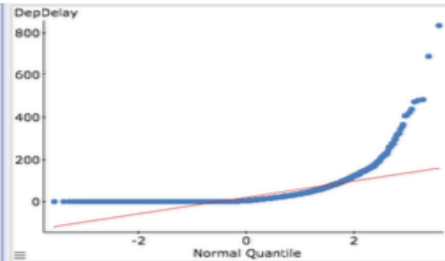
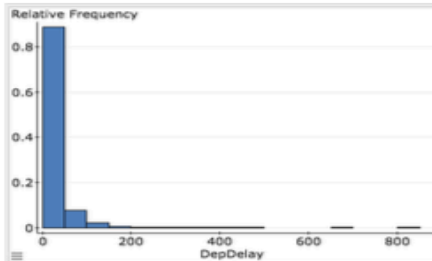
↓

$$4.032$$

$$= \left[25.25 \pm 4.032 \times \frac{20.2}{\sqrt{6}} \right]$$
$$= [0, 25.2 + 4.032 \times 8.24]$$

Question Time

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



QQplot
of sample
mean ($n=6$)

QQplot of delay.

Figure 1: H

histogram, QQplot, QQplot of the sample mean

(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.

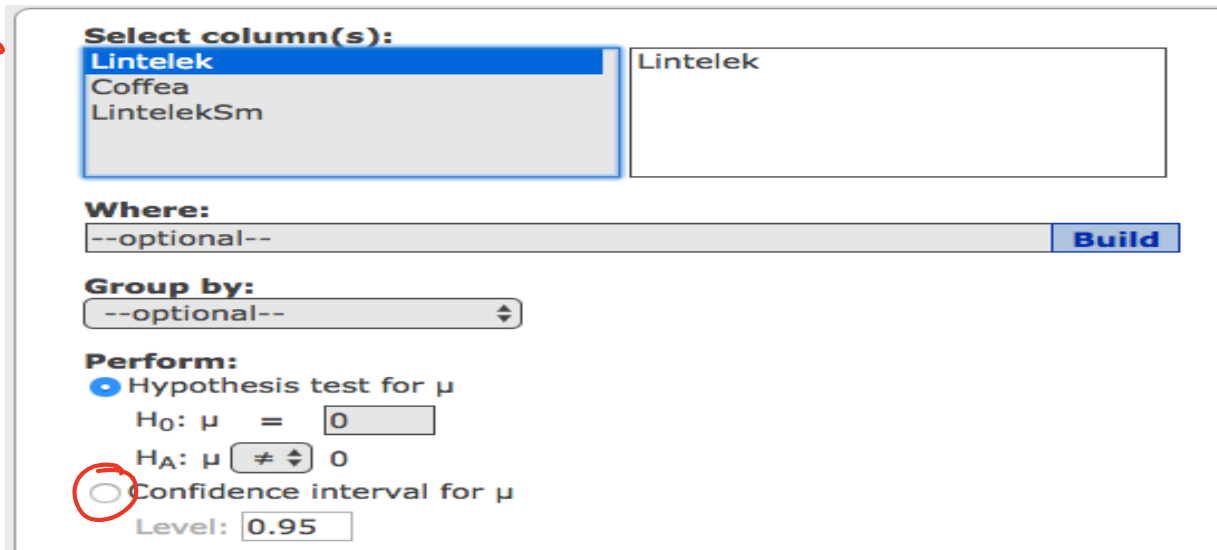
(C) The sample mean is normally distributed and we have 99% confidence in the interval.

(D) [A] and [B]

(E) [A] and [C]

Statcrunch: Confidence intervals

- ❑ Load data into Statcrunch.
- ❑ T Stats -> One Sample -> With Data
- ❑ Highlight column.
- ❑ Check the Confidence interval box and state level (the default is 0.95 = 95%)



The screenshot shows the Statcrunch interface for a one-sample t-test. The 'Select column(s):' section has 'Linteleg' selected. The 'Where:' section is set to '--optional--'. The 'Group by:' section is also set to '--optional--'. In the 'Perform:' section, the 'Hypothesis test for μ ' option is selected, with $H_0: \mu = 0$ and $H_A: \mu \neq 0$. The 'Confidence interval for μ ' option is selected, and the 'Level:' is set to 0.95. A red circle highlights the 'Confidence interval for μ ' radio button, and a red arrow points from the text in the list above to this radio button.

Select column(s):
Linteleg
Coffea
LintelegSm

Where:
--optional-- **Build**

Group by:
--optional--

Perform:
 Hypothesis test for μ
 $H_0: \mu = 0$
 $H_A: \mu \neq 0$
 Confidence interval for μ
Level: 0.95

Example 1: Amazon product scores



LINTELEK
Fitness Tracker Watch, Lintelek Smart Band Step Tracker
Calorie Counter Sleep Monitor Touch Screen Activity Health
Tracker Wearable Pedometer Smart Bracelet for iPhone
Android Smartphone
★★★★☆ 174 customer reviews | 72 answered questions

4.2 out of 5 stars

Shipping Details

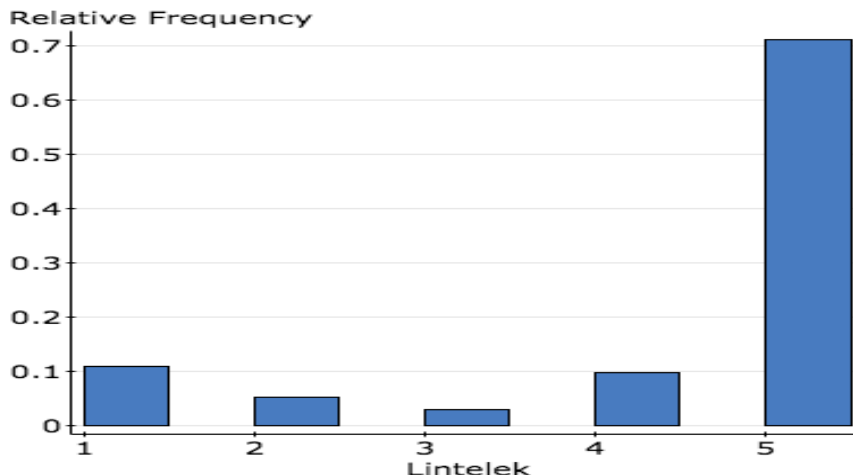
Order within 13 hrs 22 mins and choose One-Day Shipping at
ed by Amazon. Gift-wrap available.

See all verified purchase reviews >

• POWERFUL FUNCTION: Pedometer, Distance, Calories Burned Measuring, Sleep Monitor, SMS

Row	Lintelek
1	5
2	5
3	5
4	5
5	4
6	4
7	5

What the data looks like



Histogram of data, which is also given by Amazon.

Jon tends to only buy products which get over 4 stars. Can Jon be sure that this tracker would get an mean rating of over 4 stars if all customers bought it?

- Become familiar with reading Statcrunch output (it will be tested).

- Understanding the Summary Statistics

$$t_{173} (2.5)$$

Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.
Lintelek	174	4.2528736	1.8778819	1.3703583	0.10388659

$$[4.25 \pm 1.97 \times 0.103]$$

$$s.e = \frac{1.37}{\sqrt{174}} = 0.103$$

- The sample mean is 4.2
- The sample standard deviation is 1.37.
- The standard error is $1.37/\sqrt{174} = 0.104$

$$\begin{matrix} 1.979 \\ \approx 1.96 \end{matrix}$$

$$\begin{matrix} t\text{-}d.f \\ d.f = 173 \end{matrix}$$



- The Statcrunch output for the confidence interval is :

One sample T confidence interval:

μ : Mean of variable

95% confidence interval results:

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Lintelek	4.2528736	0.10388659	173	4.0478252	4.4579219

\bar{x}

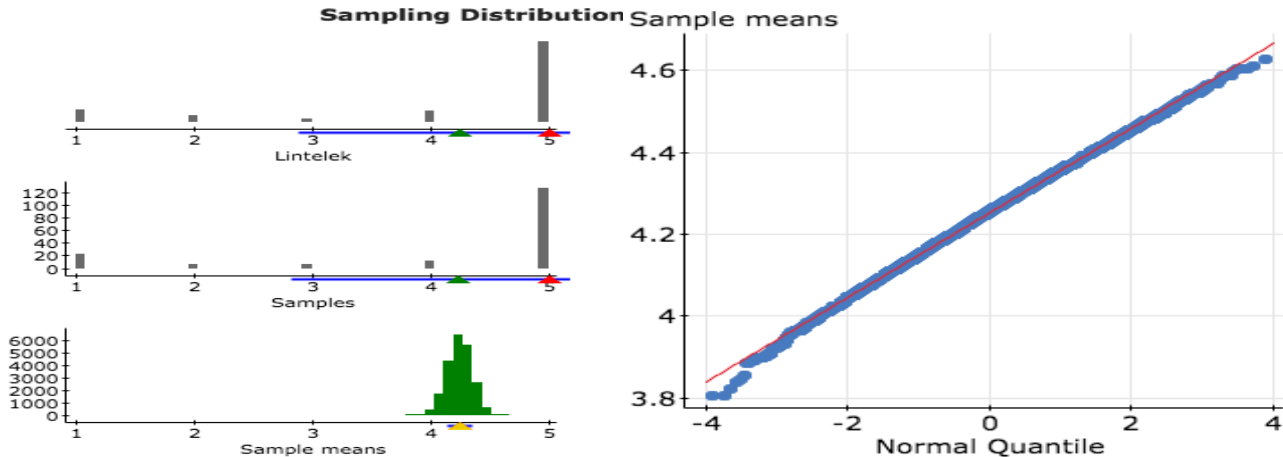
$\frac{s}{\sqrt{174}}$

$n-1$
 $n=174$

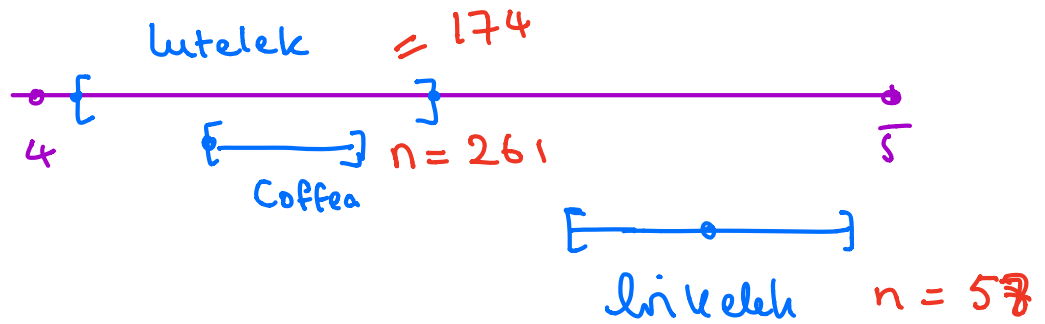
$[4.05, 4.46]$

- 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 is likely to have more than 4 stars.
- But we still need to *check normality of the sample mean*.
- To do this: Use the Statcrunch Sampling App.

- ❑ The data is not normal, so we need to check if we have 95% confidence in this interval. The observed sample mean of 4.2 is a number from the green histogram



- ❑ **Observation:** the QQplot of the green histogram, it is close to normal.
- ❑ **Conclusion:** we really do have pretty much close to 95% confidence in the interval.

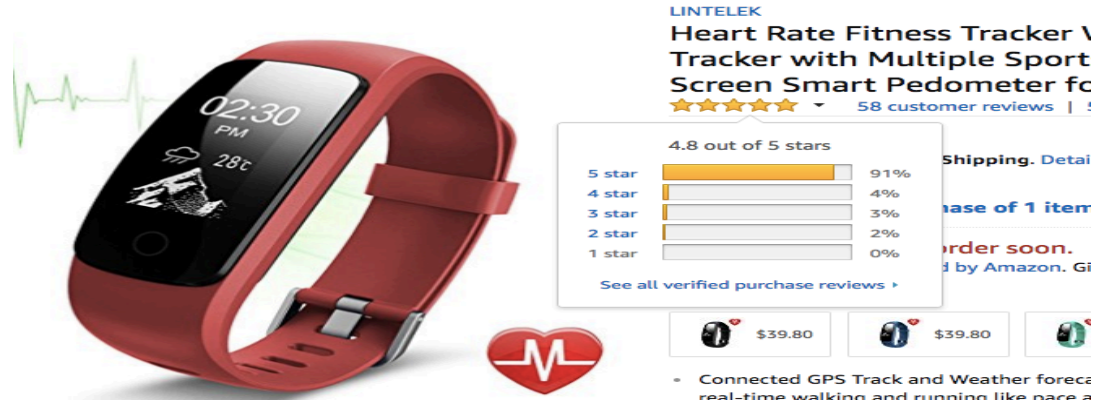


(*) Lutelelele Sm CI interval is above the other two CIs. Suggests Lutelelele Sm scores higher

(*) We still need to be cautious because sample size is small which means the sample mean is not close to normal. This means we may not have full 95% CI in the interval.

Amazon example 2

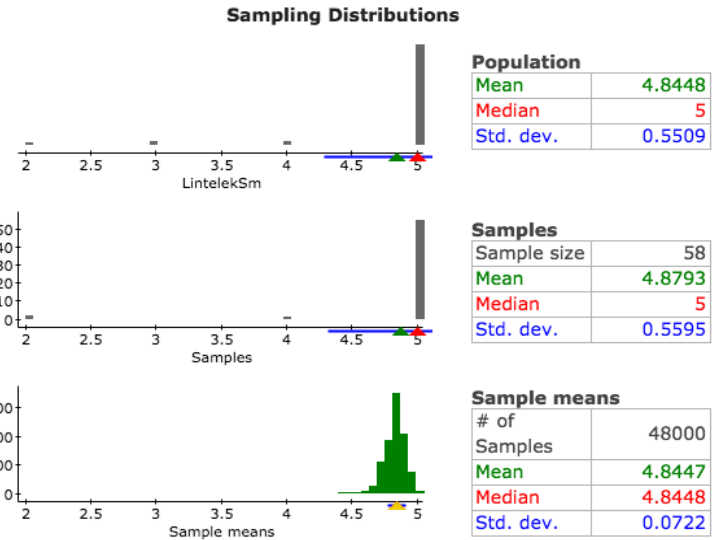
- 4.8 is the sample mean based on 58 reviews.



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

In the statistical analysis we assume this histogram is **normal**. But *it appears a little left skewed*.

Distribution for sample of size 58 =



- The 95% confidence interval for the mean is [4.70,4.99]

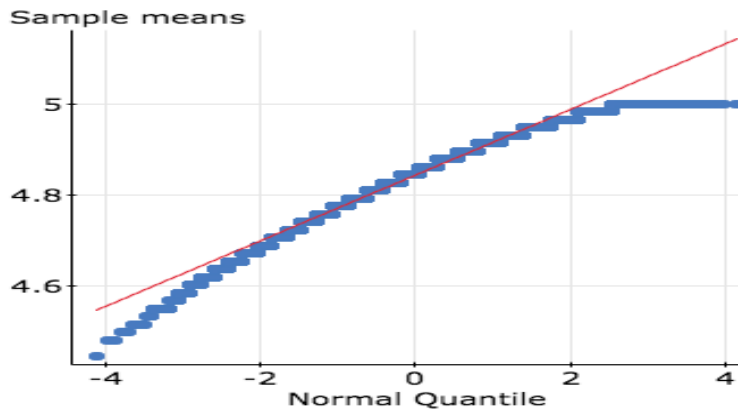
One sample T confidence interval:

μ : Mean of variable

95% confidence interval results:

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
LintelekSm	4.8448276	0.072970521	57	4.6987066	4.9909485

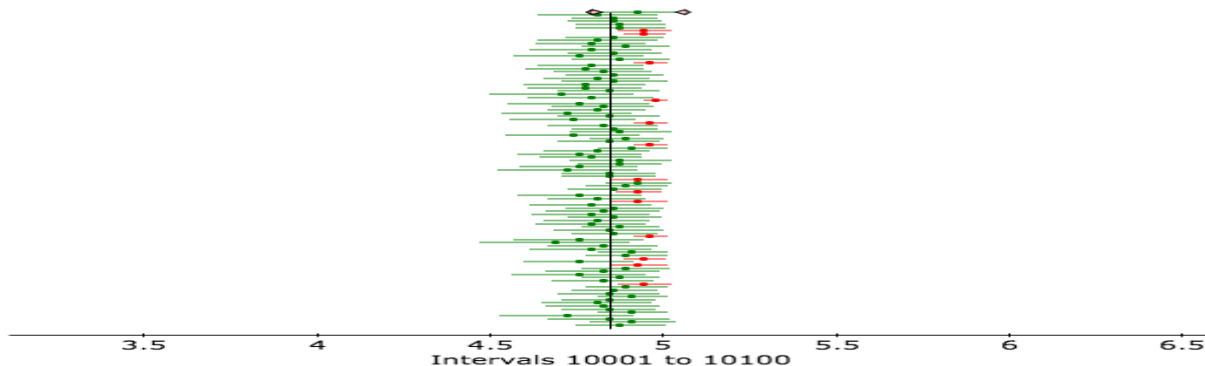
- **Warning:** The QQplot of the sample mean (the green histogram on the previous page) is not normal.



- ❑ The output gave the 95% confidence interval for the mean to be [4.70,4.99]
- ❑ However, we observe using the Statcrunch app that we only have about 89% confidence in the interval.

Confidence intervals a mean: LintelekSm ($\mu=4.845$, $\sigma=0.556$) Type=T
Sample size=58

CI Level	Containing μ	Total	Proportion
0.95	9010	10100	0.8921



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

- ❑ **Conclusion** We are **less than** 95% confident that the mean is contained in $[4.70, 4.99]$.
- ❑ **Reality check:** What do we want to learn from this interval.
- ❑ The entire interval lies far above 4.0, that a population mean of 4.0 or less seems implausible.
- ❑ Help: Draw number line with 4.0, 4.8 and the standard error of 0.07 to see this.

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:

Column	n	Mean	Variance	Std. dev.	Std. err.
Polyphenol(9)	9	5.5	6.335	2.5169426	0.83898086

$$\bar{x} = 5.5$$

$$s = 2.517$$

$$\frac{2.517}{\sqrt{9}} = 0.83$$

$$\left[5.5 \pm \frac{2.3}{0} \times 0.83 \right]$$

- 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	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
(...)												
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

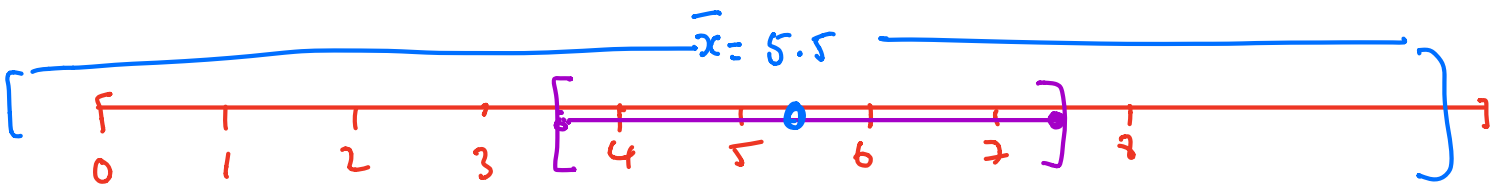
- The margin of error m is: $m = t^* \times s/\sqrt{n} = 2.306 \times 2.517/\sqrt{9} \approx 1.93$. The 95% confidence interval is 5.50 ± 1.93 , or 3.57 to 7.43.
- We can say **“With 95% confidence, the mean of percent increase is between 3.57% and 7.43%.”** The corresponding Statcrinch output is below:

One sample T confidence interval:

μ : Mean of variable

95% confidence interval results:

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Polyphenol(9)	5.5	0.83898086	8	3.5653067	7.4346933



↑
we think the mean over
all healthy males lies
here

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, 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 $df = 14 = (15 - 1)$

Since the sample size has increased, it is likely that the sample standard deviation is closer to the true standard deviation.

Fact: As the sample size grows both the sample mean and the sample standard deviation tend to get closer to the population standard deviation.

$n = 9$ previous 2.3
 \downarrow $n = 15 = 2.14$ decreases

- This is why the critical value for the t- distribution changes from 2.306 when the df is 8 (sample size is n=9).....
-to 2.145 when the df = 14 (sample size is n=15).
- With 95% confidence we believe the true mean change in polyphenol level lies in the interval

$$\left[4.3 \pm \underbrace{2.145}_{\text{t-tables 14 df, 2.5\%}} \times \frac{3.06}{\sqrt{15}} \right] = [2.6, 6]$$

$$\bar{x} \pm 2.145 \times \frac{3.06}{\sqrt{5}}$$

Reminder: Confidence intervals

- ❑ Learn to construct confidence intervals in Statcrunch
- ❑ Stat -> T Stats -> One Sample -> With Data
- ❑ You will get the following drop down menu.

The screenshot shows the 'One Sample T' window in StatCrunch. The 'Select column(s):' section has 'Polyphenol(15)' selected. The 'Where:' section is empty. The 'Group by:' section is empty. The 'Perform:' section has 'Confidence interval for μ ' selected with a level of 0.95. The 'Options' dialog box is open, showing the following results:

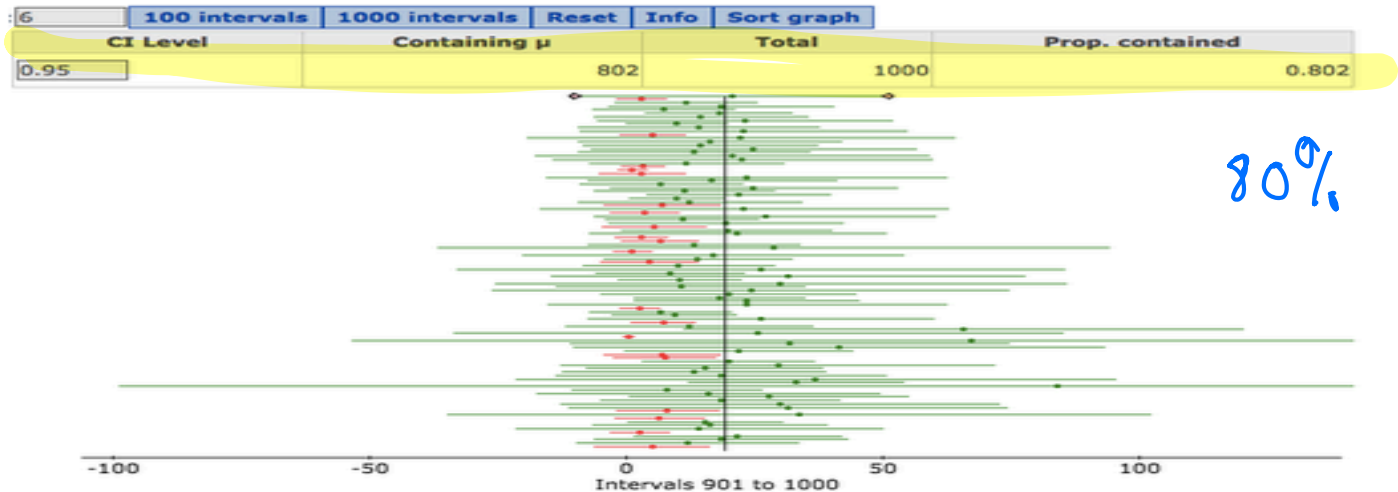
95% confidence interval results:
 μ : Mean of variable

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Polyphenol(15)	4.3	0.79162281	14	2.6021379	5.9978621

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.

Question Time

- 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].

Accompanying problems associated with this Chapter

- HW 8
- HW 9
- HW 10