Introductory Statistics with Excel

Andrew Parnell

University College Dublin

Class 4 - Confidence intervals and t-tests

Learning outcomes

In this class we will cover

- The t-distribution
- Why the tests we have met so far aren't quite right
- Running t-tests
- Creating confidence intervals

Revision from yesterday

What do we know so far?

- We have a scientific question of interest we want to answer.
- The information that might answer this question comes in the form of data obtained from an RCT, a scientific experiment, a survey, etc
- We know that the data will suffer from **uncertainty** and estimates taken from it will have a **sampling distribution**. We will not be able to estimate the true answer correctly

We are now in a position to use our tools of inference to create an answer to the question of interest by taking account of the uncertainty in our data. We will aim to make decisions about population parameters from sample data using hypothesis tests and confidence intervals.

Revision from yesterday

Yesterday we showed that:

- If we take lots of samples the sample means turn out to be **normally distributed**
- The larger the samples we take the **closer** the sample means will be to the true population mean

We know that the sampling distributions will be normally distributed because of the **central limit theorem**.

Reminder: population parameters and sample statistics

- The **population** is the entire collection of units about which we are interested
- The **population parameter** is a fixed number associated with population. We want to estimate it
- The **sample** is the collection of units about which we have data. We have a **sample size** of n units in our sample
- The **sample statistic** is a summary number computed from the population corresponding to the population parameter. It is occasionally known as a **sample estimate** or **point estimate**

The fundamental rule for using data for inference is that the available data must be **representative of the population with regards to the question of interest**

Example

We have milk production values for 14 cows in kilograms. We are interested in the mean milk production value of all cows in Ireland.

How do the following terms relate to this problem?

- population
- sample
- sample size
- population parameter
- sample statistic

Standardising statistics for sampling distributions

- Yesterday we learnt about how to **standardise** values by subtracting the mean and dividing by the standard deviation. We end up with a random variable that has mean 0 and standard deviation 1.
- We can do this in general for **any** sample statistic:

$$\text{Standardised value} = \frac{\text{Sample statistic} - \text{Population parameter}}{s.d.(\text{sample statistic})}$$

- In the real world, we don't have a population parameter, nor a standard deviation. However, we can replace the standard deviation of the parameter with the **standard error** and the population parameter with a suitable **test value**.

The t distribution

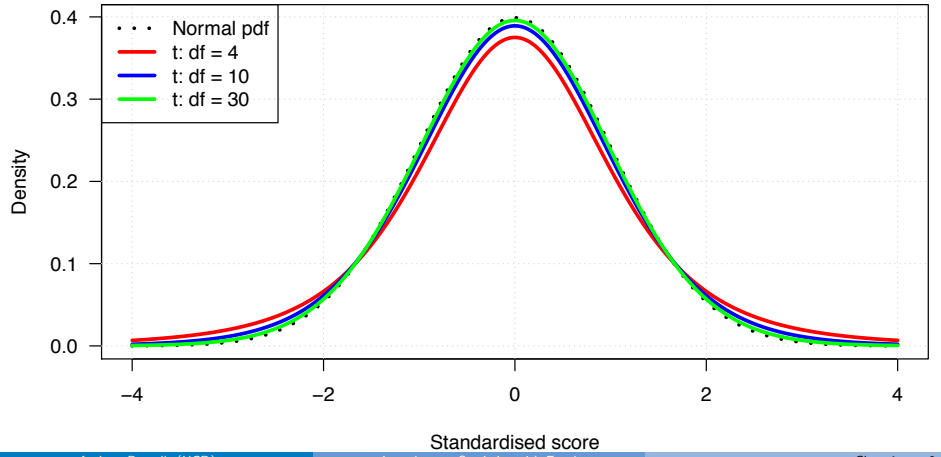
Using Student's t-distribution instead

- If we replace the population parameter and standard deviation values, we have:

$$\text{Standardised value} = \frac{\text{Sample statistic} - \text{Hypothesised population value}}{\text{sample standard error}}$$

- If we knew the true standard error of the sample statistic, then the standardised value would be normally distributed. However, when we use the standard error we end up with something that is t-distributed
- The t-distribution (or Student's t-distribution) is another bell-shaped distribution. It has an additional parameter to the Normal distribution called the **degrees of freedom** (df). For most of the examples we meet the degrees of freedom parameter is a function of the sample size
- When df$= \infty$ then the pdf of t is equivalent to the normal distribution. When df is small, the bell shape has much longer tails

Student's t-distribution: pictures



The law of large numbers and the central limit theorem

There are two key statistical results that are relevant here:

1. The **law of large numbers** states that **the sample mean will 'eventually' get 'close' to the population mean as the sample size rises**. This happens no matter how you define 'close' though 'eventually' could be a very long time
2. The **central limit theorem** (CLT) is possibly the most important theorem in statistics. It states that **if the sample size is sufficiently large, the sample means of random samples from a population with some mean and standard deviation are approximately normally distributed with the same mean and standard deviation equal to the original standard deviation divided by the square root of the sample size**

All of the results of the different situations we have met in this lecture follow from the CLT

t-tests

Reminder: the 8-steps of a Hypothesis Test

1. State the Null and Alternative Hypotheses (H_0 and H_A).
2. Identify the level of significance α.
3. State any assumptions.
4. Assume H_0 is true and compute the test statistic.
5. Identify the rejection region:
   decide if the test is upper, lower or two-tailed, then use the appropriate table and α to find critical value(s).
6. Compare the test statistic with the critical value(s).

7. Draw a conclusion:
   - If the value of the test statistic falls inside the rejection region then reject H_0 and conclude that H_A is true.
     This is either a correct decision or a type I error.
   - If the test statistic does not fall in the rejection region do not reject H_0. This does not mean that H_0 is proven, it simply means that the data do not provide sufficient evidence to reject it.
     This is either a correct decision or a type II error.
8. Interpret the conclusion: The conclusion must be stated in the context of the problem and should include the level of significance.

Example

A random sample of 14 cows was selected from a large dairy herd in Cork. The milk yield in one week was recorded, in kg, for each cow.

169.6	142	103.3	111.6	123.4	143.5	155.1
101.7	170.7	113.2	130.9	146.1	169.3	155.5

- Dept. of Ag. want to investigate the farmer's claim that the mean weekly milk yield for the herd is 120kg.

Milk yield example

1. *State the Null and Alternative Hypotheses (H_0 and H_A).*

   H_0 : population mean $= 120$ and H_A : population mean $\neq 120$

2. *Identify the level of significance.*

   We will use $0.05 = 5\%$

3. *State any assumptions.*

   Small sample so use t distribution.

Milk yield example

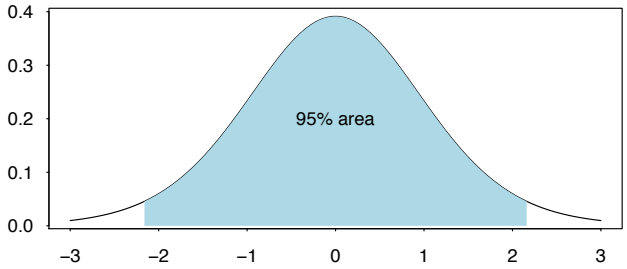
4 *Assume H_0 is true and compute the test statistic.*

$$\begin{aligned}\text{test statistic} &= \frac{\text{sample mean} - 120}{\text{standard deviation}/\sqrt{n}} \\ &= \frac{138.278 - 120}{24.58/\sqrt{14}} \\ &= 2.78\end{aligned}$$

Milk yield example

5 *Identify the rejection region: use the appropriate Excel function and significance level to find critical value(s).*

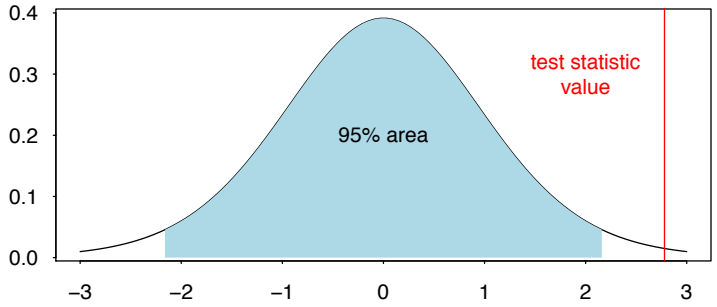
2 tailed test. $df = n - 1 = 13$.



critical value $= 2.16$ (Excel formula was `T.INV.2T(0.05, 13)`)

Milk yield example

6 *Compare the test statistic with the critical value(s).*



test statistic $= 2.78 > 2.16 =$ critical value

Milk yield example

7 *Draw a conclusion:*

Test statistic falls inside the rejection region so reject H_0 and conclude that H_A is true.

This is either a correct decision or a type I error.

8 *Interpret the conclusion:*

The mean weekly milk yield is *significantly different* from 120kgs, at the 5% level.

p-value approach to hypothesis testing

- The conclusion of a hypothesis test is dependent on the choice of the significance level (usually 0.05).
- Rather than pick a value, you could find the value associated with the test statistic value. This is the p-value approach we met yesterday.
- The p-value can be compared to the significance level, so each person can make a decision based on their own choice of value.

p-value approach to hypothesis testing

- Our test statistic value was 2.78. We can the probability of being more extreme than this value using Excel
- The formula is: `T.DIST.2T(2.78, 13)` $= 0.016$.
- This is smaller than 0.05 so if using 5% as the significance level we can say that the result is 'statistically significant at the 1.6% level'

p-values continued

- Don't forget what statistical significance means. It is not a measure of whether a result is large or important
- The American Statistical Association recently released some advice on p-values:

  By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis[1]

[1] http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108?needAccess=true

Difference between two population means

Many studies are undertaken with the objective of comparing the characteristics of two populations. In such cases we need two samples, one from each population.

Examples

- Compare the average fuel consumption of Volvo and Renault cars.
- Compare the proportion of the population < 25 yrs in the UK and Ireland.
- Compare the average IQ for males and females.
- Compare the proportion of people 'improved' following treatment for hayfever with drug A and drug B.

Difference between two population means

Population 1	Population 2
Pop mean 1, Pop sd 1	Pop mean 2, Pop sd 2
$\Downarrow$	$\Downarrow$
Sample 1	Sample 2
Sample mean 1, Sample sd 1	Sample mean 2, Sample sd 2

We are not interested in the population means themselves but in their difference

For example

- pop mean 1 − pop mean 2 $= 0$ implies pop mean 1 $=$ pop mean 2
- pop mean 1 − pop mean 2 > 0 implies pop mean 1 $>$ pop mean 2
- pop mean 1 − pop mean 2 < 0 implies pop mean 1 $<$ pop mean 2

The two samples will be **independent** or dependent (**paired**) according to the design of the experiment/study

Independent and Paired samples

Independent: if selection of items for one sample does not depend in any way on the selection for the other sample.

Paired: items in the two samples are paired in some way.

Example – Study to compare average IQ of males and females.

- 2 independent samples – select random sample of males and then a random sample of females.
- 2 paired samples – select a number of brother & sister pairs and these make up the two samples.

We only consider inferences based on independent samples

Sampling Distribution of a difference in means

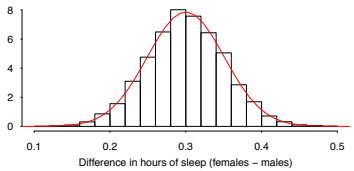
Since we are interested in the population mean difference it is natural to consider the sample means differences. We need to know its sampling distribution.

The **sampling distribution** *for a statistic is the distribution of possible values of the statistic, for repeated samples of the same size taken from the same population.*

- A sample 190 UCD students were asked "How many hours of sleep did you get last night?"
- The mean of the 100 males was 7.1 hours and the 90 females was 7.4 hours
- Suppose another 190 students were asked that same question.
- It is unlikely that the mean difference would be exactly 0.3 hours again.

Sampling Distribution of a difference in means

- Is we asked 'lots' of groups, each with 190 students (100 males and 90 females), the same question, we would get a histogram of sample mean differences:



- A normal curve is superimposed on top of the histogram – this is the **sampling distribution** of possible sample means.
- Note that it is approximately a normal distribution.

Performing the hypothesis test

- To test H_0 : population mean difference $= 0$, the test statistic is:

$$\text{test statistic} = \frac{\text{mean group 1} - \text{mean group 2}}{\sqrt{\left(\frac{\text{sd1}^2}{n_1} + \frac{\text{sd2}^2}{n_2}\right)}}$$

  where these are all calculate on the sample data
- The critical value then comes from the t-distribution with a complicated formula for the degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

More advanced version.

- To test the more complicated hypothesis:

$$H_0 : \text{population mean difference} = k$$

- The test statistic is then:

$$\text{test statistic} = \frac{\text{mean group 1} - \text{mean group 2} - k}{\sqrt{\left(\frac{\text{sd1}^2}{n_1} + \frac{\text{sd2}^2}{n_2}\right)}}$$

- ... but everything else is the same

Example - slow learners

10 pupils previously diagnosed as slow learners were taught by a new method, 12 taught by a standard method. After 6 months they are scored on a reading test with the following results:

	n	$mean$	sd
New method	10	76.4	5.84
Standard method	12	72.3	6.34

At the 5% level, do these data indicate that the new method is better?

Example - slow learners

- H_0 : population mean difference $= 0$ vs H_0 : population mean difference $\neq 0$
- Test statistic:

$$\text{test statistic} = \frac{(76.4 - 72.3)}{\sqrt{\frac{5.84^2}{10} + \frac{6.34^2}{12}}} = 1.58$$

- Compute the degrees of freedom:

$$df = \frac{\left(\frac{5.84^2}{10} + \frac{6.34^2}{12}\right)^2}{\frac{1}{10-1}\left(\frac{5.84^2}{10}\right)^2 + \frac{1}{12-1}\left(\frac{6.34^2}{12}\right)^2} = 19.77$$

- Compare with critical value 2.09 (`T.INV.2T(0.05, 20)` in Excel)
- Conclusion - do not reject H_0 at 5% level

Confidence Intervals

Confidence intervals and confidence levels

- A confidence interval is a set of two values (lower and upper) accompanied by a **confidence level**.
- The confidence level tells us how likely it is that the interval estimate contains the true value of the parameter
- Most commonly we calculate 90%, 95% and 99% confidence intervals
- As we are often assuming the sampling distribution to be normal (or t), we cannot calculate a 100% confidence interval (remember that the bell-shaped normal distribution has infinite tails from our empirical rule)

Interpreting confidence intervals

BEWARE: this is something lots of people get wrong

- Our confidence level quantifies how likely it is that the interval estimate contains the true value of the parameter
- It does not tell us how likely the parameter is to be in the interval
- This is because our parameters are **fixed** and the uncertainty is in the **sample data**
- A more complete definition is that a 95% confidence interval would include the true population value for 95% of all possible random samples from the population

How to calculate confidence intervals

A confidence interval or interval estimate (in all of the situations we meet) can be calculated as:

$$\text{Sample estimate} \pm \text{Multiplier} \times \text{Standard error}$$

The multiplier is based on the confidence level desired and the probability distribution

- Here the **sample estimate** is the statistic we can calculate from the sample data
- The **multiplier** determines the amount of confidence we will have in the result
- The **standard error** is the standard error of the statistic we are trying to calculate

A CI for a single population mean

The confidence interval for a single population mean is:

$$\text{sample mean} \pm \text{t-value} \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

Some notes:

- $\frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$ is our estimate of the standard error of the mean
- The t-value is the multiplier for the t-distribution with (sample size $-$ 1) degrees of freedom at the chosen confidence level
- We use the t distribution here because the normal distribution is invalid when the sample size is small. When the sample size is large the t and the normal are the same

Example

A random sample of 14 cows was selected from a large dairy herd in Cork. The milk yield in one week was recorded, in kg, for each cow.

169.6	142	103.3	111.6	123.4	143.5	155.1
101.7	170.7	113.2	130.9	146.1	169.3	155.5

1. Dept. of Ag. want to investigate the farmer's claim that the mean weekly milk yield for the herd is 120kg.

Milk yield example

Small Sample CI:

$$\text{sample mean} \pm (t \text{ value}) \frac{\text{standard deviation}}{\sqrt{n}}$$

$$138.28 \quad \pm \quad (2.16) \frac{24.58}{\sqrt{14}}$$

$$(124.09, 152.47)$$

Based on our sample, we are 95% confident that the mean weekly milk yield lies between 124kg and 152kg.

Where did the t-value come from?

- In the previous slide we used 2.16 as the t-value in the confidence interval.
- This value is the 97.5th percentile of the t-distribution with 13 degrees of freedom
- For a 99% confidence interval we would use the 99.5th percentile of the t-distribution. (It always helps to draw a picture)
- In Excel, we'd use `T.INV(0.975, 13)` which will return the value 2.16

What determines the width of a CI?

Three things determine the width of a confidence interval:

1. The sample size n. If n is large we will have a smaller interval because the standard error will be smaller
2. The confidence level. The more confident we want to be the wider we have to make the interval
3. The natural variability of the data. If the data are very variable then our confidence interval will be wider

Note that we have some control over (1) and (2) but no control over (3)

CI for the difference between two independent population means

Suppose have two groups and want to calculate a confidence interval for the differences between them. We use the formula:

$$\text{mean group 1} - \text{mean group 2} \pm t\text{-value}\sqrt{\left(\frac{\text{sd1}^2}{n_1} + \frac{\text{sd2}^2}{n_2}\right)}$$

where sd1 and sd2 are the standard deviations of each of the groups, as n_1 and n_2 are their sample sizes.

The extra complication in this version is that the t-value is harder to calculate

Example

Some researchers were interested in the effect of hangovers amongst college students. Students were asked whether their parents suffered from alcohol problems and asked to rate the severity and duration of their own hangovers on a 13-point scale, with 13 being the most severe. 1227 students were contacted and the data are shown below. Calculate a 95% confidence interval.

Group	Sample size	Mean	Standard deviation
Parental alcohol problems	$n_1 = 282$	mean grp 1 $= 5.9$	sd1 $= 3.6$
No parental alcohol problems	$n_2 = 945$	mean grp 2 $= 4.9$	sd2 $= 3.4$

Solution: using unequal variances

This is the formula for the degrees of freedom (yikes!):

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{3.6^2}{282} + \frac{3.4^2}{945}\right)^2}{\frac{1}{282-1}\left(\frac{3.6^2}{282}\right)^2 + \frac{1}{945-1}\left(\frac{3.4^2}{945}\right)^2} \approx 441$$

We also need

$$\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} = \sqrt{\left(\frac{3.6^2}{282} + \frac{3.4^2}{945}\right)} = 0.24$$

Finally we need `T.INV(0.975, 441)`$= 1.96$ so the 95% CI is:

$$5.9 - 4.9 \pm 1.96 \times 0.24 = (0.53, 1.47) \text{ symptoms}$$

Summary of class 4

- The t-distribution is the one to use whenever we don't know what the population standard deviation is (i.e. nearly always)
- t-tests are created following the standard hypothesis testing steps. We calculate whether our test statistic is in the rejection region or not, or use a p-value
- More useful than a hypothesis test is a confidence interval, which expresses uncertainty about how likely repeated samples are to contain the true value