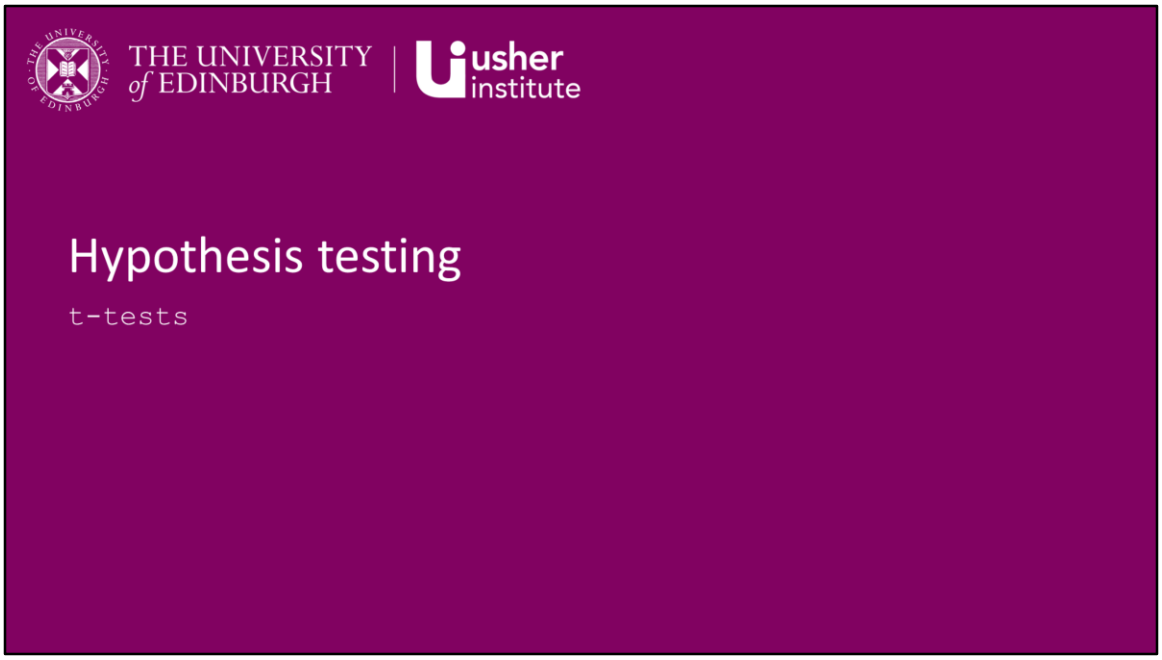
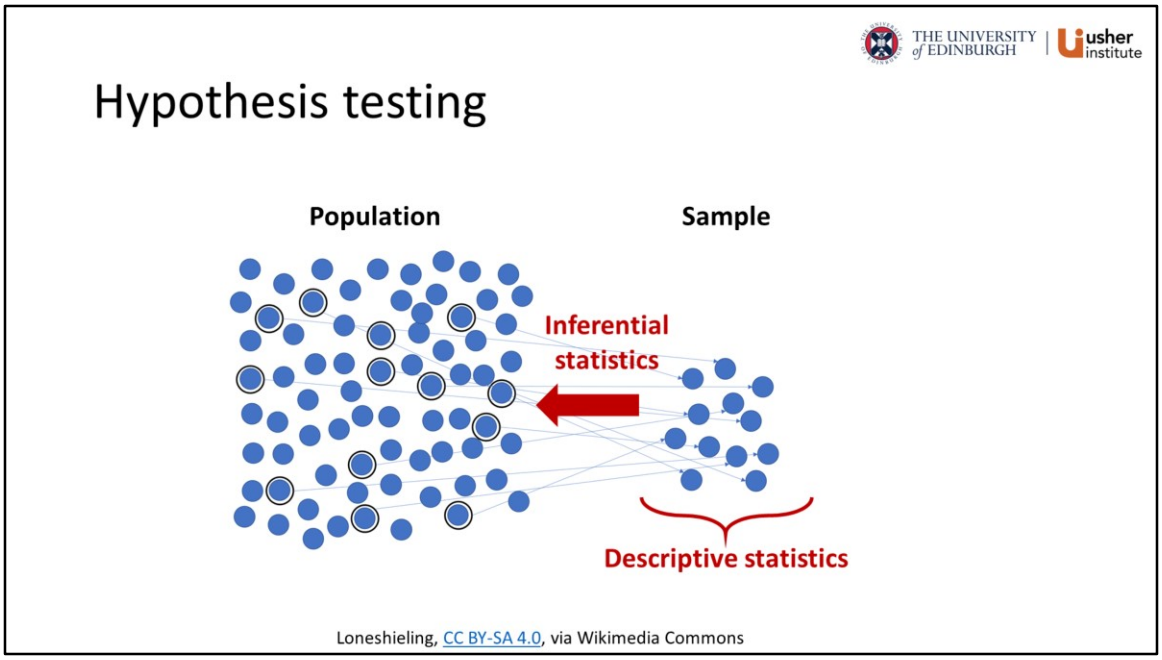
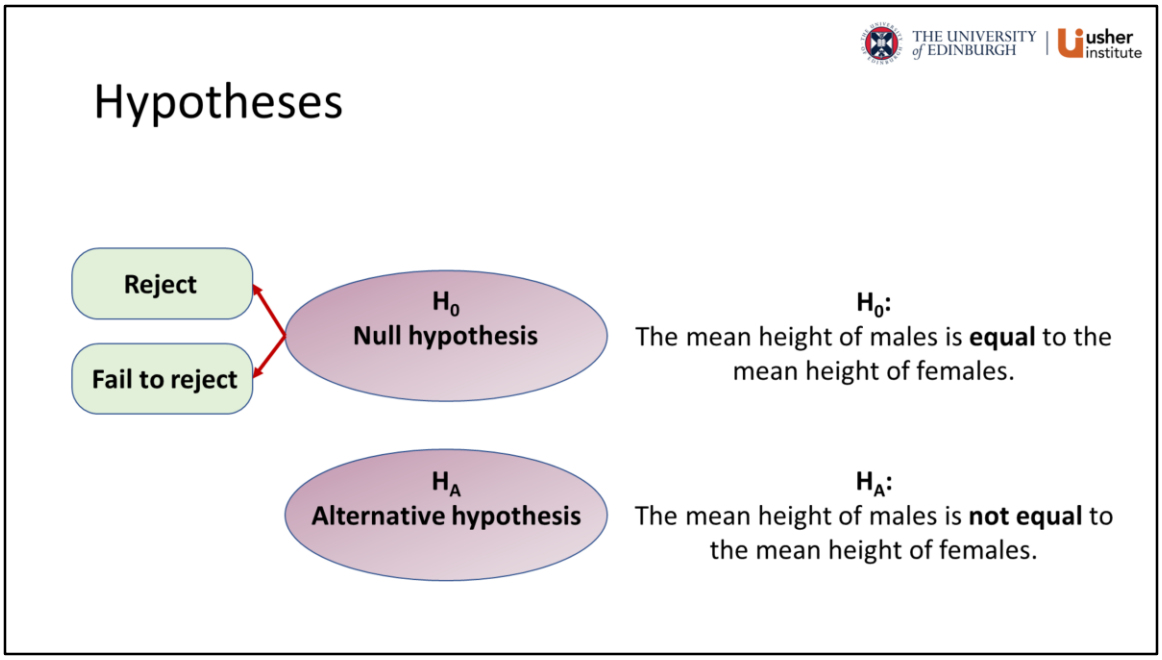


In this video, we are going to introduce a statistical method called hypothesis testing. First, we will look at what a hypothesis test is and some of the terms commonly used when describing them. We will then learn the procedure and the steps in performing a hypothesis test.

Then we will discuss some specific tests, looking at examples, and how the results from these tests should be interpreted.



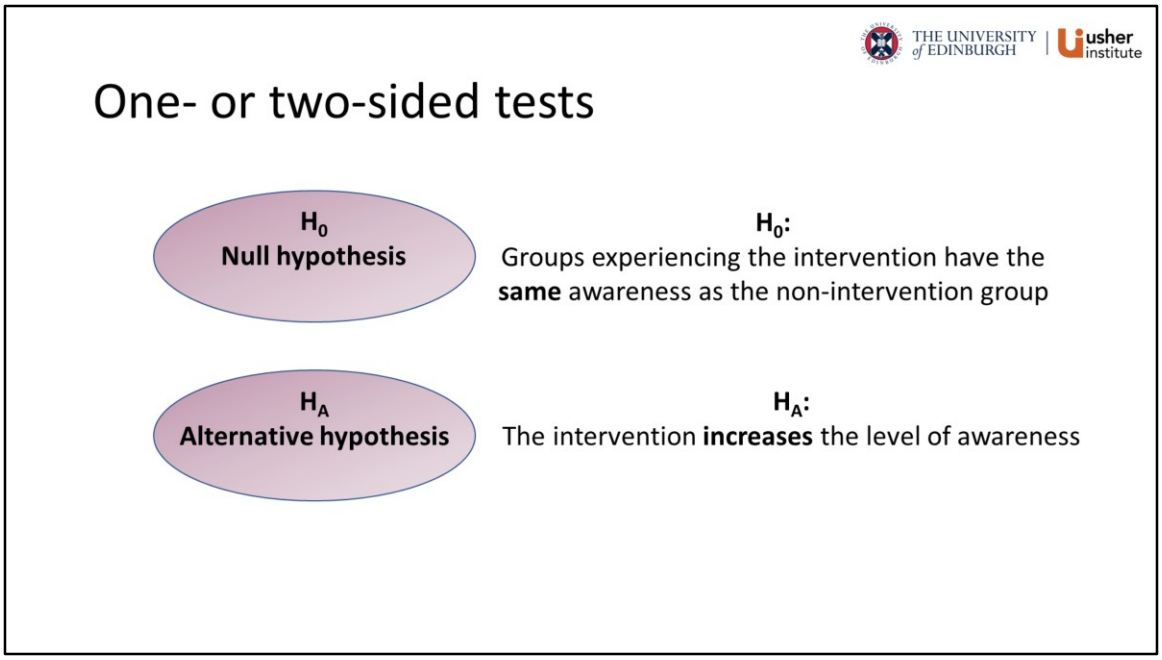
Hypothesis testing is a form of inferential statistics. It aims to examine a set of sample data, and measure the strength of evidence for a hypothesis about a population parameter. To test hypotheses, we follow a set of formal procedures, which will vary depending on the data and hypotheses.



The first step in hypothesis testing is to define the research question and state the associated hypotheses. A statistical hypothesis is a statement about a population, which may or may not be true. For each research question, there is both a null hypothesis, which we call H_0, and an alternative hypothesis, which we call H_A. H_0 is typically that some parameter (such as a correlation or a difference between means) in the populations of interest is zero. For example, the mean height of males is equal to the mean height of females. By contrast, H_A states that there is some effect of interest and so this parameter is not zero. For example, the mean height of males is not equal to the mean height of females. The step of choosing the null and alternative hypotheses we're testing relies on an understanding of the problem context.

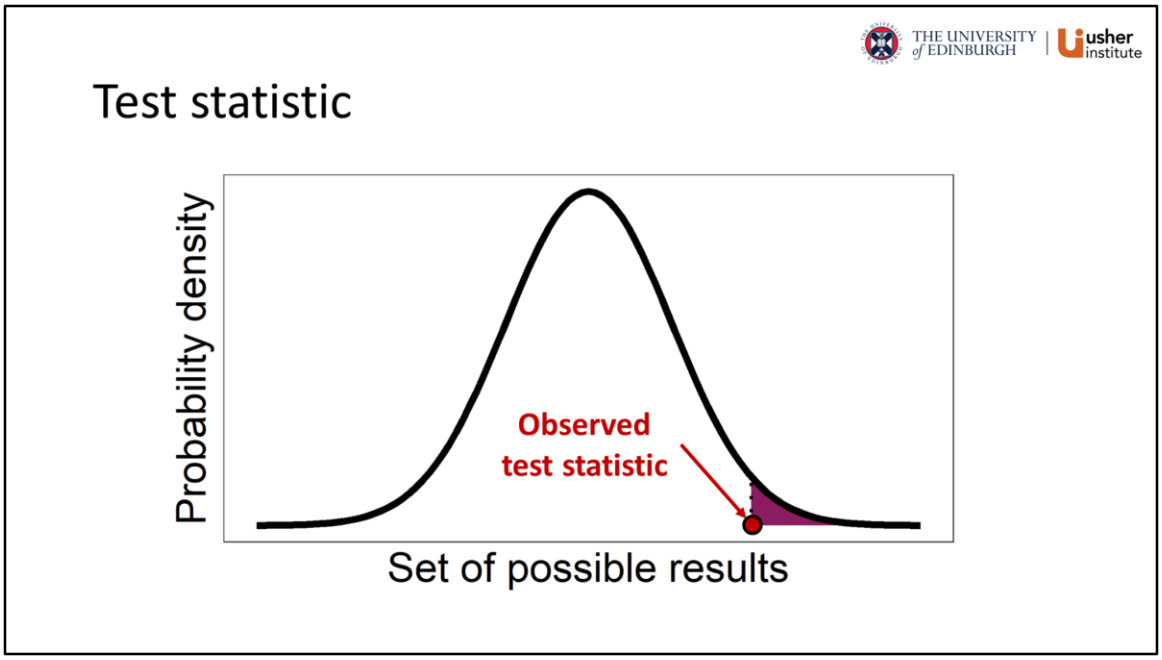
The hypothesis testing process aims to either reject or fail to reject H_0. You can think of the null hypothesis as the default theory that requires sufficiently strong evidence against it in order to reject it.

It's important to remember that hypothesis testing requires making some assumptions about the underlying model. If the assumptions are not met, then the results of the hypothesis test are invalid.



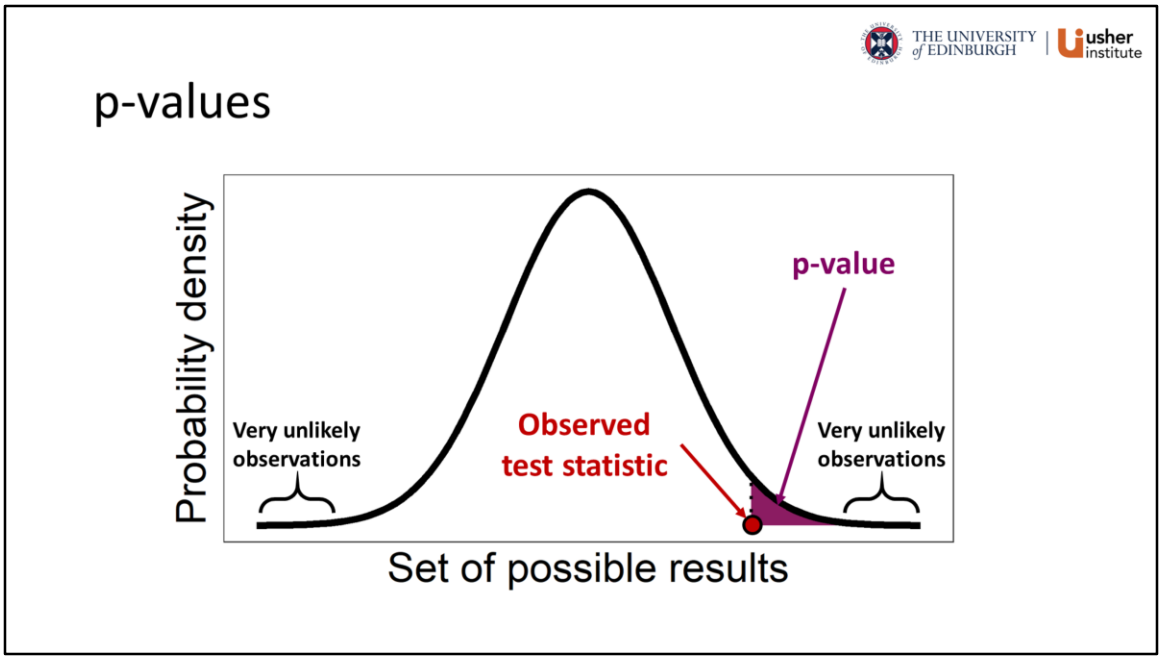
Occasionally you will encounter studies where the effect or difference can only, logically, occur in one direction. Let's take as an example an intervention to increase health awareness, which could be considered to have the same effect as doing nothing, but could not really have less effect. The null hypothesis would be that groups experiencing the intervention have the same awareness as the non-intervention group. And the alternative hypothesis would be that the intervention increases the level of awareness.
This would be a one-sided test, only looking for an effect in one direction.

However, instances where it is logical to test in only one direction are rare and there is scope for abusing this method to increase the power towards the desired result. Therefore one-sided tests should be approached with caution.
A good summary of one-sided and two sided tests, from the UCLA Institute for Digital Research and Education, can be found here: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/



A hypothesis test is typically specified in terms of a test statistic, which is a quantity calculated from the sample data. All test statistics follow known theoretical probability distributions.

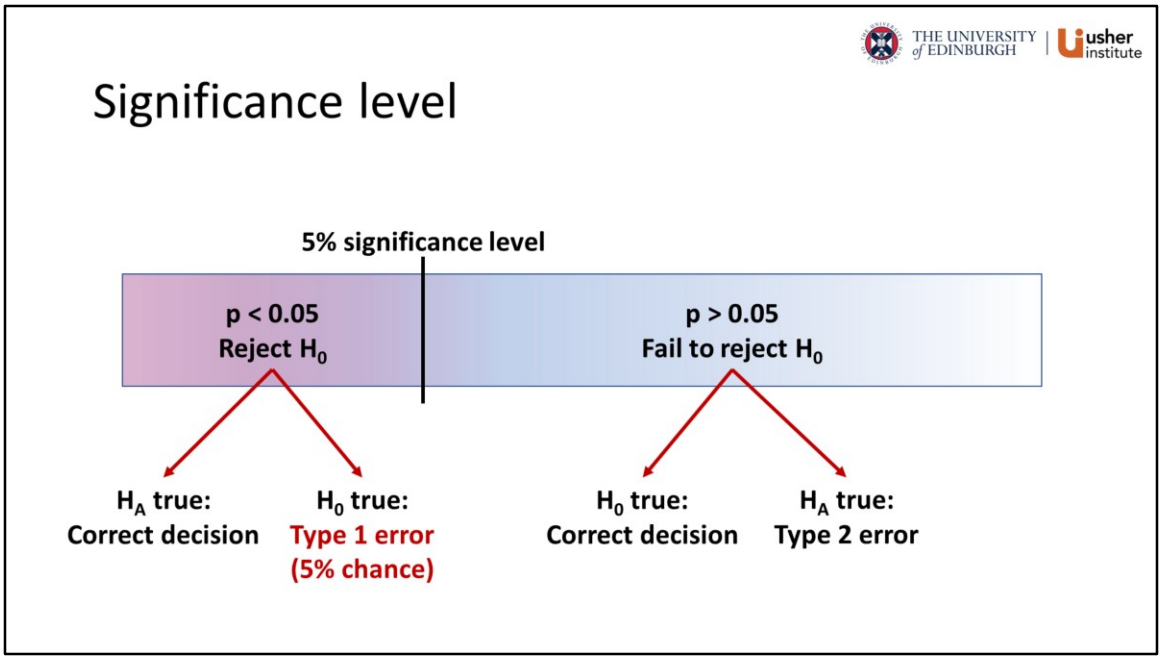
A test statistic is similar to a descriptive statistic, as they both provide a numerical summary of a dataset, and many statistics can be used as both test statistics and descriptive statistics. However, a test statistic is specifically intended for use in hypothesis testing, whereas the main quality of a descriptive statistic is that it is easily interpretable.



The p-value quantifies the statistical significance of a result, the result being the observed value of the chosen statistic. It is the probability of obtaining a test statistic at least as extreme as the test statistic actually observed, under the assumption that the null hypothesis is correct. The p-value is therefore calculated using the probability distribution of the test statistic.

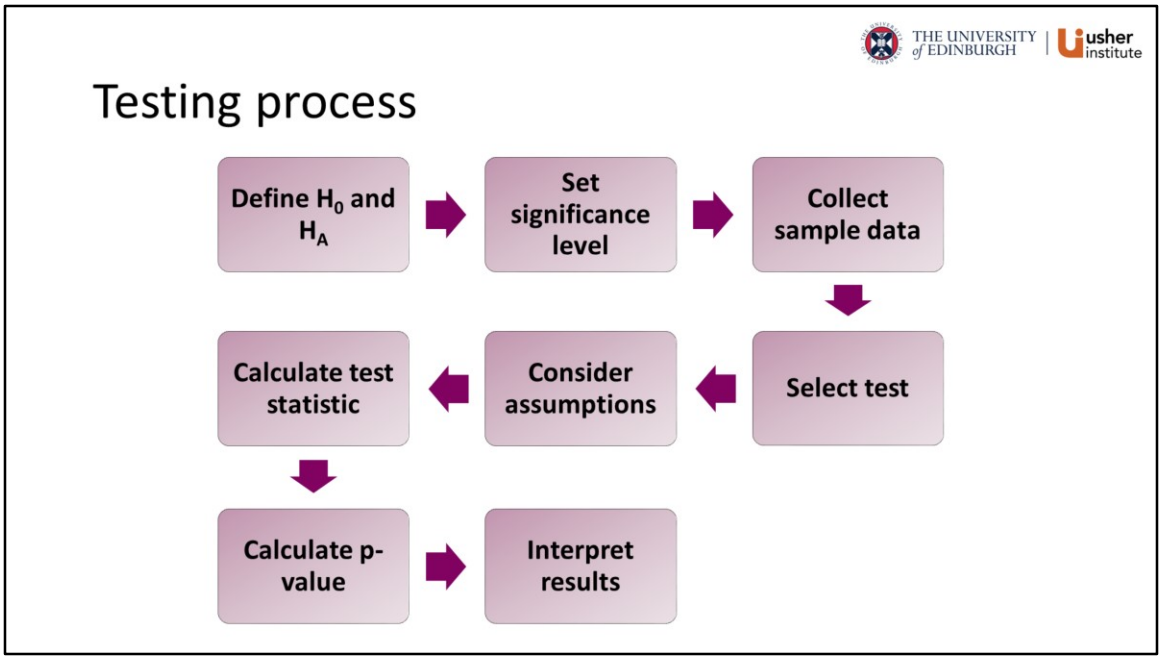
A very small p-value means that such an extreme observed test statistic would be very unlikely if the null hypothesis were true.

All other things being equal, smaller p-values are taken as stronger evidence against the null hypothesis.

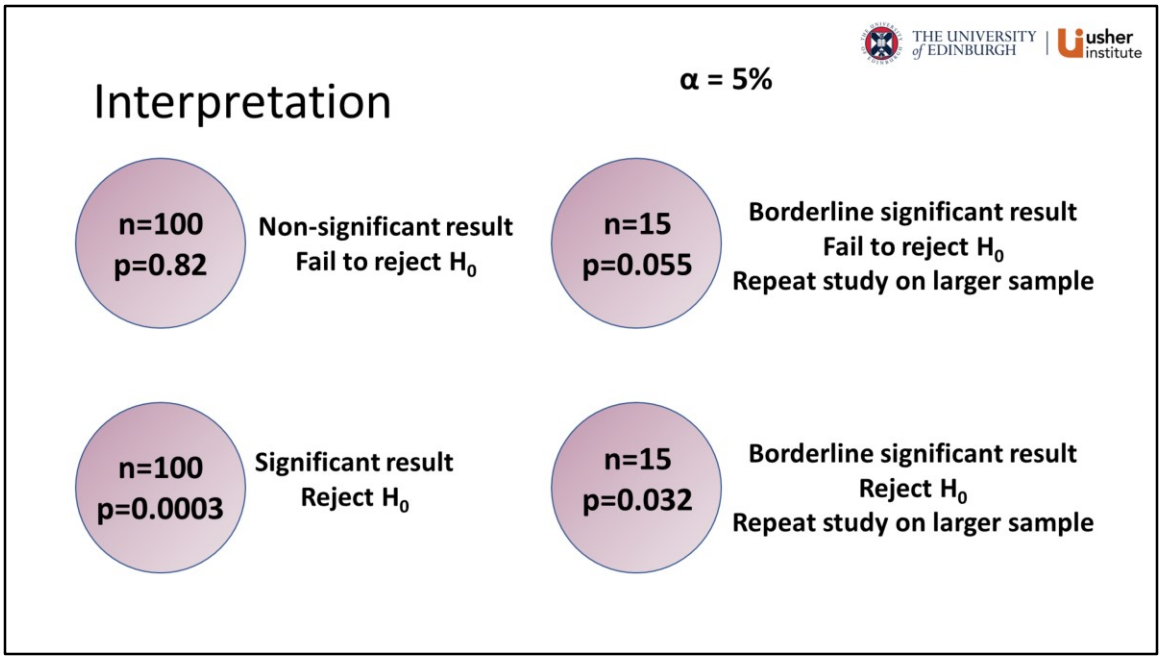


We have seen that a low p-value means that our observed test statistic is unlikely if H_0 is true. To determine how low the p-value should be to reject the null hypothesis, we need to select a significance level. This should be done once you have defined the hypotheses and before you perform the statistical test. The significance level is a probability threshold below which the null hypothesis will be rejected. The significance level, which we call α, is the chance of a type 1 error, which is when we incorrectly reject H_0 in favour of H_A. A standard in scientific studies is to use a 5% significance level, which means that there is a 5% chance that we incorrectly reject H_0 in favour of H_A.

At the 5% significance level, if the p-value is lower than 0.05, then we can reject H_0, but if the p-value is greater than 0.05, then we cannot reject H_0. A result is said to be statistically significant if it allows us to reject the null hypothesis. Loosely speaking, rejection of the null hypothesis implies that there is sufficient evidence against it.



To put all this together, we will describe the general process of hypothesis testing. The first step is to define the null and alternative hypotheses. Then we set the significance level. Then we collect relevant data from a sample. Then we need to decide which test is appropriate for our hypotheses and data. It is important to consider the assumptions that the test makes and ensure they are met. Then, we can calculate the appropriate test statistic, which then allows us to calculate the associated p-value. And finally we need to interpret the results.



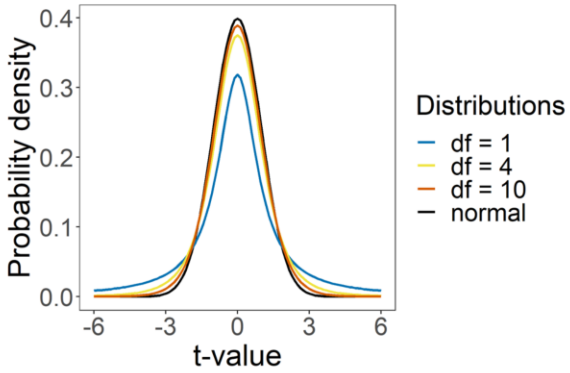
When interpreting the results of a hypothesis test, you should consider the p-value and how close it is to the significance level. It's also important to note the effect size as well as the sample size and the variability within the sample to give context to the test conclusion.

A very small p-value with a large effect size and large sample size would be straightforward to interpret as a statistically significant result that allows to reject the null hypothesis.
Conversely, a very high p-value would be interpreted as not statistically significant, and the null hypothesis can't be rejected.
But if the p-value is close to the significance level, either greater or smaller, and the sample size is small, this would be interpreted as a borderline result, and it should be recommended to repeat the study with a larger sample.

t-test

$$t = \frac{estimated\ value\ -\ hypothesised\ value}{standard\ error}$$

H_0:
The estimated value and the hypothesised value of the population parameter are **equal**.

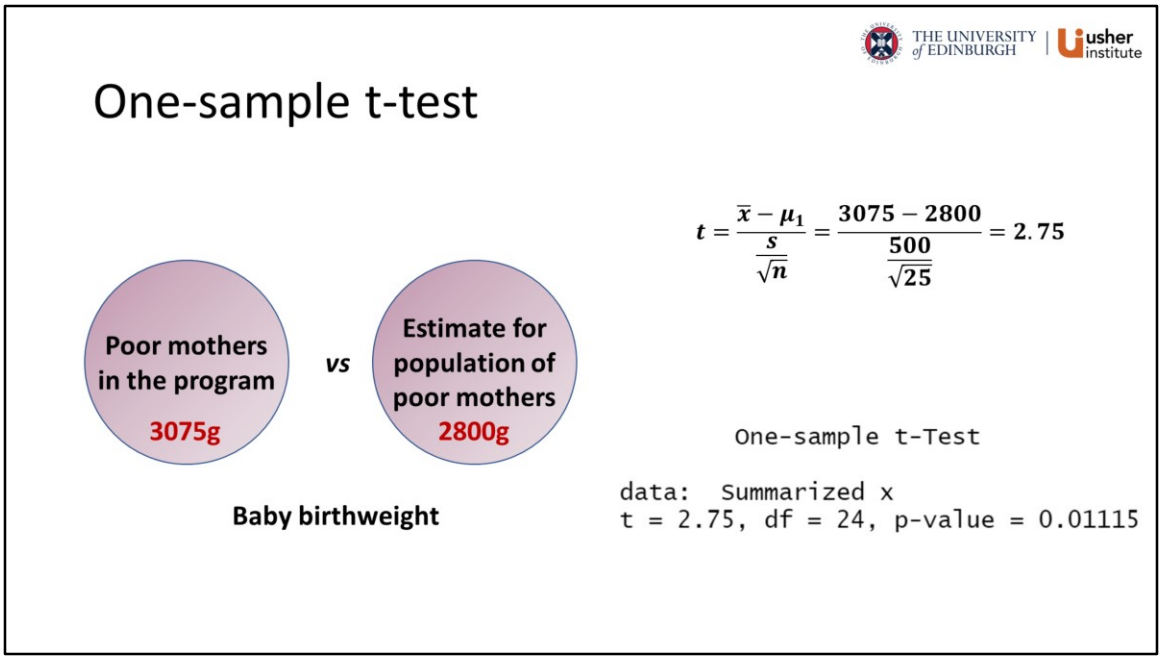


The t-test is a type of hypothesis test in which the test statistic follows the theoretical t-distribution, also called Student's t-distribution. The test statistic for the t-test is called the t-value, and it combines data into one numerical value that summarises the variable of interest. The t-distribution is similar in shape to the normal distribution, but is wider with longer tails. It is characterised by a parameter called degrees of freedom (df), and it approaches the normal distribution for higher degrees of freedom.

The t-value is calculated slightly differently depending on the type of test but it is always the ratio of the departure of the estimated value of a parameter from its hypothesised value to its standard error. The null hypothesis will be that the estimated and the hypothesised values of the population parameter are equal. The t-value gets larger as the data gets further from the null hypothesis (H_0).

A t-test is usually performed when we don't know the population variance. If the population variance is known, then a different test, the z-test, based on the normal distribution, can be used.
The t-test makes a few assumptions that need to be met to ensure your results are valid.

- First, it assumes that the variable of interest follows a normal distribution in each population considered.
- The observations also need to be independent of one another.
- If comparing two populations, they should have the same variance. This assumption is called the assumption of homogeneity of variance.



There are different types of t-tests that can be performed depending on the data and type of analysis required.

The one-sample t-test looks at one sample of measurements of a continuous variable and compares the sample mean to a hypothesised population mean. The null hypothesis is therefore that the population mean is equal to the hypothesised population mean.

For example, we could have a study looking at the link between mother poverty and babies' birthweight. To do this, a local hospital introduced a new prenatal care program to reduce the number of low birthweight babies born in the hospital. 25 mothers who live in poverty participated in this program. Data drawn from hospital records reveals that the babies born to these mothers had a birthweight of 3075 grams, with a standard deviation of 500 grams. The mean birthweight for mothers living in poverty in the US is 2800 grams. The question is whether this program has been effective at improving the birthweights of babies born to poor mothers.
Since we have only one sample and an expected value, 2800 grams, a one-sample t-test is appropriate here. The null hypothesis is that the difference between the mean birthweight for the population of program babies and the expected mean birthweight

for poor mothers is zero. The alternative hypothesis is that the difference between the observed mean birthweight for program babies and the expected mean birthweight for poor mothers is not zero.

The t-value is calculated by dividing the difference between the sample mean, 3075, and the expected mean, 2800, by the standard error, which is the standard deviation, 500, divided by the square root of the sample size, 25. So the t-value is equal to 2.75. Using a statistical package, this corresponds to a p-value of 0.01, which is below our 5% significance level, so we can reject the null hypothesis. Therefore we can conclude that there is a statistically significant difference between the observed mean birthweight for program babies and the expected mean birthweight for poor mothers, and so the prenatal care program seems effective at improving the birthweight of babies.

THE UNIVERSITY of EDINBURGH | usher institute

t-values and degrees of freedom

- $df = n - 1$

df \ α	0,25	0,2	0,15	0,1	0,05	0,025	0,01
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485

The t-test produces two values as its output: t-values and degrees of freedom. t-values are an indication of the difference between the means of two samples. The degrees of freedom (df) are the number of values in a study that have the freedom and ability to vary. The degrees of freedom are essential for assessing the validity, and importance, of hypothesis tests.

The number of degrees of freedom is equal to the sample size minus 1. Written as a formula, it looks like this:

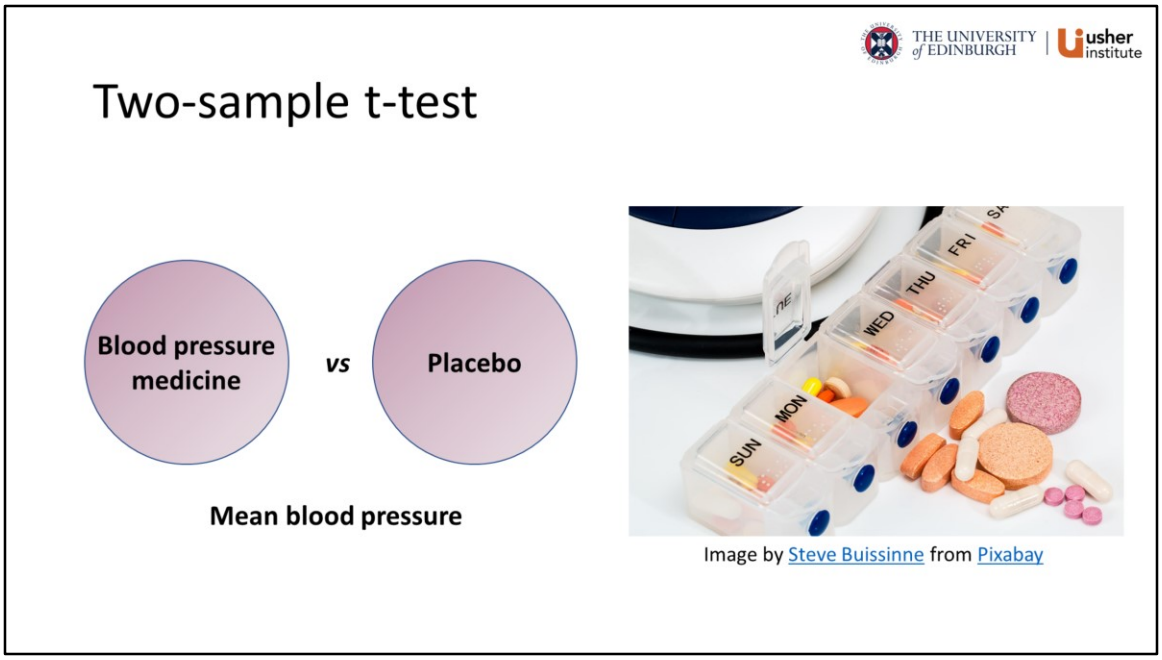
df = n-1 where n is the sample size.

In our one-sample t-test example, the sample size was 25, therefore there are 24 degrees of freedom.

The degrees of freedom can be used to determine the significance of a t-value, using a t-distribution table, as shown here. The purple column highlights the t-values that would be considered significant, at a significance level of 0.05, for the increasing (as you go down) degrees of freedom. Note that we look at the column for 0.025 because this is a two-sided test. The red box shows the critical t-value for a two-sided t-test with 24 degrees of freedom, which is 2.064. The t-value for our one-sample t-test was 2.75, which is higher than 2.064, so the difference is indeed statistically significant.

This method is not commonly used anymore, as statistical packages will calculate the

p-value instead of relying on estimating significance from a table.



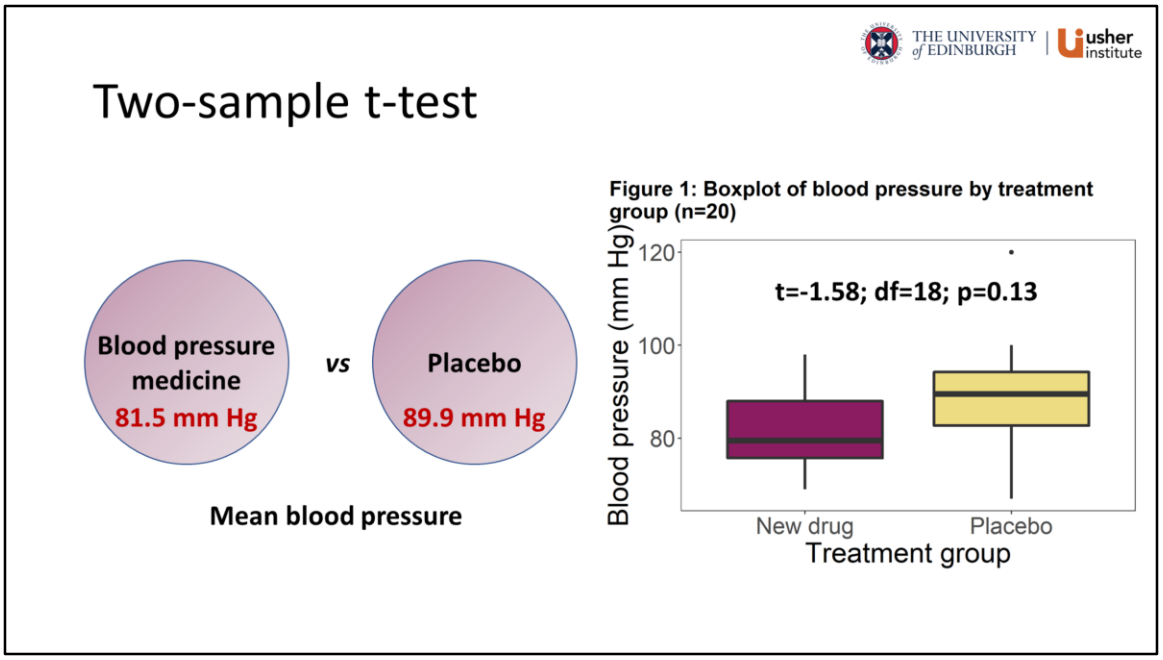
The two-sample t-test is a type of hypothesis test commonly used to determine if there is a significant difference, for a continuous variable, between the means of two unrelated groups. Essentially, the t-test allows us to compare the mean values from two sets of data and decide if they come from the same population.
For example, if we measured the systolic blood pressure in two groups of ten patients with high blood pressure, one taking blood pressure medicine, and the other taking a placebo, they would be unlikely to have exactly the same mean and standard deviation. A t-test would be used to judge whether the treatment has a real and significant influence on the measured blood pressure.

Calculating the t-value for a two-sample t-test requires three pieces of information:

- the difference between the mean values from each group (called the mean difference)
- the standard deviation of each group
- the sample size for each group

A large t-value indicates that the groups are different. A small t-value indicates that the groups are similar.
Once you have calculated the test statistic, you can use it, along with the number of

degrees of freedom, to calculate the p-value.



Going back to our example of blood pressure treatment, the null hypothesis of a two-sample t-test is that the population means for both groups are equal. The samples means are 81.5 for the group taking the blood pressure medicine, and 89.9 for the group taking a placebo.

Using a statistical package, we can calculate the t-value, degrees of freedom and p-value for this test. The t-value is -1.58 with 18 degrees of freedom. This small t-value corresponds to a p-value of 0.13, which is quite high and above the 5% significance level. Therefore, we cannot reject the null hypothesis that the population means of the two treatment groups are equal, which means that the difference between the mean blood pressure of the two treatment groups is not statistically significant. We can conclude that the treatment does not have a significant effect on blood pressure, although the sample size is quite small, and the study would be better repeated on larger samples.

Paired t-test

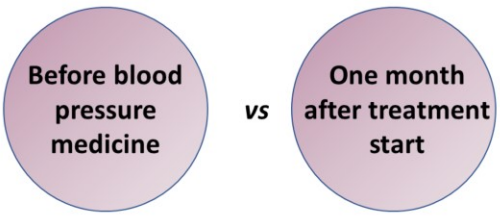


Image by Steve Buissinne from Pixabay

Another variant of the t-test is the paired t-test. This compares a numerical variable between two related samples. Each observation is paired with another, either because the variable is measured for each individual at two different points, or because the variable is measured in two groups where individuals are matched for some other variables. This is what we call paired data. Note that, in this case, the assumption is that the individual differences are normally distributed. The null hypothesis of a paired t-test is that the mean difference between the two related samples is equal to zero for the population.

For example, in the context of the blood pressure example, you might measure the blood pressure of a group of patients before their treatment starts, and then measure their blood pressure another time, one month after starting the treatment.

Summary

- **Hypothesis testing process**
  - Null vs alternative hypotheses
  - Test statistic
  - p-value
- **Interpretation**
  - Distribution
  - Significance level
- **t-tests**
  - One-sample
  - Two-sample

In this video, we have seen what a hypothesis test is and the different steps involved in the process, in particular, defining the null and alternative hypotheses, calculating the appropriate test statistic and the associated p-value. We have also discussed how to interpret hypothesis tests based on the p-value and set significance level. And finally, we looked at different types of one- and two-sample t-tests.