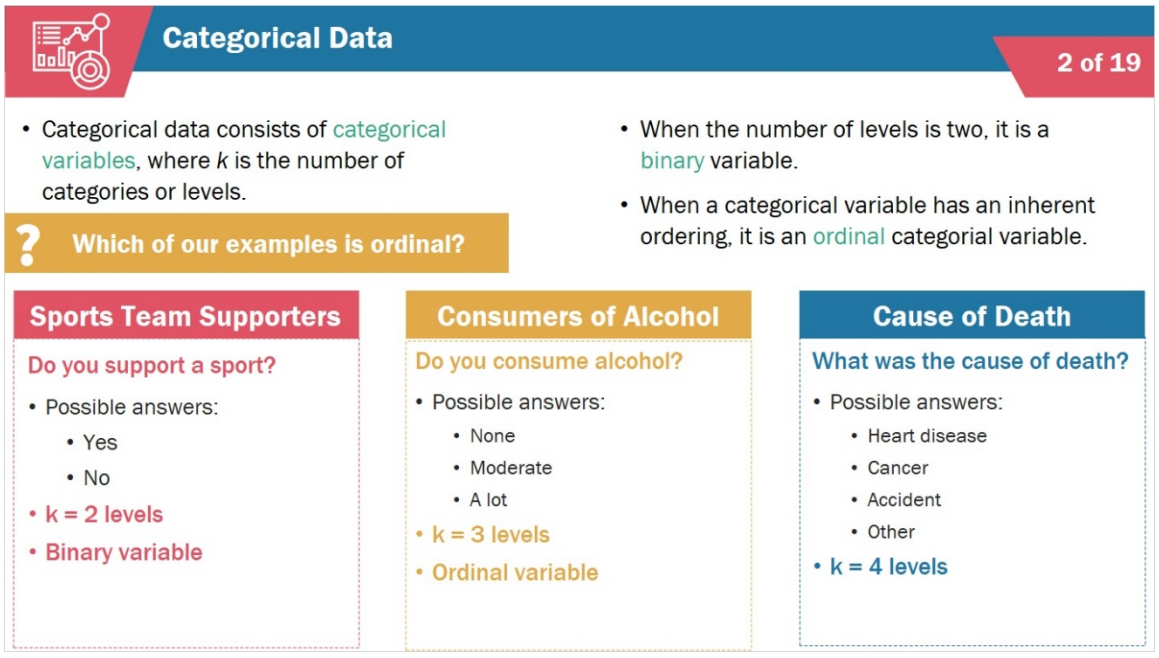
Analysing Count Data

Slide 1: Introduction

Welcome to this presentation on analysing count or categorical data. My name is Caroline Brophy. In this session we will explore statistical methods for analysing count or categorical data variables.

We will study hypothesis testing and confidence intervals for a single proportion, and for comparing two proportions. We will also explore contingency tables for two categorical variables and a chi-square test of independence for two categorical variables.

Slide 2: Categorical Data



In earlier sessions in this module, the focus was on analysing continuous data. For example, data might have been collected on the height of a particular plant species, at random from the population of the species. In that case, the hypothesis test or confidence interval aims to make statements about the true population mean height of the tree species.

In this session, we will examine categorical data, where *k* is the number of categories or levels.

Let's take a look at some examples of categorical data variables.

Suppose a random sample of people from a population are asked if they a supporter of a particular sports team. The possible answers are 'yes' or 'no' and the response for each person is recorded. This is an example of a categorical data variable, with *k* = two levels. A study such as this might be carried out to find out what proportion of people in the true population are supporters of the sports team.

Another example of a categorical variable is the amount of alcohol a person consumes, with possible answers 'none', 'moderate', or 'a lot'. This time the categorical variable has three levels. And of course, the decision as to how to label the categories, or how many categories to have, will often lie with the person conducting the study.

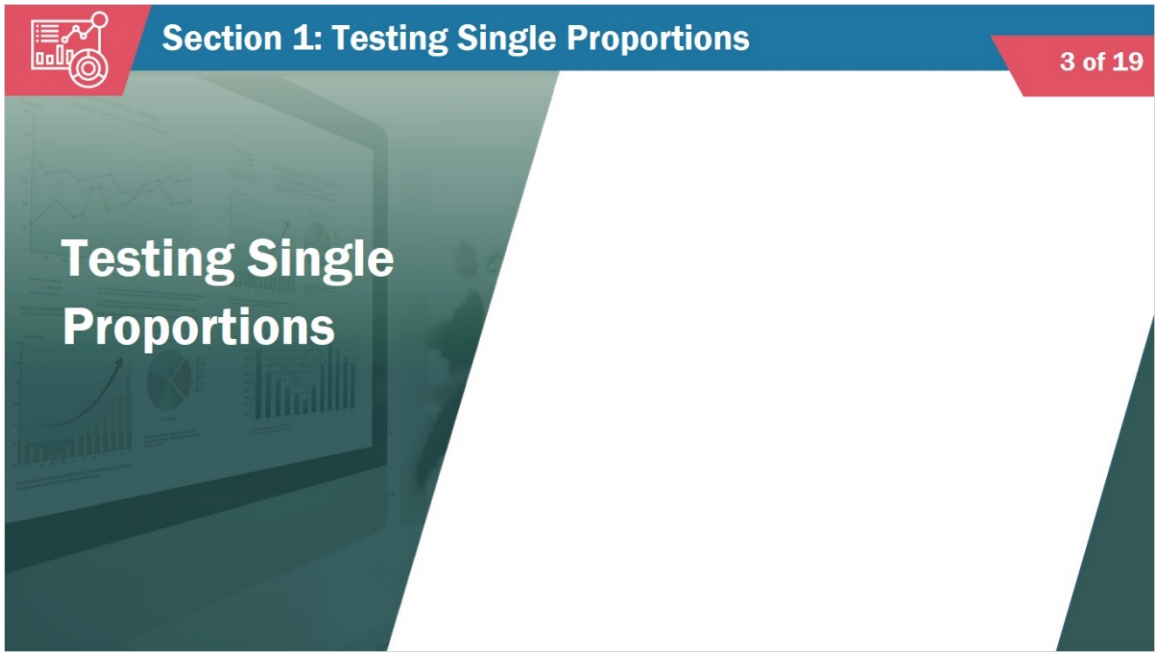
A third example is cause of death where possible answers are 'heart disease', 'cancer', 'accident', and 'other'. This variable has four levels, so *k* is equal to 4.

When the number of levels is two, we call this a binary variable. In the examples, the supporter of a sports team is an example of a binary variable.

When the ordering of the levels of a categorical variable has meaning, we call it 'ordinal'.

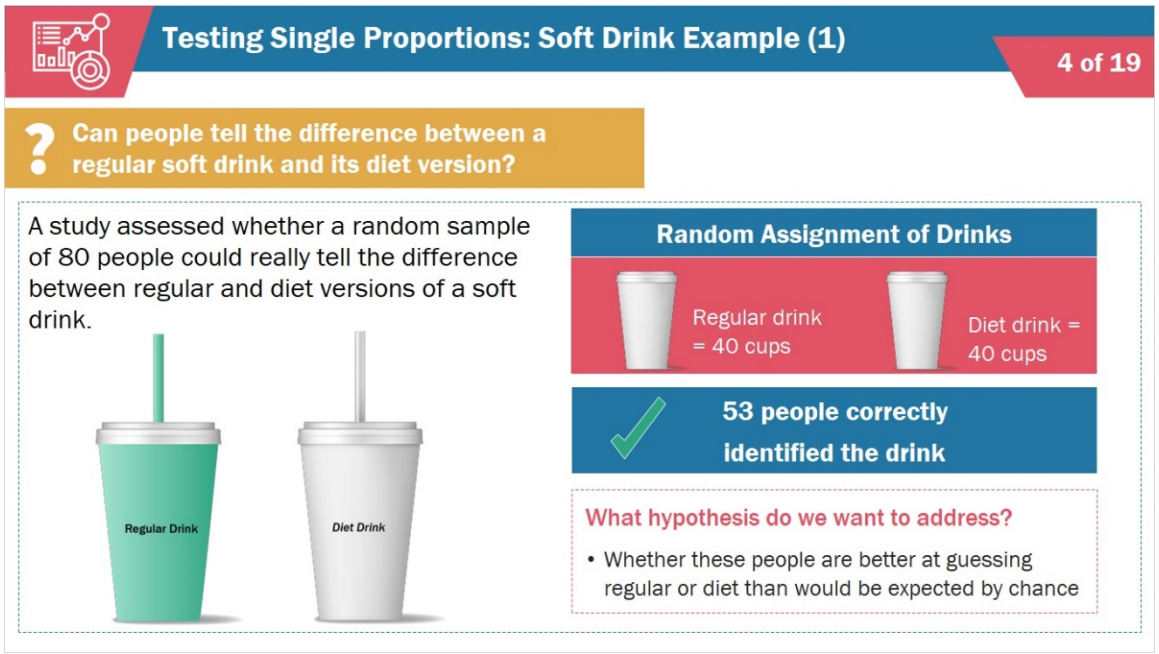
Let's consider our three examples, are any of them ordinal? The sports team binary variable is not: it does not matter whether yes or no comes first. The causes of deaths could be listed in any order, for example, it would not matter if 'accident' came first instead of third. However, for the alcohol question, the levels of the category are increasing in magnitude of consumption and the order has logic to it. This variable is an example of an ordinal variable.

Slide 3: Section 1: Testing Single Proportions



We will first explore how to analyse a single proportion.

Slide 4: Testing Single Proportions: Soft Drink Example (1)



We will start with an example which involves a soft drink.

Here is a question for you to consider: can people tell the difference between regular and diet versions of soft drinks? If you were given a soft drink, for example, a glass of 7-up, and asked if it was diet 7-up or not, do you think you could you answer correctly?

A study was carried out to assess whether people who believe they can tell the difference between a regular soft drink and a diet version of it, can actually tell the difference.

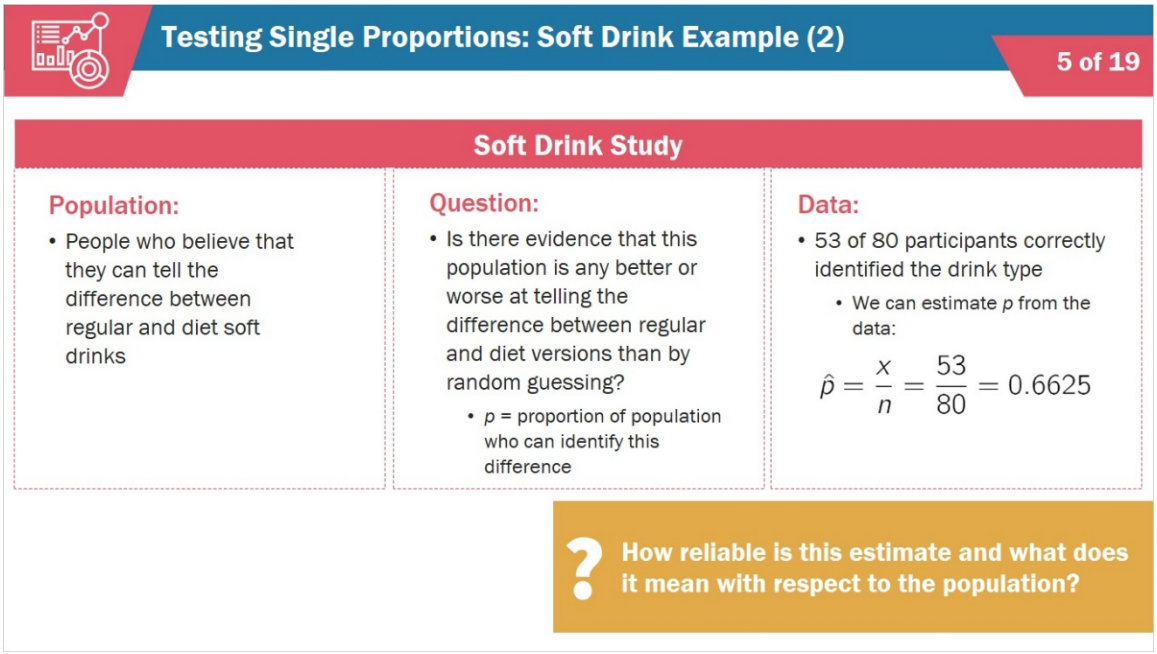
Some people believe that they can tell the difference between a regular soft drink and a diet version of it. A researcher wanting to test this claim randomly sampled 80 such people. Eighty plain white cups were filled with the soft drink, half diet and half regular through random assignment. Each person took a sip from their cup and was asked to identify the drink as diet or regular. Of the 80 participants, 53 correctly identified the drink.

What is the question or hypothesis that we want to address here? Really, we want to know if these people are better at guessing regular or diet than would be expected by chance.

And what would be expected by chance?

Well, we would expect that a person would guess correctly 50% of the time.

Slide 5: Testing Single Proportions: Soft Drink Example (2)



The population here is people who believe that they can tell the difference between regular and diet soft drinks.

The question here is really whether or not this population of people are any better or worse than random guessing at telling the difference between a diet and regular soft drink. We use *p* to denote the proportion of people in the population who can identify the difference between a diet and regular soft drink.

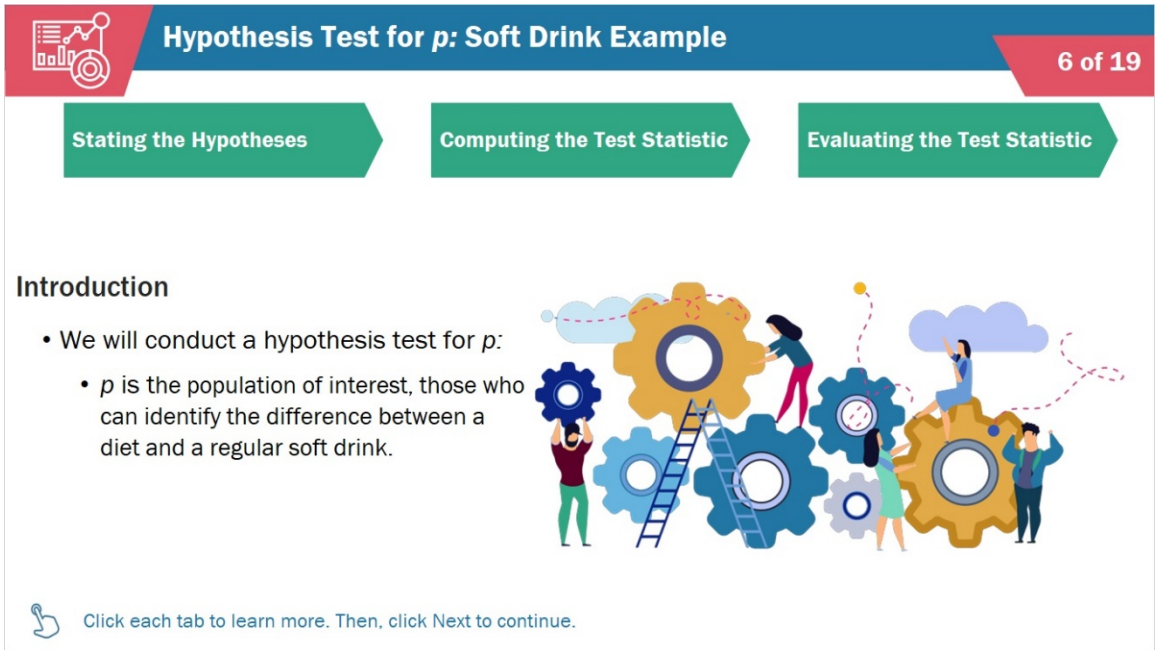
The data arising from the study showed that 53 of 80 participants correctly identified the drink type. We can use this data to find an estimate of *p*, our parameter of interest.

Here we find that the estimate of *p* is 53 divided by 80 which equals 0.6625. Or we estimate that 66% of the population can indeed tell the difference.

Note the use of the hat symbol over the p; this is to indicate that it is an estimate from data. When we say *p,* we mean the true population parameter, and when we say *p_hat* we mean the estimate of the proportion from the data.

Of course, with any estimate comes uncertainty. We will now look at ways to explore how reliable our estimate is and understand what it means with respect to the population.

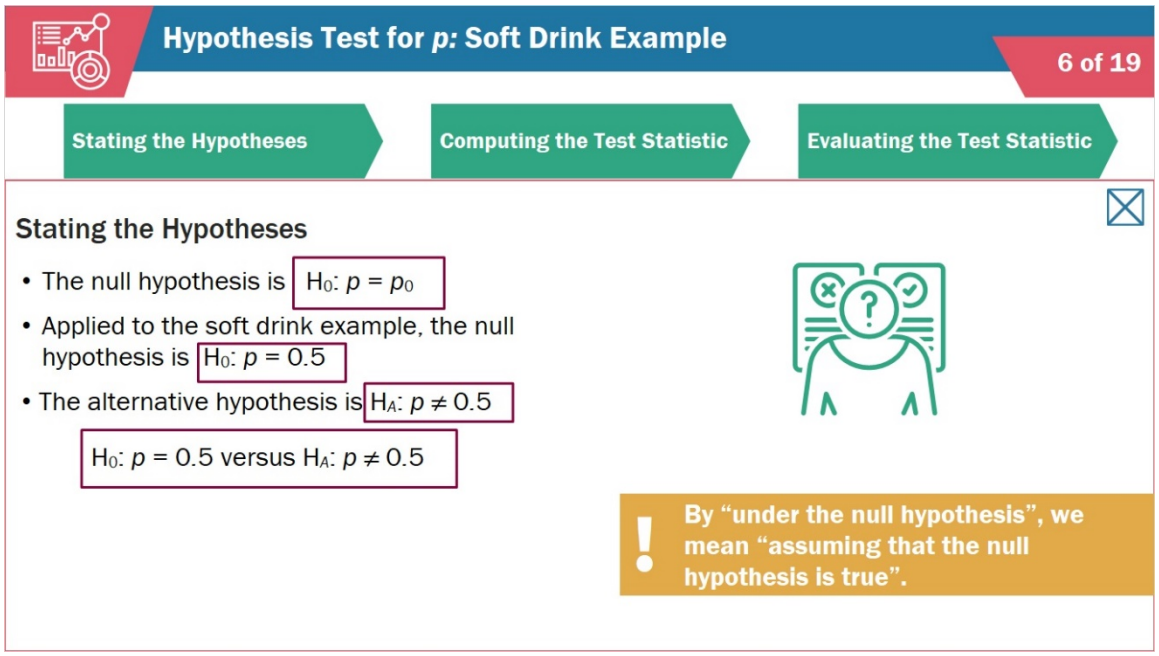
Slide 6: **Hypothesis Test for *p*: Soft Drink Example**



We are going to conduct a hypothesis test for *p*, the proportion of people in the population who can identify the difference between a diet and regular soft drink. The logic we will follow to do this will be similar to that used in previous sessions for testing the population mean. We will specify our hypotheses, compute our test statistic, evaluate the test statistic against the reference distribution, and give our conclusions.

Click each tab in turn to learn more. When you are ready, click next to continue.

Tab 1: Stating the Null Hypothesis

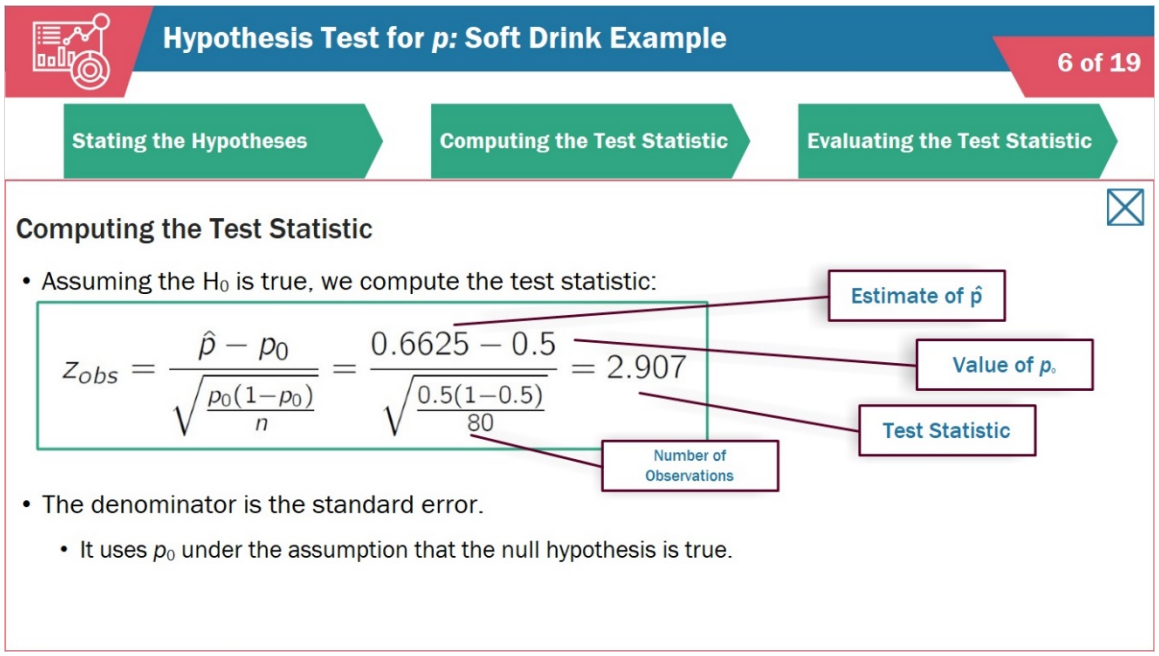


We start by stating the null hypothesis; p is the population proportion of interest, p_0 is the hypothesised value of p. So, our null hypothesis is that p equals p_0.

In this example, the null hypothesis is that these people are no better or worse than random guessing at telling the difference between diet and regular soft drinks. With random guessing, there is a 50:50 chance of getting it right, so under the null hypothesis, we hypothesise that the population proportion is equal to 0.5. By saying "under the null hypothesis, we mean "assuming that the null hypothesis is true".

The alternative hypothesis is that the population proportion does not equal 0.5.

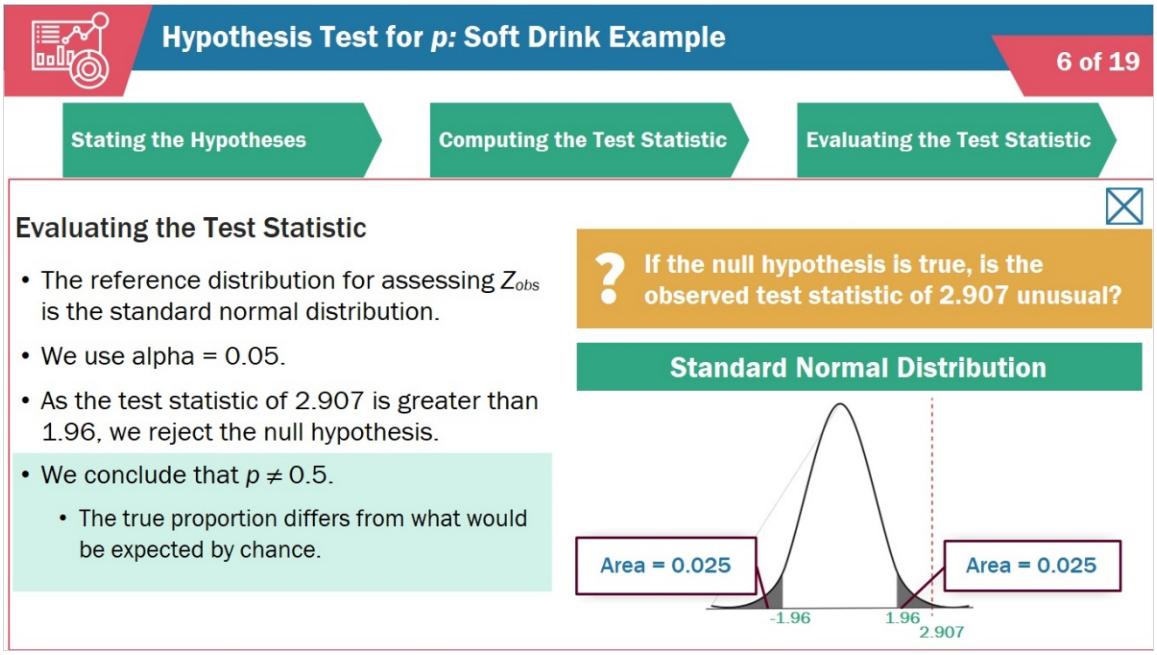
Tab 2: Computing the Test Statistic



The test statistic takes a similar form to what you saw in previous sessions: the parameter estimate from the data, minus the value of the parameter under the null hypothesis, divided by the standard error. Since the test statistic is constructed assuming that the null hypothesis is true, p_0 is used in computing the standard error on the denominator.

We fill in the values: our estimate p hat is 0.6625, our value of p under the null hypothesis is 0.5 and n the number of observations in the sample is 80. Our test statistic value works out to be 2.907.

Tab 3: Evaluating the Test Statistic



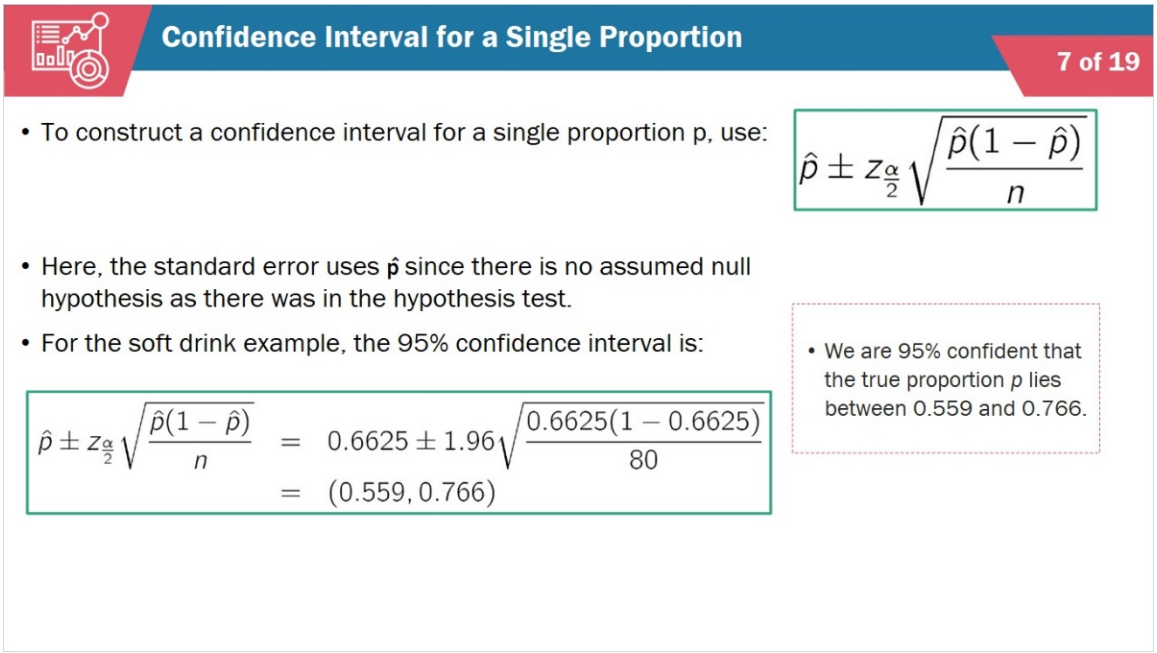
If the null hypothesis is true, the test statistic is a random draw from a standard normal distribution. If the null hypothesis is true, is the observed test statistic of 2.907 unusual? Let's take a look.

Here we can see the standard normal distribution. Is our test statistic value unusual for this distribution? Our observed test statistic, or our observed z value, is equal to 2.907. What defines unusual? Is 2.907 unusual?

We will use alpha = 0.05 here, and since we are doing a two-sided test, we find the critical values such that the area under the curve outside our critical values sum to 0.05, with each one equalling 0.05/2 = 0.025. So, the shaded grey area to the left of -1.96 is equal to 0.025 and the same for the area to the right of 1.96, it is also equal to 0.025. We say that any test statistic that lies outside these bounds is unusual for the distribution.

We have observed a test statistic of 2.907, which is greater than 1.96, and so we reject the null hypothesis and conclude that *p* does not equal 0.5. We can say that the true population probability differs from what would be expected based on chance alone.

Slide 7: Confidence Interval for a Single Proportion



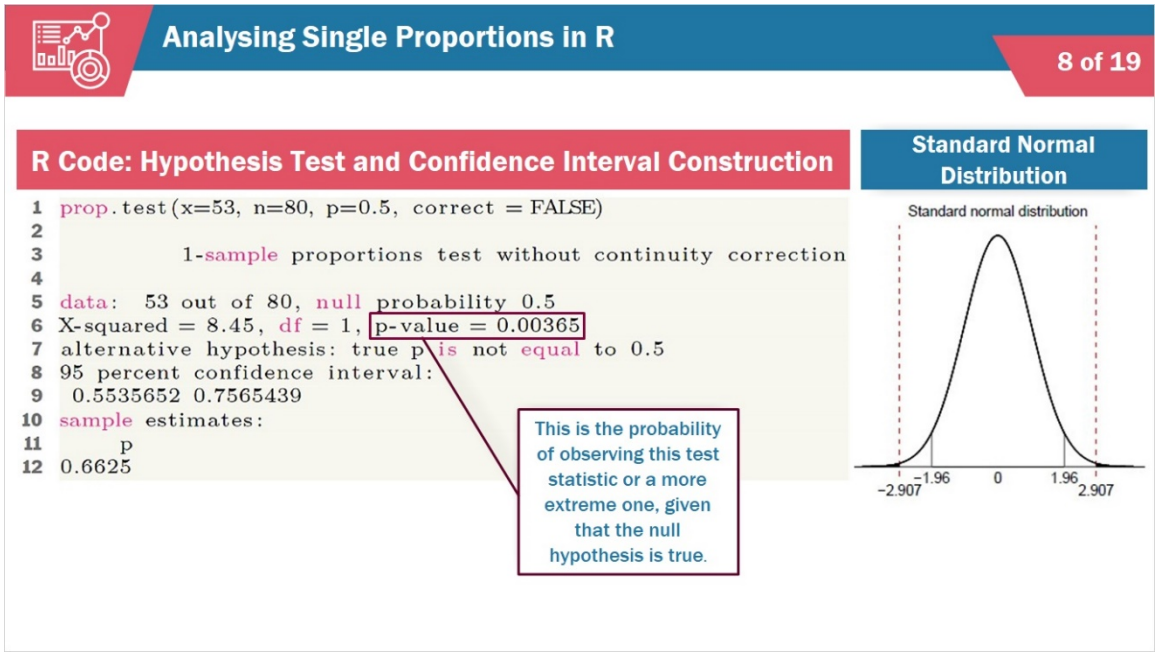
We may also wish to construct a confidence interval for *p*, our population parameter of interest. The structure of the confidence interval is similar to what you saw in previous sessions: the parameter estimate from the data, plus or minus the critical value times the standard error.

We use *p* hat to compute the standard error in the confidence interval, as opposed to p_0 for the hypothesis test standard error. This is because the test statistic is computed assuming the null hypothesis is true, but for a confidence interval, there is no null hypothesis assumption.

We plug in our estimates, *p_hat* equals 0.6625, *n* = 80 and our critical value is 1.96. Our interval goes from 0.559 up to 0.766.

We are 95% confident that the true population parameter *p* lies between 0.559 and 0.766. Remember, the interpretation of a confidence interval is always in relation to the population parameter of interest.

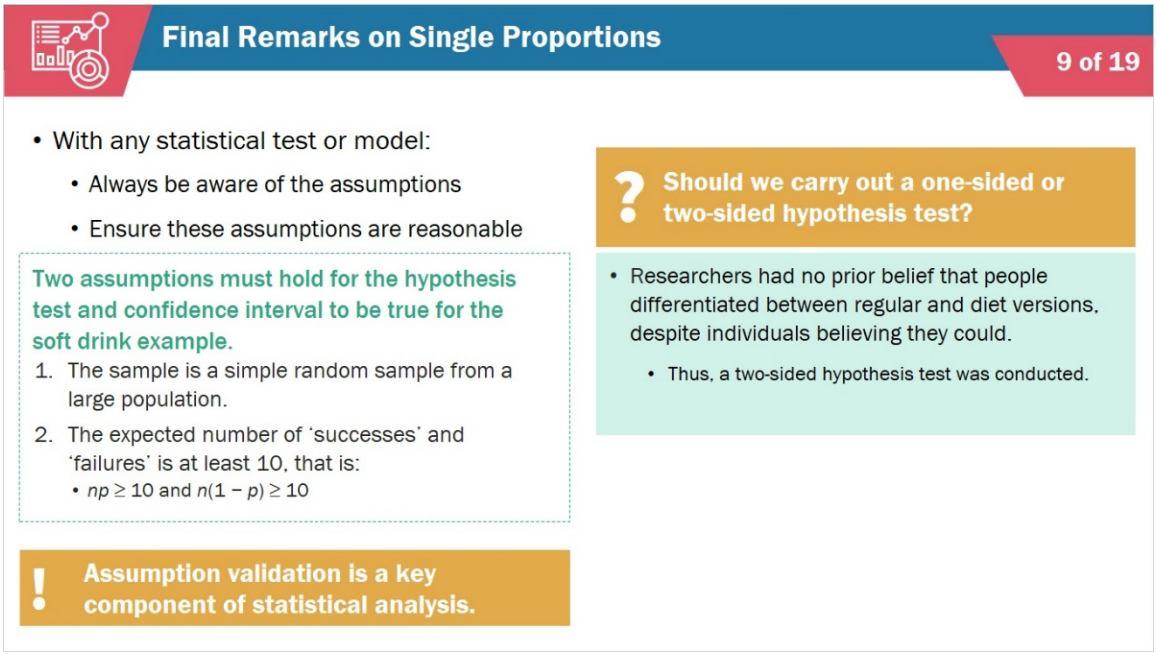
Slide 8: Analysing Single Proportions in R



We can do the hypothesis test and construct the confidence interval using the R code shown. We use the function prop.test() and specify the number of successes, the total number of observations and the value of *p* under the null hypothesis.

We see that our p-value is 0.00365 and this is the probability of observing our test statistic or a more extreme value, assuming that the null hypothesis is true. Graphically, 0.00365 is the area under the curve to the right of 2.907 plus the area under the curve to the left of -2.907. Effectively, we condition on the null hypothesis being true, and examine the probability that our observed test statistic or a more extreme value comes from the null distribution.

Slide 9: Final Remarks on Single Proportions



With any statistical test or model, there will be assumptions on which the validity of the inference from the analysis depends. It is a statistician's responsibility to firstly, know what these assumptions are, and secondly, to ensure they are reasonable.

Here, there are two assumptions to check for the validity of the hypothesis test and the confidence interval.

The data must be a simple random sample from a large population. It is crucial to determine how a study was carried out and assess its validity. In this case, we are told people were randomly sampled from the population of interest.

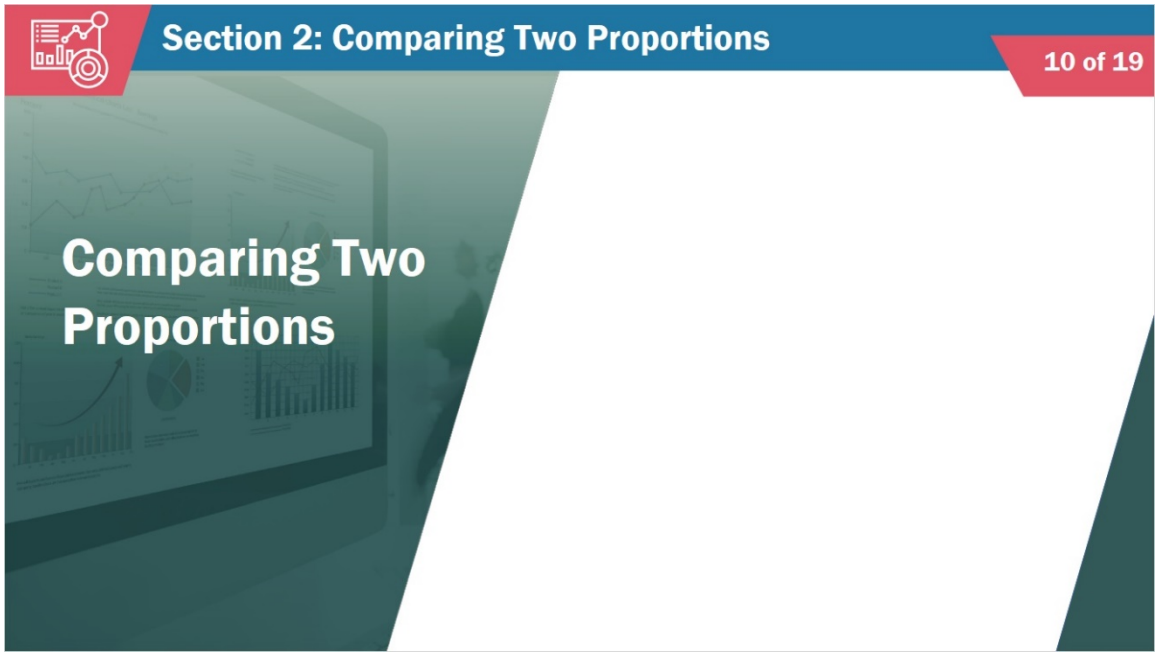
The expected number of successes and failures must be at least 10. We check this for the hypothesis test using p_0 and for the confidence interval using *p*_hat. For the hypothesis test, $np_0 = n(1 - p_0) = 80*0.5 = 40$, which is greater than 10. For the confidence interval, n**p_hat* = 53 and *n**(1 – *p_hat*) = 27, which are both greater than 10.

It is crucial that these assumptions are checked since the analysis results may not be valid if the assumptions are not met. Assumption validation is a key component of any statistical analysis.

When performing a hypothesis test, you must decide whether a one-sided test or two-sided test is best. The only reason to do a one-sided test is if the experimenters have an *a priori* belief that the outcome is in one particular direction. When analysing data, we never decide on the direction of the alternative hypothesis, or the value of the parameter under the null hypothesis, based on our observations in the data.

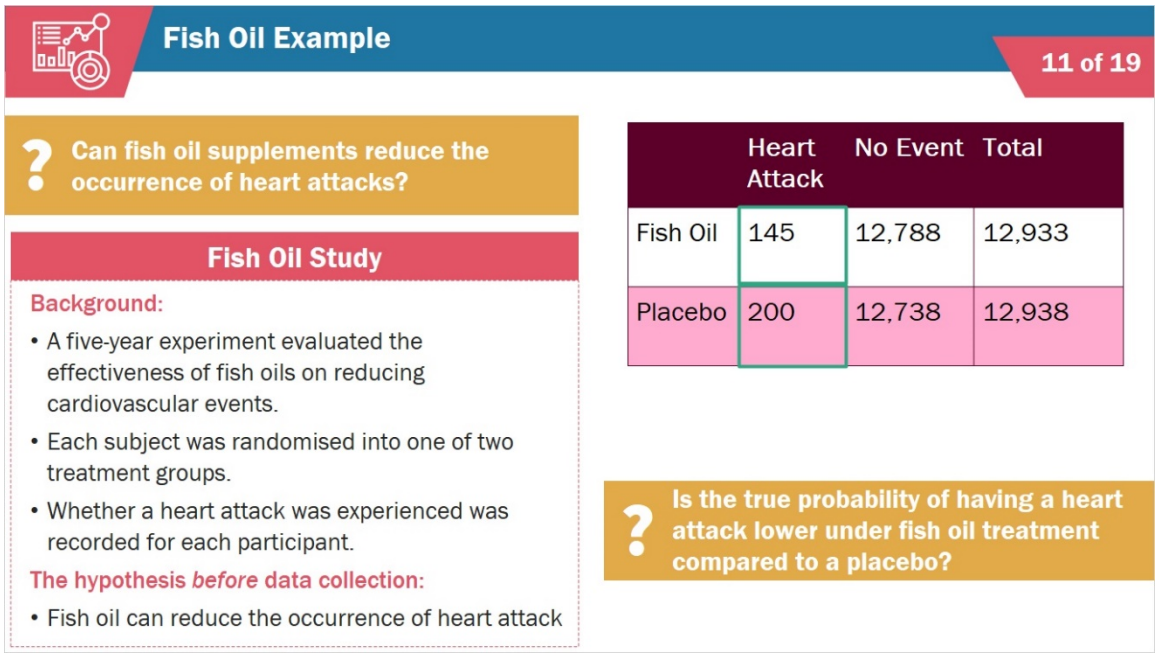
In the soft drink example, the researchers conducting the study had no prior belief that people could tell the difference between regular and diet soft drinks, despite the individuals believing they could. Thus, a two-sided hypothesis test was conducted.

Slide 10: Section 2: Comparing Two Proportions



We will now move on to comparing two proportions.

Slide 11: Fish Oil Example



Let's start with a motivating example: can fish oil supplements help reduce the occurrence of a heart attack? An experiment was carried to address this question.

Here are the details of the experiment:

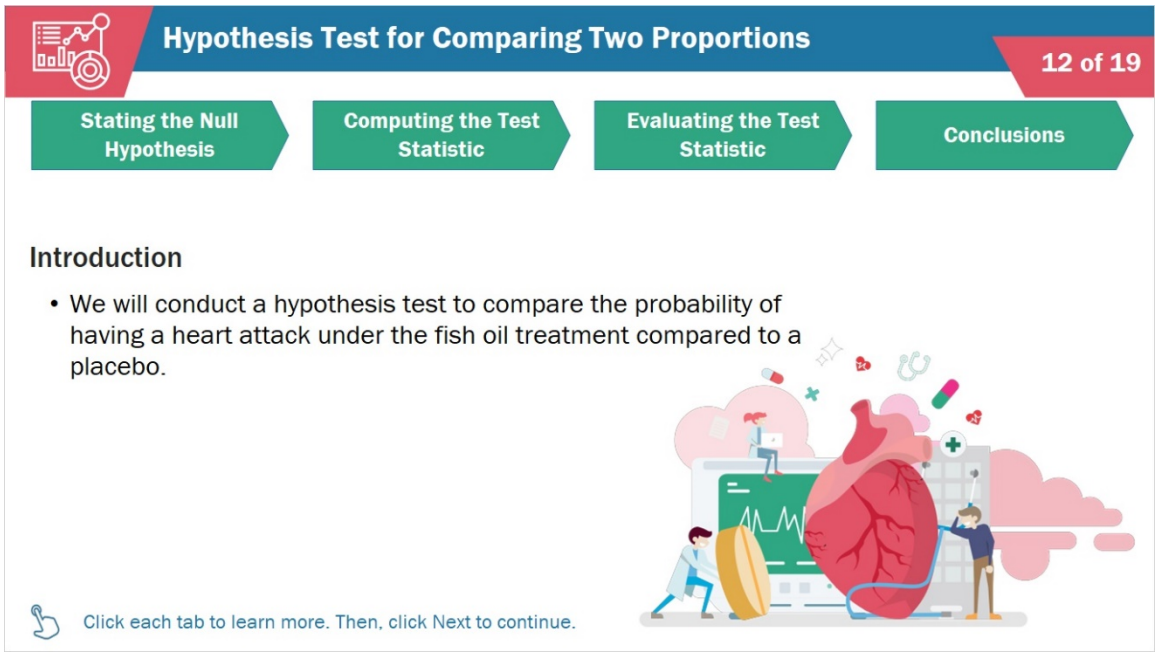
A 5-year experiment was conducted to evaluate the effectiveness of fish oils on reducing cardiovascular events, where each subject was randomised into one of two treatment

groups. Whether a heart attack was experienced was recorded for each participant and shown in the following table. Before the data was collected, it was hypothesised that fish oil could reduce the occurrence of heart attack.

The data are displayed here in a contingency table. Participants are categorised according to whether they received the fish oil supplement or a placebo and whether or not they had a heart attack during the five years of the study. For example, there were 145 people who had heart attacks and were on the fish oil treatment, while there were 200 on the placebo treatment who had a heart attack. There were approximately 13,000 participants in each group.

We want to know if the probability of having a heart attack is lower under the fish oil supplement treatment compared to placebo.

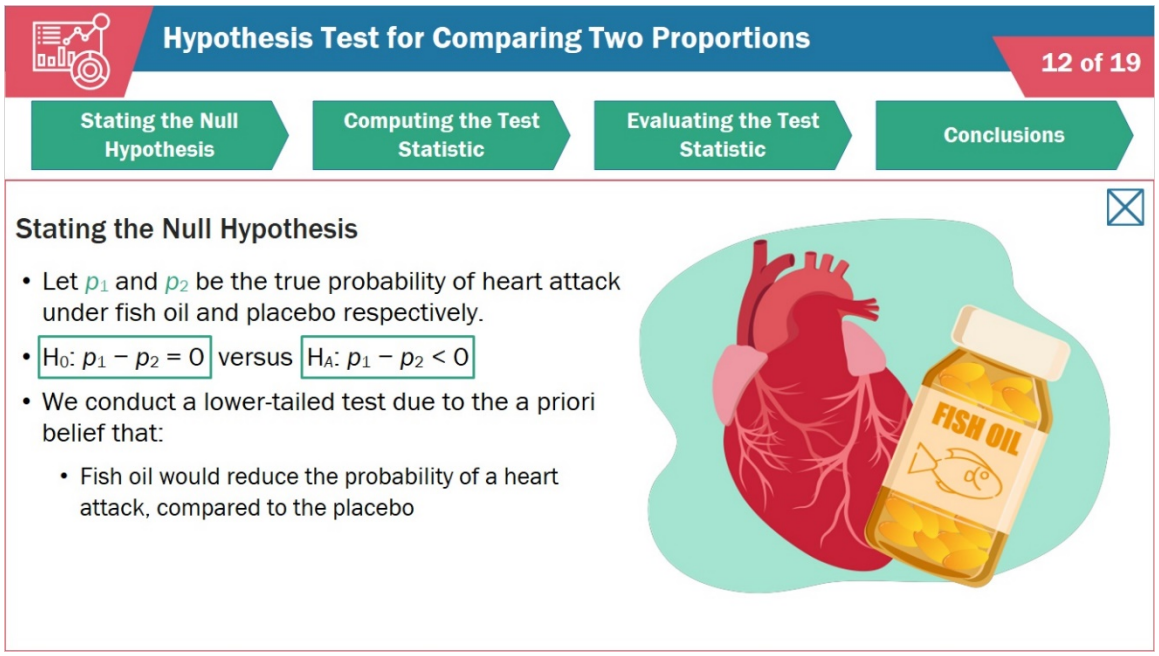
Slide 12: Hypothesis Test for Comparing Two Proportions



We will conduct a hypothesis test to compare the probability of having a heart attack under the fish oil treatment compared to placebo. Once again, we will specify our hypotheses, compute our test statistic, evaluate the test statistic against the reference distribution, and give our conclusions.

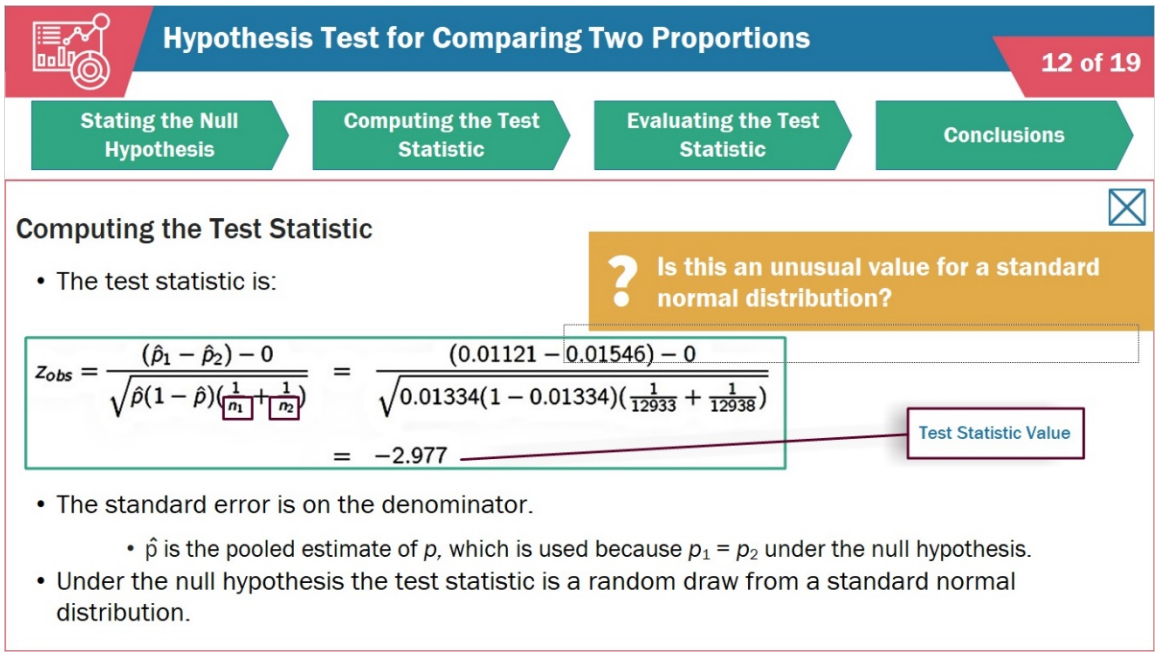
Click each tab in turn to learn more. When you are ready, click next to continue.

Tab 1: Stating the Null Hypothesis



We start off by defining our notation. We let p1 be the probability of a heart attack under fish oil, and similarly p2 is the probability under placebo. Under the null hypothesis, the two probabilities are the same, or their difference is equal to 0. The alternative hypothesis is that p1 is lower than p2, or that p1 – p2 is less than 0. Remember, we were told in the example that it was hypothesised prior to doing the experiment that taking the fish oil would reduce the probability of a heart attack, compared to the placebo group. This *a priori* belief is the reason that we are doing a lower tailed test (as opposed to a two-sided test).

Tab 2: Computing the Test Statistic

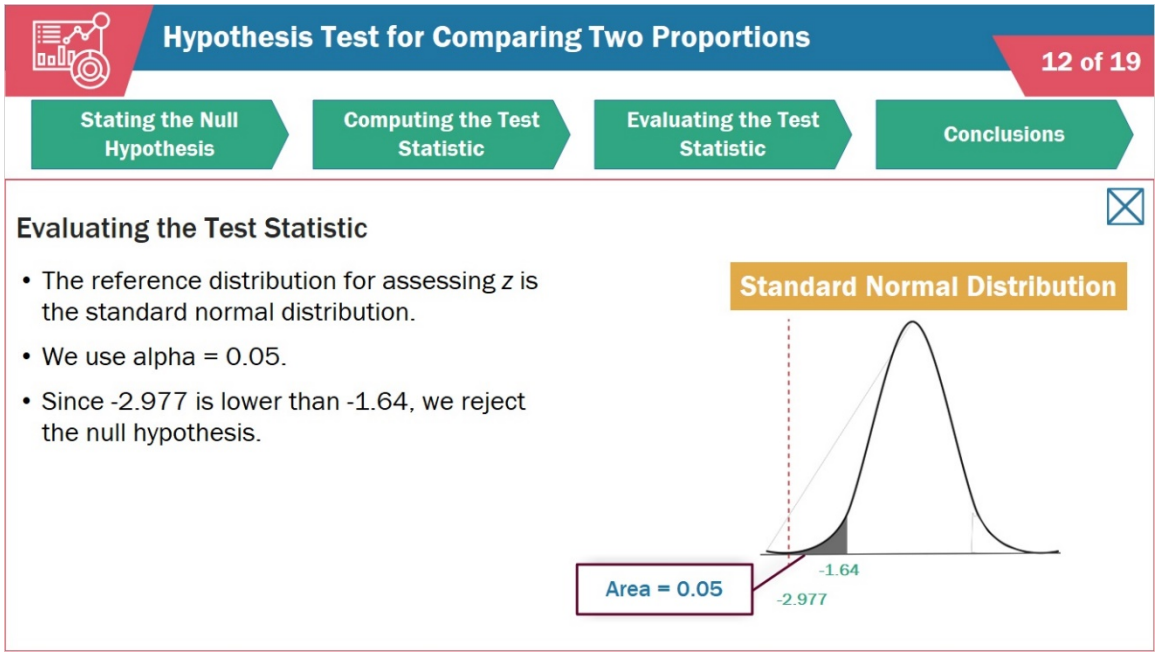


This is the formula for the test statistic. On the numerator, we have the difference between p1_hat and p2_hat, which are the estimates of each probability from the data, minus 0, the difference under the null hypothesis. The standard error is on the denominator, where p_hat is the pooled estimate of p, assuming that there is no difference between the two groups (which is what is assumed under the null hypothesis). n1 and n2 are the number of participants in the fish oil and placebo groups respectively.

The test statistics value works out to be -2.977.

If the null hypothesis is true, the test statistic is a random draw from a standard normal distribution. We have observed -2.977, and the question now is whether or not this is an unusual value for a standard normal distribution.

Tab 3: **Evaluating the Test Statistic**

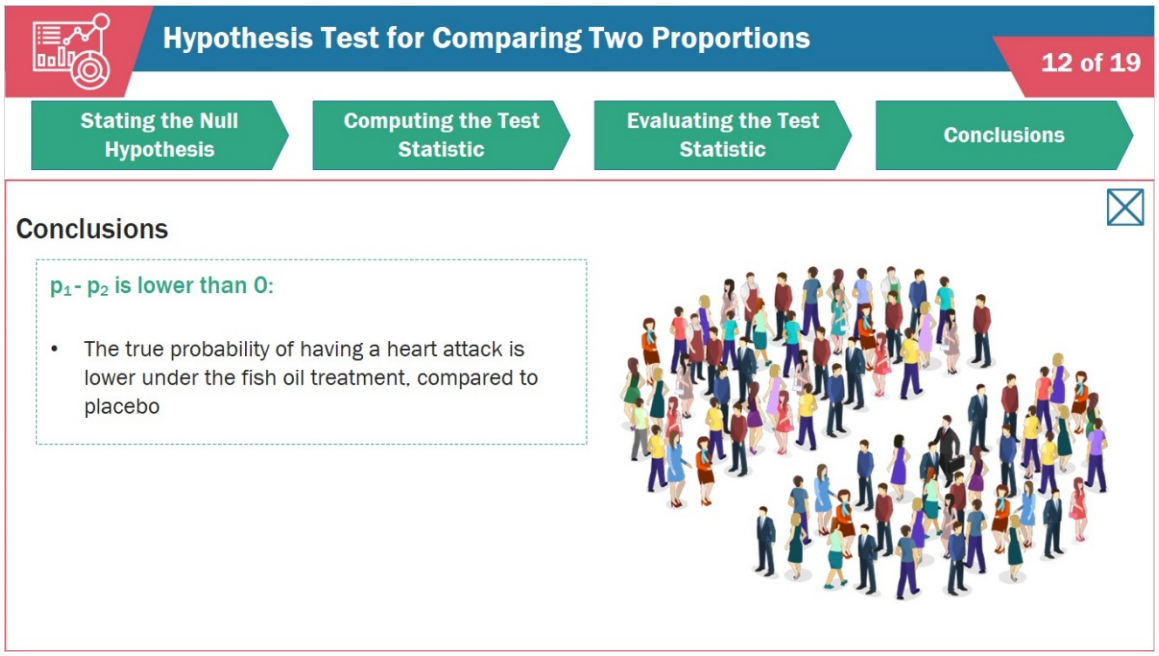


Let's examine the standard normal distribution. What you can see here, is the probability density function for the standard normal distribution. The area under the curve to the left of -1.64 has been marked out and this shaded grey area is equal to 0.05 or 5% of the total area under the curve. We say that the critical value for alpha = 0.05 is equal to -1.64. In this case, any value lower than -1.64 is said to be unusual for the distribution, while anywhere else along the axis to the right of -1.64 is considered 'typical' for the distribution.

We have observed -2.977, which is considered unusual for this distribution and not consistent with the null hypothesis.

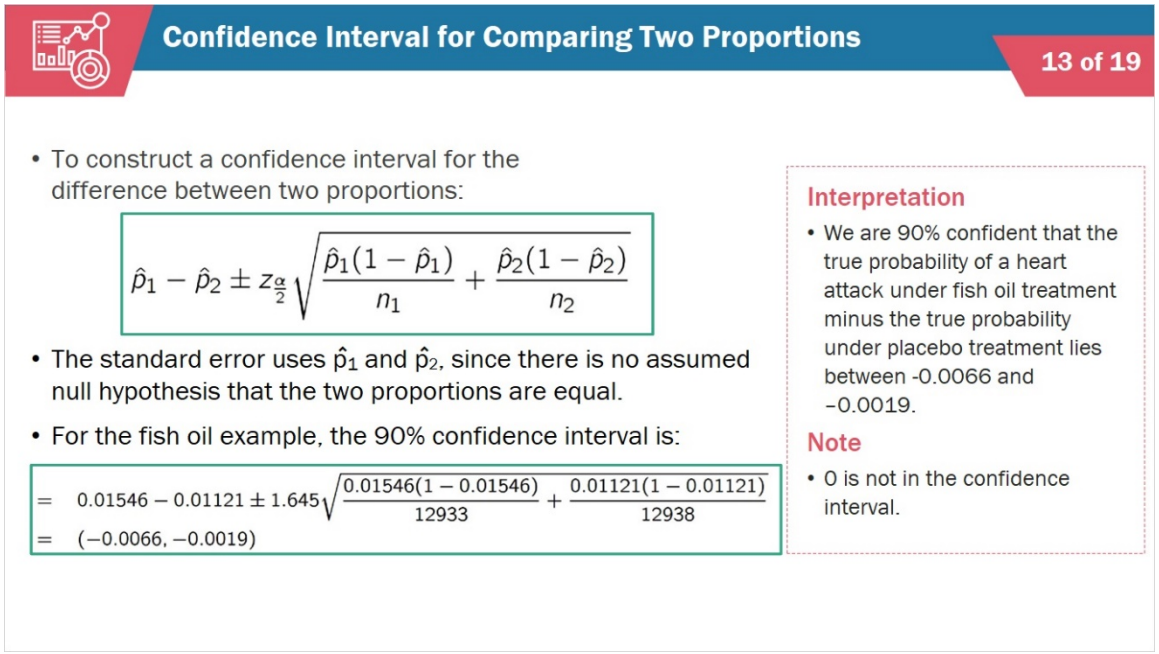
Since -2.977 is lower than -1.64, we reject the null hypothesis.

Tab 4: Conclusions



We conclude that p1 – p2 is lower than 0, meaning that the true probability of having a heart attack is lower under the fish oil treatment, compared to placebo.

Slide 13: Confidence Interval for Comparing Two Proportions



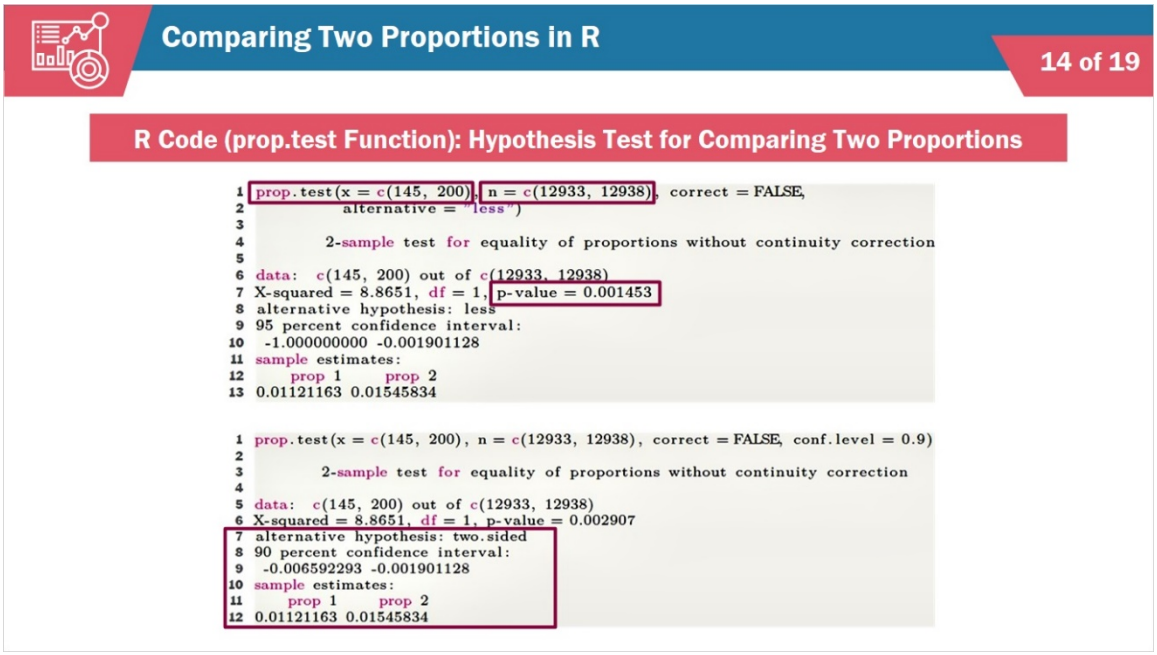
Let's now construct a confidence interval for p1 – p2. The structure of the confidence interval is as before: estimate plus or minus the critical value, times the standard error. You'll see that this time we use p1_hat and p2_hat individually in the standard error, as opposed to p_hat the pooled estimate that we used in the hypothesis test. This is because in the hypothesis test, we assume the null hypothesis is true and if p1 equals p2, the best estimate of it is the pooled estimate. For a confidence interval, there is no

null hypothesis, and so we use p1_hat and p2_hat separately and make no assumption that p1 and p2 are equal.

Going back to the fish example, we fill in the values: p1_hat = 0.01546, p2_hat = 0.01121, the critical value is 1.645, which is for alpha equal to 0.1/2 = 0.05 for a 90% confidence interval, and n1 and n2 are 12933 and 12938 respectively. Working out the values, the 90% confidence interval goes from -0.0066 to -0.0019. What is the interpretation of this confidence interval?

We are 90% confident that the true probability of a heart attack under the fish oil treatment minus the true probability under the placebo treatment lies by between -0.0066 and -0.0019. We note that 0 is not in the confidence interval.

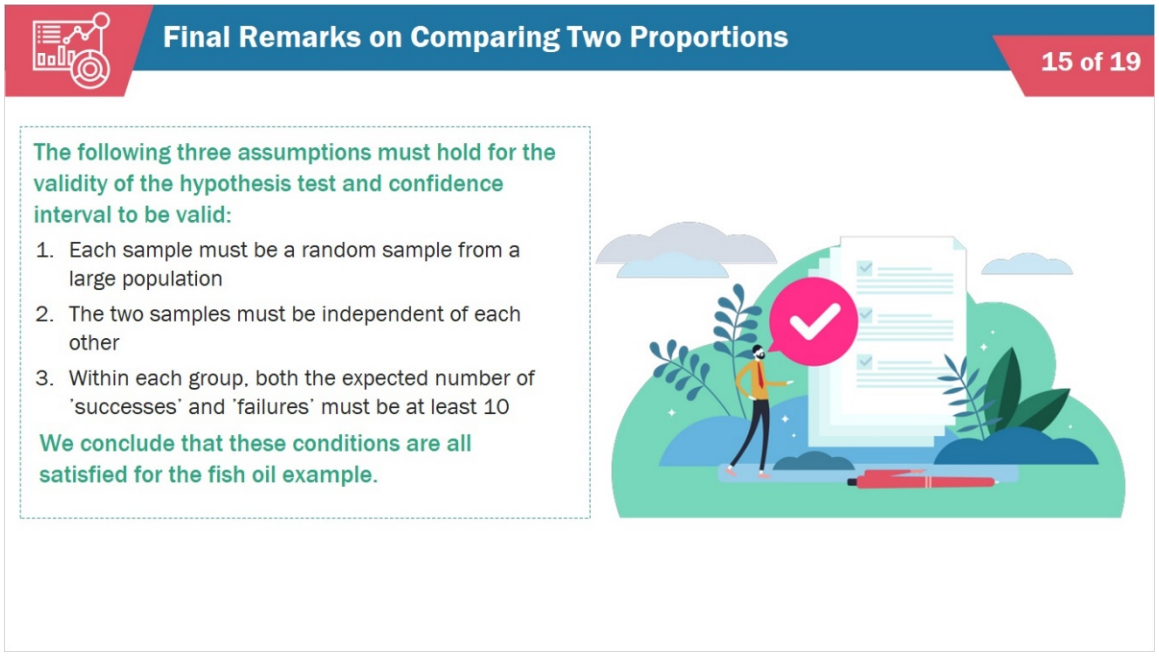
Slide 14: Comparing Two Proportions in R



We can use R software to perform a hypothesis test for comparing two proportions. We use the prop.test() function. The x vector shows the number of heart attacks under the two treatments, while the n vector shows the number of participants in each group. In the first segment of code, a lower-tailed test is performed. The p-value for the test is 0.001453; this is the probability of observing this test statistic or more extreme, conditioning on the null hypothesis being true. Since this is less than alpha = 0.05, we reject the null hypothesis and conclude that the true probability of having a heart attack is lower under the fish oil treatment, compared to placebo.

The second example uses a two-sided alternative, only so that we can see the full confidence interval.

Slide 15: Final Remarks on Comparing Two Proportions



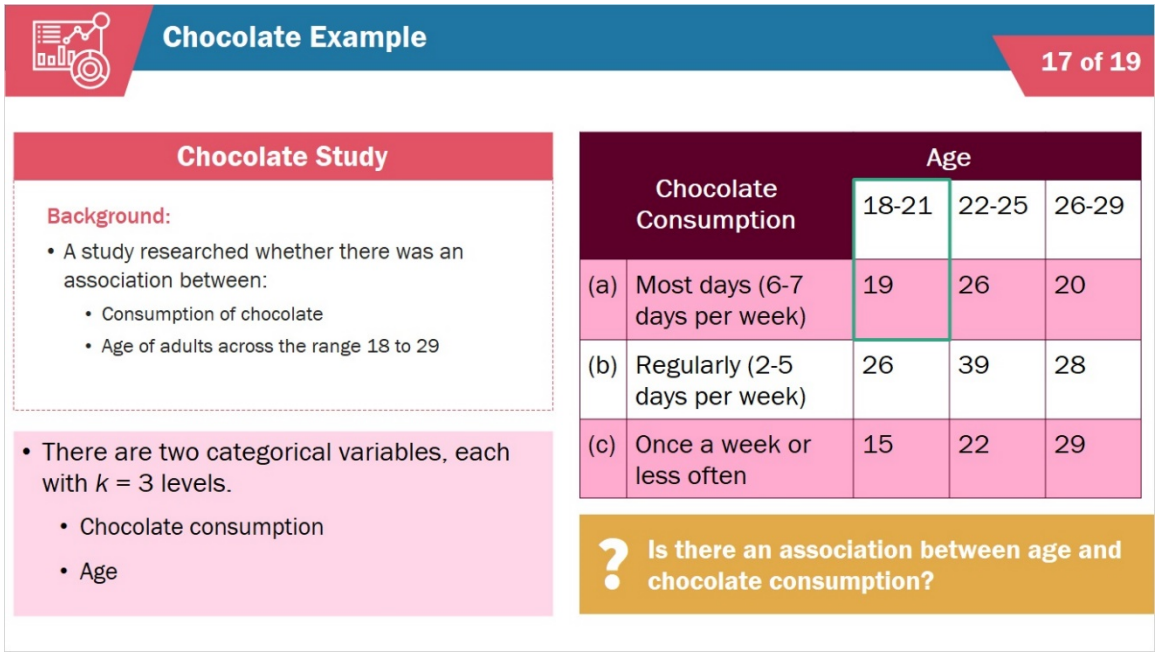
The validity of the hypothesis test and confidence interval for comparing two proportions relies on three assumptions being valid. First, we need each sample to be a simple random sample from a large population. Second, we need the two samples to be independent of each other. Third, we need both the expected number of successes and failures to be at least 10.

These conditions are all satisfied for the fish oil example.

Slide 16: Section 3: Contingency Tables and Chi-square Tests

Up until now, we have analysed binary variables, that is, categorical data variables with just two levels and we assessed the probability of a 'success'. We will now explore how to analyse categorical data variables with more than two levels.

Slide 17: Chocolate Example

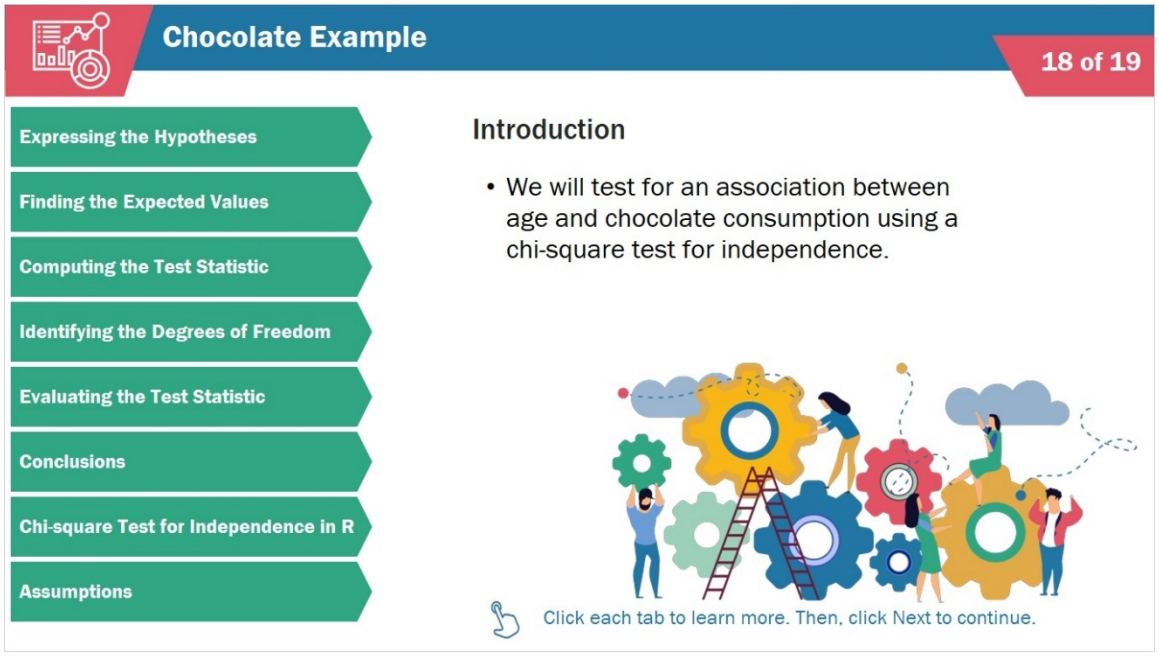


We'll start with a motivating example. A study was carried out to see if there was an association between consumption of chocolate and the age of adults across the range 18 to 29. Each participant was categorised by their age: 18-21, 22-25 or 26-29, and by their chocolate consumption: (a) Most days (6-7 days per week), (b) Regularly (2-5 days per week) or (c) Once a week or less often.

The data from the study is displayed here in a two-way contingency. For example, there were 19 participants in the 18-21 age category that eat chocolate most days.

Here we have two categorical variables: chocolate consumption and age, each with $k = 3$ levels. We want to know if there is an association between age and chocolate consumption.

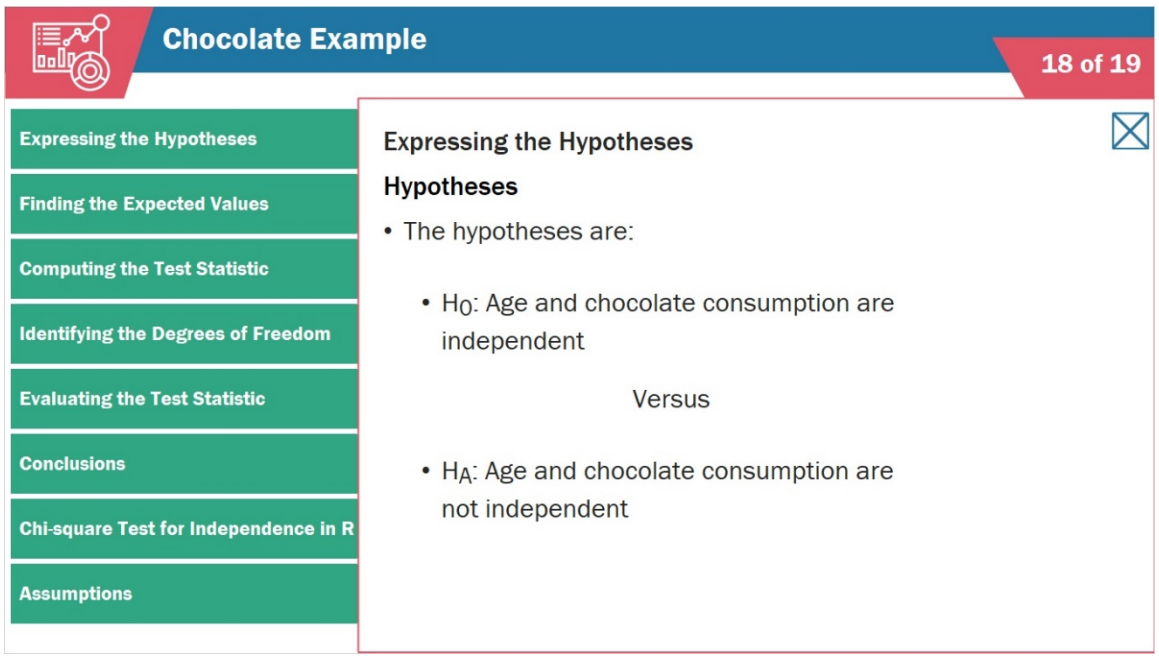
Slide 18: Chocolate Example



We will test for an association between age and chocolate consumption using a chi-square test for independence. We want to know, for example, if as age increases does chocolate consumption change. We will once again specify our hypotheses, compute our test statistic, evaluate the test statistic against the reference distribution, and give our conclusions.

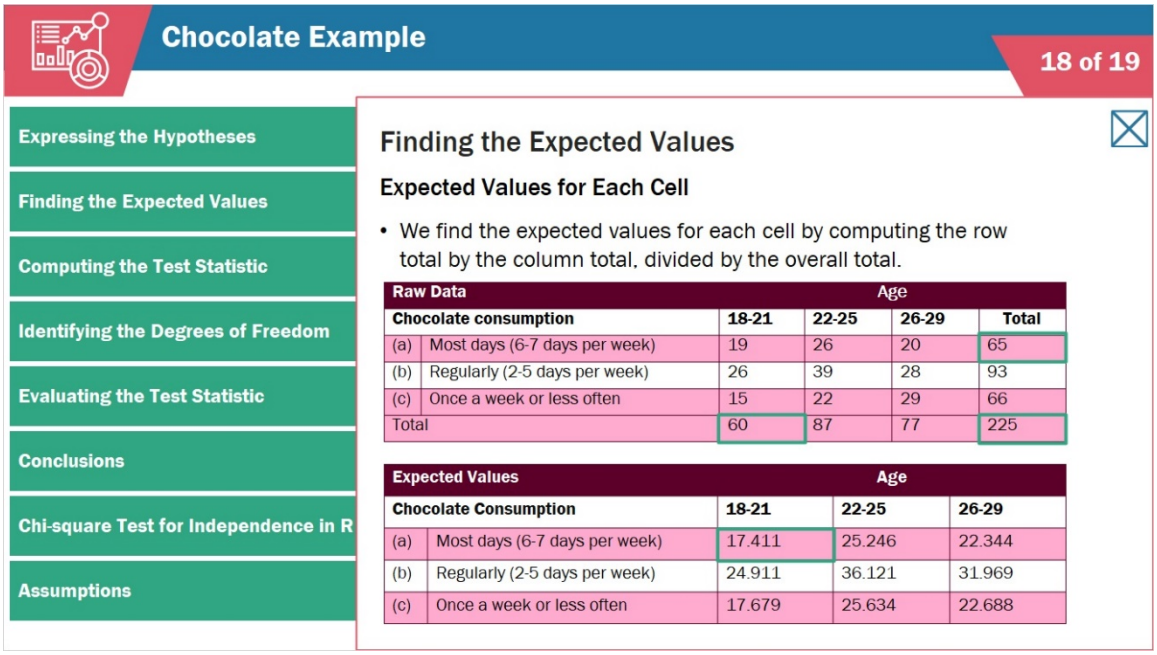
Click each tab to learn more. When you are ready, click next to continue.

Tab 1: Expressing the Hypotheses



Let's start by expressing the hypotheses. The null hypothesis is that age and chocolate consumption are independent of each other, and the alternative is that they are not independent.

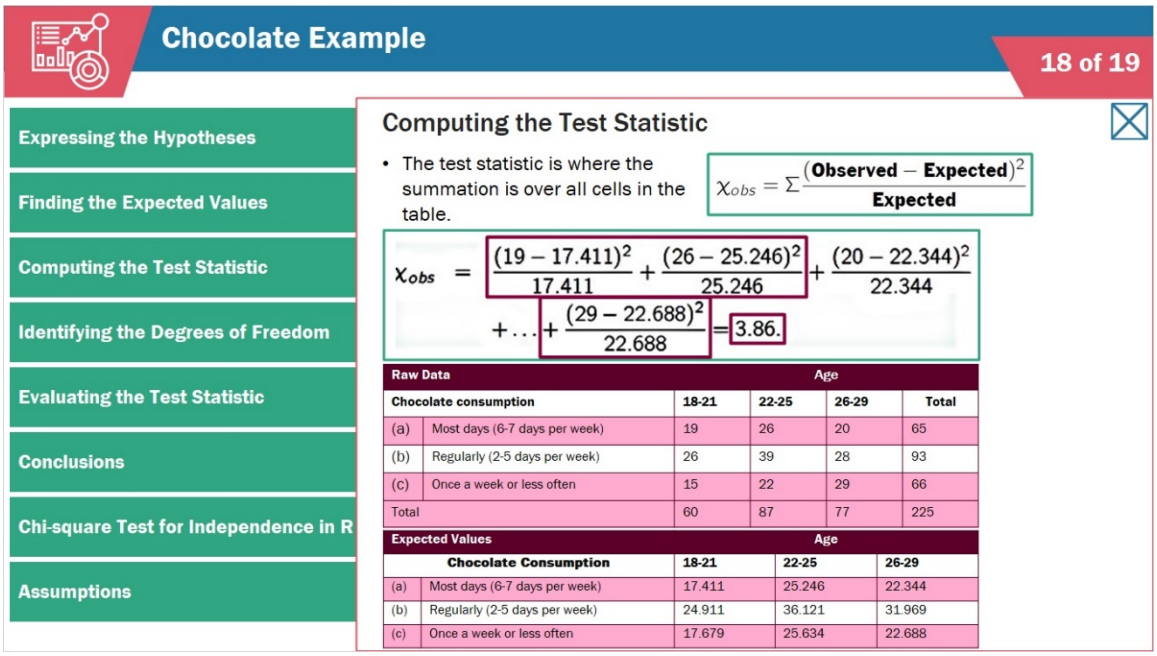
Tab 2: Finding the Expected Values



Before we can compute our test statistic, we need to find expected values for each cell, whereby expected value we mean the expected value assuming that the null hypothesis is true. We do this by multiplying the row total by the column total and dividing by the overall total, where each total is from the raw data table. For example, to find the expected value in the first cell we multiply 65 by 60 and divide by 225 giving 17.411.

Here are the expected values for each cell. Each value is computed in a similar way from the raw data table: row total by column total divided by overall total.

Tab 3: Computing the Test Statistic

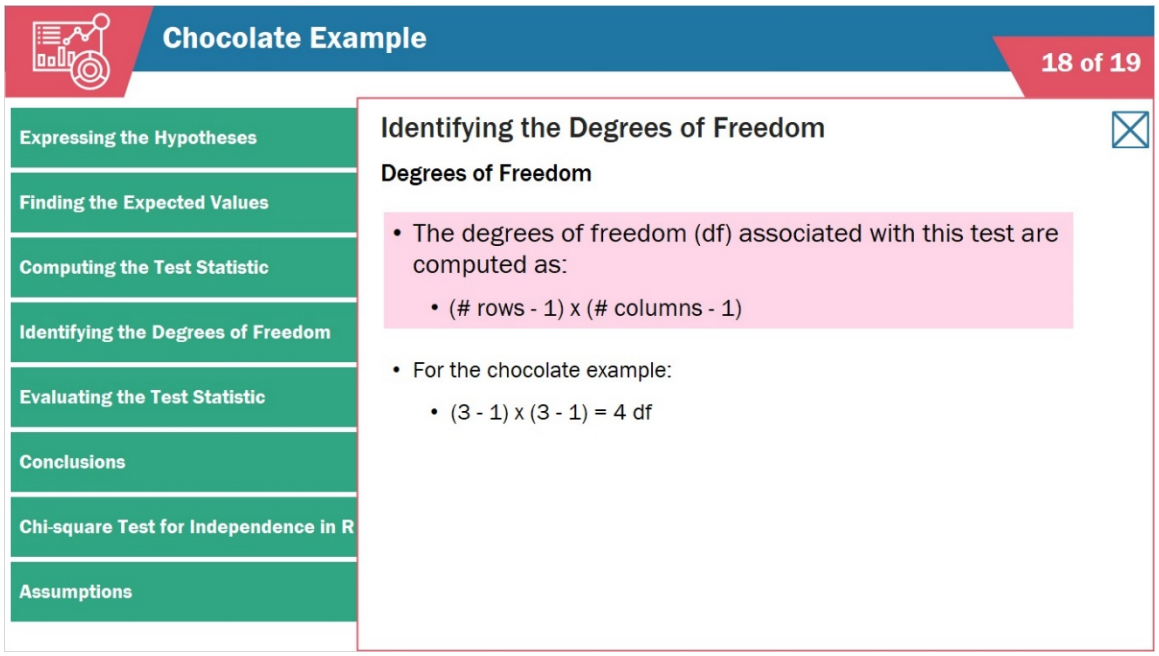


The test statistic is calculated by computing observed value minus expected value all to be squared, divided by expected for each cell, and then summing over all cells. Observed counts are the raw observed data counts. 'Expected values' are the values that would be expected under the null hypothesis, that is, assuming that the null hypothesis is true.

Going back to our example; in the first cell the observed cell count is 19 and the expected value is 17.411. We compute 19 – 17.411, square it, and divide by 17.411. For the second cell, it is 26 – 25.246 squared, divided by 25.246, and we continue on for each cell down to the last cell which is 29 – 22.688 squared, divided by 22.688. When we add up the values for each cell, we get 3.86.

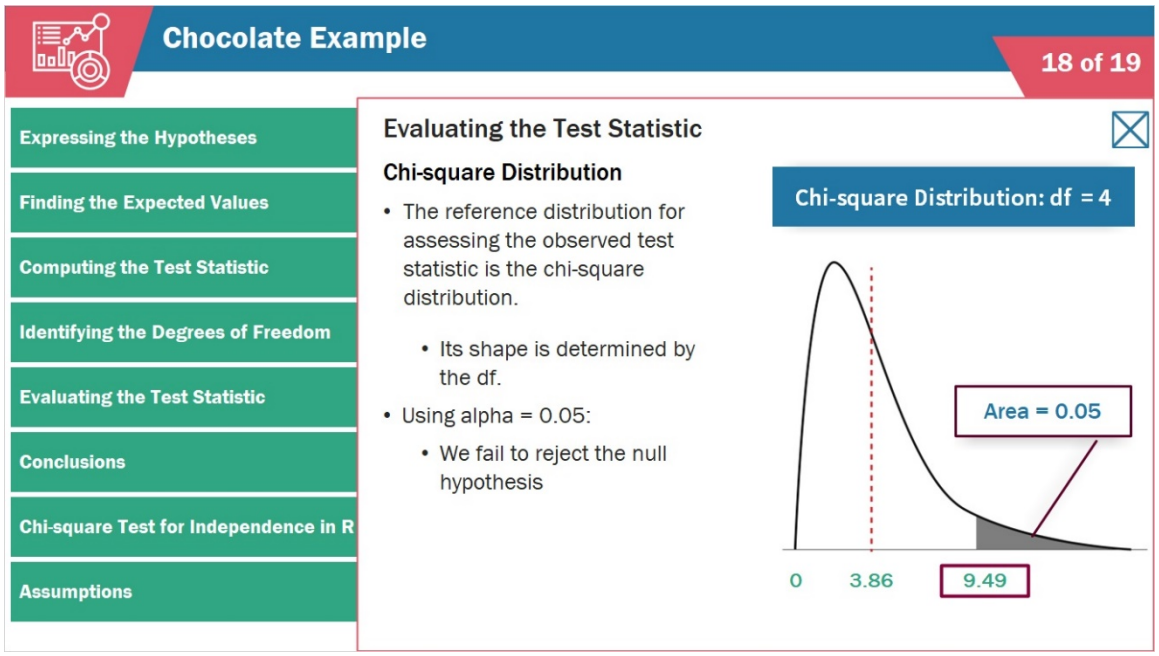
Remember that the expected values are what is expected under the assumption that the null hypothesis is true. The numerator of each cell calculation in the test statistic is effectively measuring how far each observed cell count is from what is expected under the null hypothesis. If the data are consistent with the null hypothesis, we would expect a small test statistic value, because observed and expected will be close together; and if the data depart from what is expected under the null hypothesis, then the test statistic will be large.

Tab 4: Identifying the Degrees of Freedom



Before we can evaluate the test statistic, we must identify the degrees of freedom. The degrees of freedom (df) associated with this test are computed as the # rows – 1 times the # columns – 1. In this example, there are 3 – 1 times 3-1 equal to 2 by 2 equals to 4 degrees of freedom.

Tab 5: Evaluating the Test Statistic

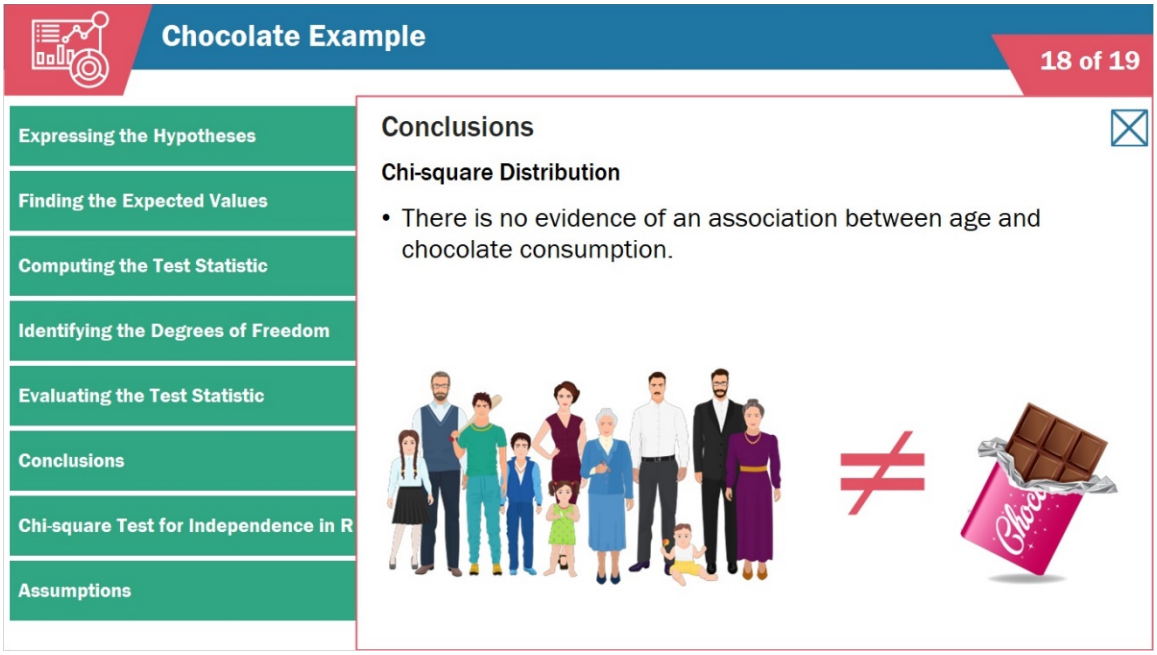


The reference distribution for assessing the observed test statistic is the chi-square distribution, the shape of which is determine by the degrees of freedom. In our example, we have four degrees of freedom and here we can see the shape of the chi-square distribution with 4 degrees of freedom. The value 9.49 marked in on the graph is the

critical value such that the area under the curve to the right of it is equal to 0.05. So, the area in grey is 5% of the total area under the curve.

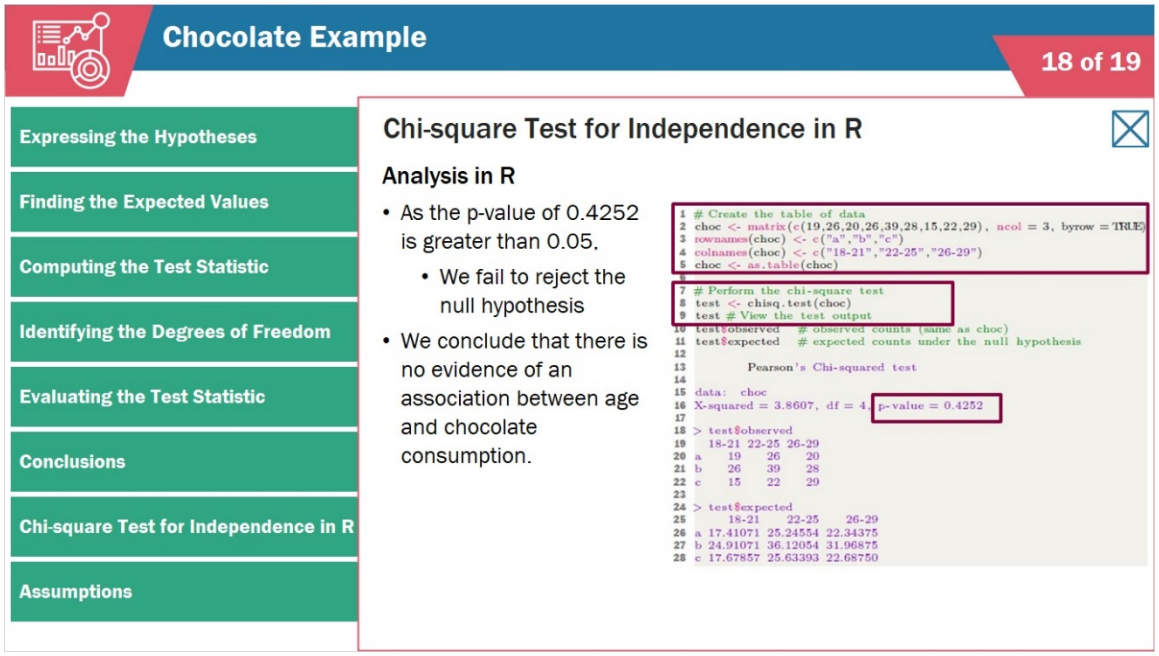
We use this cut-off of 9.49 to determine if our observed test statistic is extreme or not, anything larger than the cut-off is considered extreme for the distribution. Since our test statistic is equal to 3.86, which is less than 9.49, it is not extreme for this distribution, using alpha = 0.05, we fail to reject the null hypothesis.

Tab 6: **Conclusions**



Tab 7: Chi-square Test for Independence in R

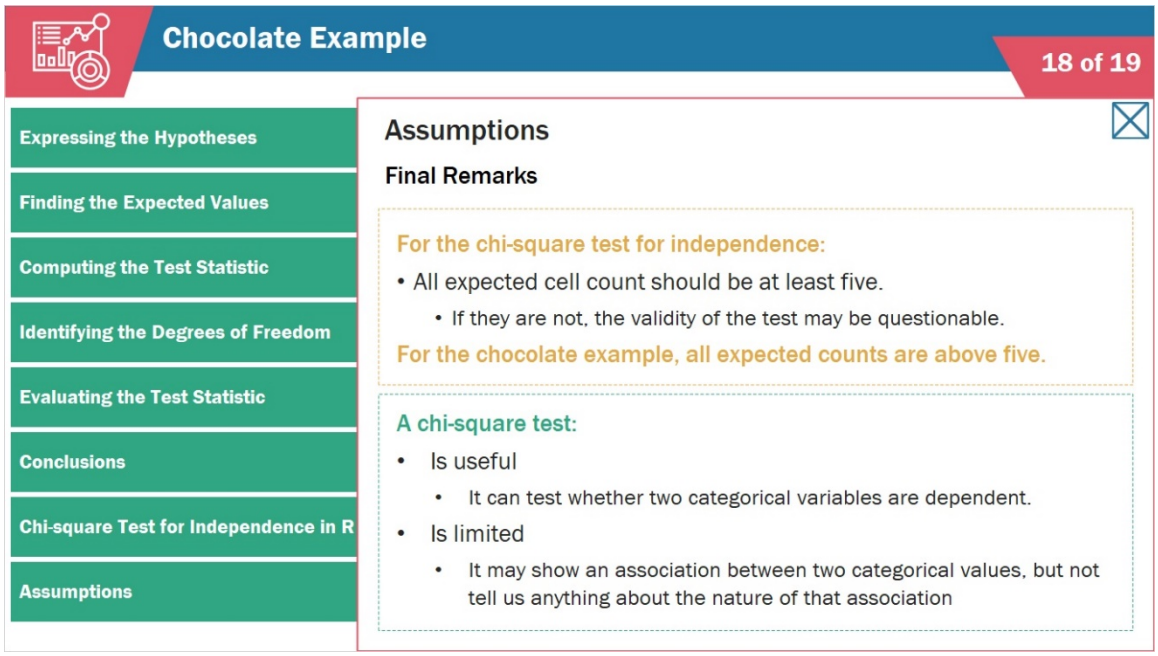
Slide 19:



We can do this analysis in R. First, we set up a table in R containing the data. We can label the rows and columns appropriately using the row names and col names functions. The first four lines of code are for reading in the data. We can then carry out a chi-square test for independence using the chisq.test function.

In the output, we can see the value of our test statistic equals 3.8607 with p-value 0.4252. The p-value is the probability of observing this test statistic or a more extreme one, assuming that the null hypothesis is true. Since the p-value 0.4252 is greater than 0.05, we fail to reject the null hypothesis, and conclude that we have no evidence of an association between age and chocolate consumption.

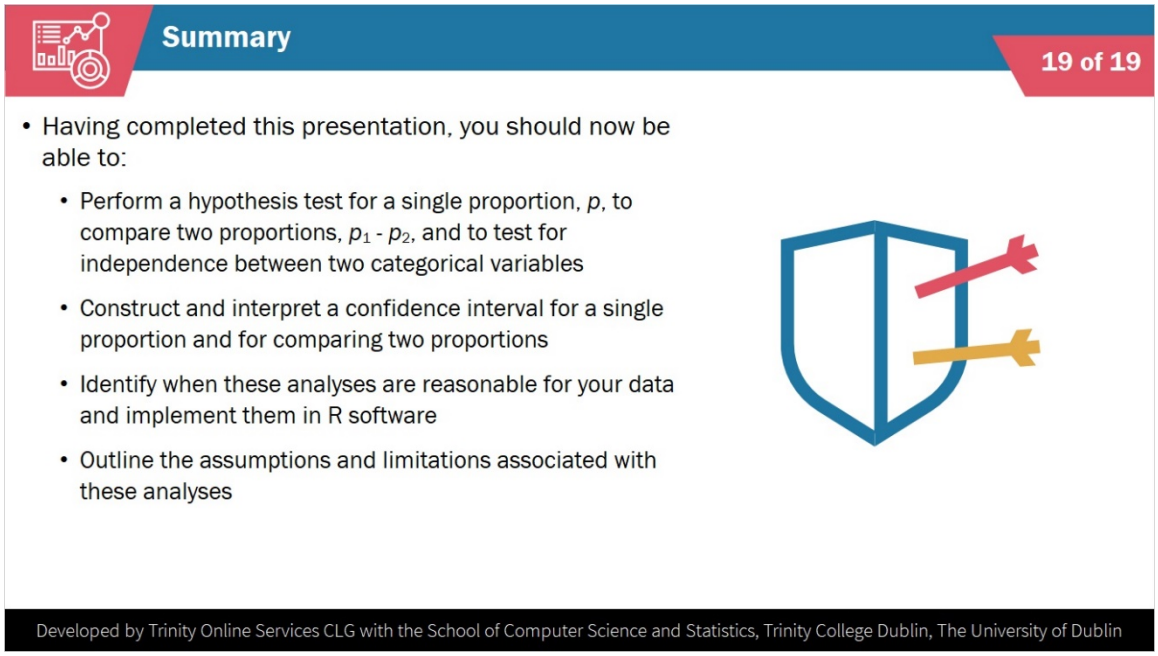
Tab 1: **Assumptions**



For the chi-square test for independence, all expected cell counts should be at least 5, if they are not, the validity of the test may be questionable. In the chocolate example, the lowest expected cell count is 17.411 and so this condition is satisfied.

We have seen how a chi-square test can test whether or not two categorical variables are independent; however, there is a limitation to this test. If the result of the test identifies an association between two categorical variables, it will not tell us anything about the nature of that association. But it is a useful starting point for jointly analysing two categorical variables.

Slide 20: Summary



Having completed this presentation, you should now be able to:

- Perform a hypothesis test for a single proportion, *p*, to compare two proportions, *p1 – p2*, and to test for independence between two categorical variables.

- Construct and interpret a confidence interval for a single proportion and for comparing two proportions.

- Identify when these analyses are reasonable for your data and implement them in R software.

- Outline the assumptions and limitations associated with these analyses.