Association Between Two Categorical Variables: Chi-Squared Test

Before you Watch

This video will describe the concept of comparing two categorical variables and formally testing whether responses to one variable are associated with responses from another variable. This builds on the Random Variables video where variable types are described, as well as the Visual Displays- Two Variables video where exploratory visual comparisons are described.

Reminder: a categorical variables is one where the possible responses fall into categories. For example, smoking status could be classified as a past or a current smoker.

The Video Content

We are interested in whether the response to one categorical variable is associated with a particular response to another categorical variable.
For example, is there an association between gender and having diabetes?

Is there a relationship between how much students exercise (low, medium, or high) and how much their parents exercise (low, medium, or high)? Is there a relationship between whether one has ever suffered from tonsillitis and whether they snore?

In all these situations, we are considering whether, across a large number of assessed people or items, we see patterns between the responses to the two categorical variables, on average.

The first step would be to present the data in a table. Let’s consider the pair of variables tonsillitis (with possible responses of either yes, or no) and snoring (also having responses either yes, or no) to see if they are related.

In one study, there was information collected on 50 children. Here are the data for the first ten children in the study. For each child, two responses were recorded, one relating to the presence of tonsillitis and one identifying whether they snored or not.

So for each child, we have information on whether or not they snore and whether or not they have tonsillitis.

This child snored and had tonsillitis. The second child in the study snored and did not have tonsillitis. We record these data for all 50 children in the study.

We can summarise the information for all 50 children using a cross tabulation as displayed here.

This is a table showing the information for both variables, Snoring and Tonsillitis, simultaneously and shows how many times each combination of responses appears.

Of the 50 children, 30 had suffered from tonsillitis during their childhood while the remaining 20 children had not suffered from this condition. Of the 50 children, 18 snored while the remaining 32 did not. So how can we use this summarised information to decide if there is indeed an association between having tonsillitis and snoring?

Is it true that those with tonsillitis are more likely to snore? First, let's consider those with tonsilitis. Of the 30 children with tonsillitis,
15 did not snore and 15 did snore. Of the 20 children without tonsillitis, 17 did not snore and 3 did snore.

Since we have different numbers of children in our study with and without tonsillitis, we consider the relative frequency through row percentages.
Out of the 30 children with tonsillitis, 15 children (or 50%) snore. So 50% of those with tonsillitis do snore. What percentage of those without tonsillitis do snore? It's 3 over 20 or 15%.

It is now much easier to compare the prevalence of snoring between those with and without tonsillitis. 50% of those with tonsillitis do snore.
So it appears that those with tonsillitis may be MORE likely to snore, or that tonsillitis is related to snoring.

Is the relationship statistically significant? If we took another sample of 50 we wouldn't necessarily get exactly the same numbers and percentages. The percentage of children with tonsillitis who snore may not be exactly 50% for example. We refer to this as sampling variation.

We perform a formal test, or hypothesis test, to check whether there is a statistically significant relationship between tonsillitis and snoring. When we are interested in determining whether two categorical variables are related we may use the chi-squared test.

The null hypothesis is stated as ‘no relationship between tonsillitis and snoring’
The chi-squared test compares the observed numbers with what we would expect to see if there was no relationship between tonsillitis and snoring.

But what would we expect to see if there was no relationship? These are referred to as the expected counts if the variables are not related
So we need to work out the expected count in each cell of the table.

Let’s start with the cell corresponding to those with tonsillitis and who snore. We observed 15 people with tonsillitis who snore. If snoring and tonsillitis are not related, how many would we expect to see here, and here?

If response to Snoring is independent of tonsillitis then the proportion of those with tonsillitis, which is 30 out of 50 should be reflected in the proportion who say yes to snoring and in the proportion who say no to snoring. So we expect these values to be 10.8 and 19.2. We repeat this for the other row of values and get 12.8 and 7.2

The chi-squared test evaluates the probability of observing such differences in observed and expected counts if there is in fact no association between the variables. This is a p-value, described in the video about hypothesis testing.

In this case the p value would be 0.01 indicating a low chance of seeing such differences in our data in there was no relationship. At a 5% significance level we would reject the null hypothesis and conclude there is a statistically significant association between tonsillitis and snoring.

Now What?

This video introduced how to determine if two categorical variables are related to each other, and whether that relationship is statistically significant or not. You can look to further develop your understanding of this concept by looking to the Other Links listed below. Alternatively, how to determine if two continuous variables are related is covered in Association Between Two Continuous Variables: Correlation.

But, when am I going to use this?

Statistics is essential in the study of systems of situations where there is an unpredictable random element. This includes a huge number of situations, such as any system that involves living things, which always have a degree of unpredictability. In fact, any studies that involve people involve statistics: for example, medical processes, education and economics. Other areas statistics can be applied to include quality control, stars, nature, or how wear and tear affects machinery.

Other Links

A YouTube video made by the Statistics Learning Centre in NZ gives an overview of data types using animation that is easy to understand and fun. Note that numerical data types has been given alternative names. Interval and ratio data are used instead of discrete and continuous. It is a slightly different approach but the same principle.

https://www.youtube.com/watch?v=hZxnzfnt5v8

If you really want to understand types of data, along with appropriate statistics and graphs, you can learn on our new Snack-size course. Takes about an hour, and lots of fun!

http://www.statslc.com/snack/