Correlation and Regression: Correlation

Before you Watch

This video will describe the concept of comparing two continuous numerical 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 numerical variable is one where the possible responses are numbers which reflect the quantity, or amount, of something. Take for example height, in centimeters, or weight, in kilograms.

The Video Content

This video explains how to determine whether or not the response to one numerical variable is related to, or associated with, the response to another numerical variable. We term this as whether the two numerical variables are related.

For example, is there a relationship between amount of time spent exercising per week and resting blood pressure? We may be interested in the relationship between the two continuous variables blood pressure and body mass index. Both variables are measured for each individual.

Or whether the amount of time studying is related to the result on the final exam. Is the length of stay in hospital related to the age of the patient? Do older people, for example, spend longer in hospital, on average compared to younger patients? These situations are all assessing the relationship or association between two numerical variables. In many cases it is important to identify which is the independent (or predictor) variable and which is the dependent (or outcome) variable for each relationship.

Let's consider the relationship between Weight (in kg) and Cholesterol (in milligrams per decilitre). We may be interested in whether cholesterol may be predicted by weight, in which case weight is the Independent variable and cholesterol is the Dependent variable.

Scatterplots enable a visual assessment of relationships between two numerical variables. They are a graphical display with a horizontal line (known as the x-axis) and a vertical line (known as the y-axis),

The independent variable, Weight, is displayed on the horizontal or x-axis. And the dependent variable, Cholesterol, is displayed on the vertical or Y-axis.

Next we display the cholesterol and weight data on the scatterplot. For our example, we have data on 20 patients. Let’s consider the first five patients.

For each patient, we have a measure of their weight and their cholesterol. To represent the five patients’ data, we will plot five points on the scatterplot.

We must retain the pairing of data. Take our first patient. Their weight is 81 kg and their Cholesterol is 205. So along the horizontal axis we go to 81, and along the vertical axis we go to 205, and where these intersect we place a dot.

We then repeat this process for each of our remaining 19 patients and end up with the following scatterplot of cholesterol against weight

From the graph, we notice that as weight increases, cholesterol tends to increase too. So those with lower weights are more likely to have lower cholesterol values and those with higher weights are more likely to have higher cholesterol values. We can describe this association, or relationship, as linear and POSITIVE.

We can describe the strength of the linear association numerically using a Correlation coefficient, which can be represented by the letter r .

The correlation coefficient value can range from minus one to positive one. Positive values suggest positive relationships like the one we displayed between Weight and Cholesterol.

While negative correlations represent negative relationships, that is that the response to one variable decreases on average while the response to the other variable increases.

If there does not appear to be any linear association, like in the third graph, this would be described as no association and the correlation coefficient would be close to zero.

We observe from the scatterplot an approximately linear association between weight and cholesterol since a straight line can represent the nature of the relationship. In this case we could draw a line of best fit through the points and end up with a line like this.

The closer the points lie to a straight line, the closer the correlation coefficient is to one (if positive relationship) or minus one (if negative relationship). Here we have a correlation coefficient of 0.85, which is close to positive one, suggesting a strong positive linear association. The closer the correlation to zero, the weaker the linear (or straight line) association.

This fitted line of best fit is called a regression line. This is covered in the next video on Regression. It is also possible for the relationship between variables to be non-linear, like in this graph.

Now What?

This video introduced how to determine if two numerical 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, you can learn how to determine if two categorical variables are related, which is covered in Association Between Two Categorical Variables: Chi-Squared Test.

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/