Australian Bureau of Statistics

Statistical Language - Correlation and Causation

Statistical Language

Correlation and Causation

Content on this page requires Adobe Flash Player to be viewed.

Download Adobe Flash player

Alternatively, read the transcripts, attached below, containing a text version of the information displayed in the Flash Animation.

This animation explains the concept of correlation and causation. If you are unable to access the video a Transcript (.doc 26kb) *has been provided. The animation requires* *Adobe Flash Player* *to run. The animation contains no audio.*

What are correlation and causation and how are they different?

Two or more variables considered to be related, in a statistical context, if their values change so that as the value of one variable increases or decreases so does the value of the other variable (although it may be in the opposite direction).

For example, for the two variables "hours worked" and "income earned" there is a relationship between the two if the increase in hours worked is associated with an increase in income earned. If we consider the two variables "price" and "purchasing power", as the price of goods increases a person's ability to buy these goods decreases (assuming a constant income).

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

Theoretically, the difference between the two types of relationships are easy to identify — an action or occurrence can *cause* another (e.g. smoking causes an increase in the risk of developing lung cancer), or it can *correlate* with another (e.g. smoking is correlated with alcoholism, but it does not cause alcoholism). In practice, however, it remains difficult to clearly establish cause and effect, compared with establishing correlation.

Why are correlation and causation important?

The objective of much research or scientific analysis is to identify the extent to which one variable relates to another variable. For example:

- Is there a relationship between a person's education level and their health?
- Is pet ownership associated with living longer?
- Did a company's marketing campaign increase their product sales?

These and other questions are exploring whether a correlation exists between the two variables, and if there is a correlation then this may guide further research into investigating whether one action causes the other. By understanding correlation and causality, it allows for policies and programs that aim to bring about a desired outcome to be better targeted.

How is correlation measured?

For two variables, a statistical correlation is measured by the use of a Correlation Coefficient, represented by the symbol (r), which is a single number that describes the degree of relationship between two variables.

The coefficient's numerical value ranges from +1.0 to –1.0, which provides an indication of the strength and direction of the relationship.

> If the correlation coefficient has a negative value (below 0) it indicates a negative relationship between the variables. This means that the variables move in opposite directions (ie when one increases the other decreases, or when one decreases the other increases).
>
> If the correlation coefficient has a positive value (above 0) it indicates a positive relationship between the variables meaning that both variables move in tandem, i.e. as one variable decreases the other also decreases, or when one variable increases the other also increases.
>
> Where the correlation coefficient is 0 this indicates there is no relationship between the variables (one variable can remain constant while the other increases or decreases).

While the correlation coefficient is a useful measure, it has its limitations:

> Correlation coefficients are usually associated with measuring a linear relationship.
> For example, if you compare hours worked and income earned for a tradesperson who charges an hourly rate for their work, there is a linear (or straight line) relationship since with each additional hour worked the income will increase by a consistent amount.
>
> If, however, the tradesperson charges based on an initial call out fee and an hourly fee which progressively decreases the longer the job goes for, the relationship between hours worked and income would be non-linear, where the correlation coefficient may be closer to 0.
>
> Care is needed when interpreting the value of 'r'. It is possible to find correlations between many variables, however the relationships can be due to other factors and have nothing to do with the two variables being considered.
> For example, sales of ice creams and the sales of sunscreen can increase and decrease across a year in a systematic manner, but it would be a relationship that would be due to the effects of the season (ie hotter weather sees an increase in people wearing sunscreen as well as eating ice cream) rather than due to any direct relationship between sales of sunscreen and ice

cream.

The correlation coefficient should not be used to say anything about cause and effect relationship. By examining the value of 'r', we may conclude that two variables are related, but that 'r' value does not tell us if one variable was the cause of the change in the other.

How can causation be established?

Causality is the area of statistics that is commonly misunderstood and misused by people in the mistaken belief that because the data shows a correlation that there is necessarily an underlying causal relationship .

The use of a controlled study is the most effective way of establishing causality between variables. In a controlled study, the sample or population is split in two, with both groups being comparable in almost every way. The two groups then receive different treatments, and the outcomes of each group are assessed.

For example, in medical research, one group may receive a placebo while the other group is given a new type of medication. If the two groups have noticeably different outcomes, the different experiences may have caused the different outcomes.

Due to ethical reasons, there are limits to the use of controlled studies; it would not be appropriate to use two comparable groups and have one of them undergo a harmful activity while the other does not. To overcome this situation, observational studies are often used to investigate correlation and causation for the population of interest. The studies can look at the groups' behaviours and outcomes and observe any changes over time.

The objective of these studies is to provide statistical information to add to the other sources of information that would be required for the process of establishing whether or not causality exists between two variables.

Further information

ABS:
1500.0 - A guide for using statistics for evidence based policy
Literacy Stats: Using ABS Statistics: Telling the right story

Return to Statistical Language Homepage

This page last updated 3 July 2013