Relationships Between Variables

Table of Contents

INTRODUCTION

If you have a large enough data set, you may be interested in doing more complex analyses looking at the relationships between different variables. This document presents some common relationships between variables: association, independence, correlation, and causation.

ASSOCIATION

If there is an association between two variables, then knowing something about one variable gives you some information about the other. There is no requirement for either variable to be numerical or categorical.

> Example 1: Cold beverage sales and sunglasses sales are associated; both tend to increase at the same time.
>
> Example 2: Sex and height are associated; men are taller on average than women.

An association does not mean that knowing one variable automatically tells you another. In Example 1, cold beverage sales and sunglasses sales will not always increase at the same time. In Example 2, some women are tall and some men are short.

Another term for association is 'dependence' because what we know about one variable depends on what we know about the other.

INDEPENDENCE

An independence between two variables means that knowing one variable tells you nothing about the other variable. Like association, there are no restrictions on the types of variables involved; the variables could be both numeric, both categorical, or one of each.

> Example 3: Coin flips and gas price changes are independent. Flipping a coin and getting tails doesn't tell you anything about whether gas prices will go up or down today. Likewise, knowing that gas prices are going down doesn't tell you anything about your next coin flip.

CORRELATION

Correlation is a special kind of association that only happens between two numeric variables. When two variables are correlated, one of two things happens

1. Positive correlation. The higher one variable is, the higher we can expect the other variable to be.
2. Negative correlation. The higher one variable is, the lower we can expect the other variable to be.

> Example 4: Hours of sunlight in a day, and average temperature in a day have a positive correlation. The longer the sun is up, the hotter we can expect the day to be.
>
> Example 5: Height and weight have a positive correlation. The taller someone is, the more you can expect them to weigh.
>
> Example 6: Time spent on a phone and time spent outdoors are negatively correlated. The more time people spend outdoors, the less time, on average, they spend on their phones.

Two variables may have weak or strong correlations in positive and negative directions. Example 5 may have a strong positive correlation, whereas it is possible that example 6 has a weak positive correlation.

Correlations, even strongly positive or strongly negative correlations, do not constitute “proof”. Instead, you can draw upon your data and analyses to state, for example, that “evidence indicates there is a strong relationship between average sunlight and temperature”. Read more below on causation.

CAUSATION

A causation is a special kind of association where changes in one variable actively cause changes in another variable.

There are no restrictions on the type of variables that can form a causation, but there are very strict requirements for something to be a causation. Statistical analysis alone cannot provide enough evidence for a causation; additional information about what the variables actually represent is needed. A good checklist of requirements is the Bradford Hill Criteria.

Associations and correlations work in both directions; if variable X is correlated with Y, then Y is also correlated with X. However, causation can only work in one direction; if X causes Y, then Y cannot cause X.

Example 4: Daylight causes average temperature. Sunlight warms the ground creating heat, which causes an increase in average temperature. Note that average temperature does not cause changes in sunlight; the sun will rise and fall regardless of what is happening on the ground. Also note that causations can be correlations (but do not need to be).

Often, if there is an association, then there is a common cause behind changes in both variables.

Example 1: Cold beverage sales and sunglasses sales do not cause each other; however, both products are useful in warm weather, so sales tend to increase at the same time.

In most cases, causation is extremely hard to prove and is beyond what this material will cover.