



Correlation and Causation

Review - 1

Two types of correlational study

When same items have values on two score variables, correlate the scores on one with the scores on the other

Measure degree of correlation in terms of Pearson coefficient r

Predict value on one variable from that on the other using the regression line: $y=ax+b$

When one nominal variable divides a population into two or more sub-populations, compare the two (or more) populations on another (score) variable in terms of their central tendencies

If the means are different, predict the value on the score variable depending on the value of the nominal variable

Review - 2

In both types of correlational studies, one commonly makes inferences from a sample to an actual (total) population

Does what is found in the sample apply to the actual population?

Addressed in terms of *statistical significance*

Is the result in the sample one that would be *unlikely* to happen by chance if there *weren't* a correlation or a difference in the actual population?

The p value specifies the likelihood of the result in the sample happening by chance (in drawing the sample)

$p < .05$ indicates there is less than 5% chance of the result happening by chance

Clicker Question

A study based on a sample of 100 UCSD students reported a difference in interest in partying between men and women ($p < .01$)

This result is not reliable because of the small sample size

This result is not reliable because of the small p-value
There is less than 1 in 100 likelihood that there is a difference in the actual population

There is less than 1 in 100 likelihood that the difference in the sample is due to chance

Review - 3

In testing a claim about differences in the means of two sub-populations, one tests the null-hypothesis

There is no difference in the means

The strategy is to try to reject the null hypothesis using the results in the sample

If the difference in means in the sample is statistically significant (at a chosen level), one infers that the null hypothesis is false

Therefore, **the means differ in the real populations**

If the differences in means in the sample are not statistically significant (at the chosen level), one cannot reject the null hypothesis

Whatever differences there might be, they will not have been detected.

5

Clicker Question

If the attempt to find a difference in means based on a sample is reported to be non-significant, that means

The probability that the null hypothesis was true was greater than 5%

The probability that the null hypothesis was true was less than 5%

There is no difference between the means in the actual population

The result is not important

Review -3 *!*

No significant difference does not mean there is no difference

There may well be a difference, but one that has not been detected given the tests employed

All we can say is that we have not detected any difference

Compare (better, contrast)

We have not found the person who killed the Prime Minister

No one killed the Prime Minister

Caught Between Two Errors

Type I error (over confidence): Thinking there is a difference between means when there is none

Use higher significance levels: instead of requiring only $p < .05$, require $p < .01$ or even $p < .001$

Type II error (humility): Thinking there is no difference between means when there is one

Use a larger sample, which has a greater chance of finding a significant difference if one is to be found

8

Two dangers - 2

	H_0 is true	H_0 is false
Did not reject H_0	Correct failure to reject	Type II error (β)
Did reject H_0	Type I error (α)	Correct rejection

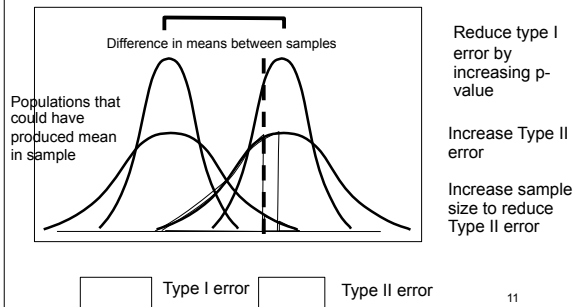
9

α and β levels

- α -level is the probability of rejecting the null hypothesis when it is true
 - Statistical significance and p-value
- β -level is the probability of failing to reject the null hypothesis when it is false
 - $(1 - \beta)$ is probability that the researcher will correctly reject the null when the null is indeed false
 - The statistical *power* of the test

10

Two types of error - 2



11

Clicker Question

Under what conditions should one focus more on reducing type II errors than type I errors?

- When it is critical not to claim a difference when there isn't one
- One should always be more concerned with type I errors
- When it is critical not to miss a difference when there is one
- When there is little worry about being wrong

Clicker Question

In which type of situation would you most likely expect that a Type II error has been committed?

When the difference between means in a small or moderate-sized sample is not found to be statistically significant

When an extremely large sample has been used

When the difference between means in a sample has been found to be significant ($p < .01$)

When the difference between means in an extremely large sample is not found to be statistically significant

Clicker Question

To reduce the likelihood of a Type II error, one should

Always insist on using p-values $< .01$

Not worry about the p-value and just look at the differences produced in the sample

Use a large enough sample so that if there is a difference, it will produce a significant difference in the sample

Use a small sample since then if there is a significant difference, there is likely to be a large difference in the real population

Science without Error?

One can reduce the risk of type I and type II errors to whatever level one desires

If one is willing to use a large enough sample

But one cannot eliminate the risk of error

It is always possible that there is no difference in means despite obtaining a significant result in one's sample

It is always possible that there is a real difference in means, but the difference in the sample is not significant

This is one more example of how scientific knowledge remains fallible!

Clicker Question

Is the following a good argument for confirming a correlational claim based on a sample:

If there is a difference between means in the population, the result in the sample will be statistically significant ($p < .X$)

The result in the sample is statistically significant ($p < .X$)

∴ There is a difference between means in the population

Yes, the argument is valid

Yes, the argument is sound

No, the argument affirms the consequent

No, the argument denies the antecedent

The Logic of Correlational Research

To confirm or falsify a correlational claim based on a sample, we use *modus tollens*. The first premise in each case, though, is different

Confirming a correlational claim:

If there is *no* difference between means in the population, then there will *not* be a statistically significant ($p < ?$) difference in my sample

There is a statistically significant difference ($p < ?$) in means in my sample

∴ There is a difference between means in the population

We pick the level of significance in the first premise according to how great a risk of error in our conclusion we can accept

The Logic of Correlational Research - 2

Falsifying a correlational claim

If there is a *detectable* difference between means in the population, then there will be a statistically significant difference ($p < ?$) in my sample

There is no statistically significant difference ($p < ?$) in means in my sample

∴ There is no *detectable* difference between means in the population

The truth of the first premise depends upon using a large enough sample

NOTE: The conclusion refers to *DETECTABLE* differences

Quest for finding causes

When something happens, we ask "Why?" We want to know what caused the event

Why are we interested in causes?

Knowing the causes frequently provides understanding

Knowing causes empowers us to intervene

These two tend to go together

Why do these barrels produce better beer?

Learning the reason is more hops provides understanding

And a procedure for making better beer

How does HIV cause AIDS?

Knowing about protease inhibitors explains

And tells us a good place to intervene



What is a cause?



The roots of talk of causation is found in our doing something to produce an effect

We want to move a rock, so we push it

We want to see a friend so we walk to her apartment

We want to stay warm so we put on a jacket

Independent of our own action, a cause is something which brings about or increases the likelihood of an effect

The cause of the explosion was the spark from the generator

Correlation and Causation

A major reason people are interested in correlations is that they might be indicative of causation

Correlations *per se* only allow you to predict

The correlation of unprotected sex with having a baby nine months later allows someone who has unprotected sex to predict that they are more likely to have a baby nine months later

Causation tells you how to change the effect

Knowing that unprotected sex causes (increases the likelihood of) having a baby nine months later allows you to take action to have or not have a baby

Correlations Point to Causation

Statistical relations between variables that exceed what is statistically expected are typically due to causal relations

Although not necessary direct causal relations

Examples:

Consumption of red wine and reduced heart attacks

Books that have a green cover and books that do not sell many copies

Good study habits and good grades

Major Difference: Correlation Symmetrical; Causation Asymmetrical

Being run into in a traffic accident might be a cause for the big dent in your car



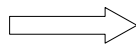
Having a big dent in your car is correlated with having a car accident, but it is not the cause of having a car accident

Causation is directional, correlation is symmetrical

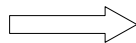
So when correlation points to causation, we still need to establish the direction

Challenge of Establishing Directionality

Does watching violence on TV result in aggressive behavior in children?



Or do the factors that generate aggressive behavior cause children to watch more violence on TV



Causal Loops

Sometimes X causes Y and then Y causes more X
The causation here is still directional, but works in both directions

Back pain may be the cause of a person limping
but walking with a limp may cause further back pain



Snoring and Obesity

There is a positive correlation between obesity and snoring

Does obesity cause (increased) snoring?

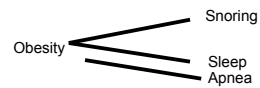
Yes—via fat buildup in the back of the throat

But fat build up also causes sleep apnea

Sleeper stops breathing momentarily and wakes up

As a result of sleep apnea, sufferer is tired and avoids physical activity

Thereby getting more obese



Relating Correlation and Causation

Establishing correlation does not establish causation
But it is a big part of the project!

If X causes Y, then one expects a correlation between X and Y

The greater the value of X (if X is a score variable),
the greater the value of Y

Individuals exhibiting X (if X is a nominal variable)
will have greater values of Y

Independent/Dependent Variables

- **Independent variable**

The variable that is thought to be the cause

The variable that is altered/manipulated in an experiment

The treatment in a clinical trial

- **Dependent variable**

The variable that is thought to be the effect

The variable that one is trying to predict/explain

The outcome in a clinical trial

The dependent variable *depends on* the independent variable

Clicker Question

If average driving speed is the independent variable in an experiment then

Its value depends upon the dependent variable

It is the variable that is manipulated in the experiment

It is the variable that is affected by the manipulation

It is to be explained by finding the cause

Measured versus Manipulated

The strongest tests of causation claims involve manipulation of variables → Experiments

In some contexts, a researcher does not or cannot manipulate the independent variable

Immoral to assign people to categories such as having unprotected sex

Cannot assign people to categories such as being female

If we are nonetheless considering causes in such a case, we refer to a *measured independent variable*

When it is possible to manipulate the independent variable (conduct an experiment), we speak of a *manipulated independent variable*

Clicker Question

Which of the following makes no sense?

- Manipulated independent variable
- Measured independent variable
- Manipulated dependent variable
- Measured dependent variable

Measures (Operational Definitions) and Data

Often causal relations are specified in general terms:

Violence on TV causes violent behavior in school

The variables used to operationally define such variables are sometimes referred to as *measures*. The specific values on these variables are *data*

“The number of gun firings on a given TV show is a good *measure* of violence on the show. We have related *data* on gun firings to *data* on two *measures* of aggressive behavior by those watching the show.”

The measure: Violence operationally defined as
of gun firings

Data on # of gun firings

Correlation without direct causation

Sometimes one variable is directly related causally to another

But sometimes the causation is via some other link



Correlations without direct causation

Ice cream sales and the number of shark attacks on swimmers are correlated

SAT scores and college grades are correlated

Skirt lengths and stock prices are highly correlated (as stock prices go up, skirt lengths get shorter).

The number of cavities in elementary school children and vocabulary size have a strong positive correlation



When causation suspected

Driving red cars is positively correlated with having traffic accidents

Why? Several possible causal scenarios

accident-prone drivers prefer red

people become more aggressive when driving red cars

more dangerous cars tend to be painted red (sports cars)

the color red is harder to see and is more likely to be involved in a 2-car accident

the color red is easier to see, and that leads more drivers to steer towards the red car

Country Music and Suicide

Out of 49 metropolitan areas studied, suicide rates are significantly higher in those in which more country music is played on the radio

Does listening to country music cause suicides?

Or?

Suicidal people choose to live in cities with more country music played on the radio

Country music is popular in cities with high poverty levels and it is the latter that causes higher suicide rates

Or?



Extraneous Variables

Given the number of possible variables to consider, in any given *inquiry* some variables will be correlated with the dependent variable of interest

If these are not the variables we are focusing on, we term them *extraneous*

But

What we term *extraneous* may in fact be the causally relevant variable

So, in testing a causal hypothesis, care must be taken to rule out any causal link between these extraneous variables and the dependent variable

Limits of correlation



Fluoride in water is correlated with lower rate of tooth decay

But why?

Fluoride reduces cavities

People in cities with fluoride enjoy better diets

People in cities with fluoride practice better dental hygiene

People in cities with fluoride have better genetics

Water in cities with fluoride contains other minerals (calcium) that help prevent tooth decay

These additional variables are extraneous from the point of view of the first hypothesis, but they might be the true causes

Telling Causal Stories Can be Fun

Correlation: Amount of ice cream sold correlates with increased deaths by drowning:

"Increases in nuclear power generator accidents (Chernobyl, Three Mile Island...) have resulted in greenhouse gas increases, ozone layer reduction, average world temperature rise and increases in the fraction of heavy water in rain. Concerns about nuclear catastrophe have resulted in increases in eating disorders, especially among those with a genetic predisposition to obesity. Heavy water in rain has resulted in an increase in the specific gravity of cream produced by cows, while the increasing world temperature has resulted in an increasing attendance at beach resorts, coupled with increased consumption of ice cream. The increased weight of fat worried people whose centre of gravity has been lowered by a rising consumption of heavy ice cream has caused an increased number of deaths by drowning." Dr. Paul Gardner, Monash University, Australia

Telling Causal Stories can be Fun - 2

Correlation: Number of fire trucks and amount of fire damage:

“While this could be another case of intentionally starting fires in effort to attract the fire people, this seems highly unlikely. Firefighter salaries are modest. The only logical explanation is that the community just feels so darn safe knowing that there are more fire trucks around, that they simply are not as careful and concerned with fire safety. They feel so confident that a truck would rescue them in an instant, before a fire could spread very far, so they are just careless.

With this inappropriate assumption and subsequent increase in fires, the firefighters are even less able to arrive at a scene on time. Thus, more damage occurs.” Katie Brandt, Purdue University Indianapolis

Beyond causal story telling

If a causal relation exists between two variables, then if we can directly manipulate values on one (the independent variable), we should change values on the other (the dependent variable)

An experiment is precisely an attempt to demonstrate causal relations by **manipulating** the independent variable and **measuring** the change on the dependent variable.

Clicker Question

Does the following argument represent the logic of experimental confirmation?

If X is a cause of Y, then there will be a statistically significant difference in Y when X is present

There is a statistically significant difference in Y when X is present

∴ X is the cause of Y

No, the first premise is usually false

No, one cannot determine statistical significance in an experiment

No, the argument affirms the consequent

No, the argument form is modus ponens whereas modus tollens should be used

The Logic of Causal Research

To confirm or falsify a causal claim based on a correlation, we use *modus tollens*. The first premise in each case, though, is different

Confirming a causal claim:

If X is not a cause of Y [and there is no alternative plausible hypothesis], then there will not be a statistically significant difference in Y when X is present

There is a statistically significant difference in Y when X is present [and there is no alternative plausible hypothesis]

∴ X is a cause of Y

Whether the first premise is true depends critically on how we set up the test of the causal hypothesis—whether we make it very unlikely that anything else could produce a difference in Y

The Logic of Causal Research - 2

Falsifying a causal claim

If X were the cause of Y [and auxiliary assumptions are true and the experimental set up is adequate], then there would be a statistically significant difference in Y when X is present

There is no statistically significant difference in Y when X is present [and auxiliary assumptions are true and the experimental set up is adequate]

∴ X is not the cause of Y

The truth of the first premise depends critically on how we set up the test of the causal claim
