Evaluating experiments

Andy J. Wills

Spurious correlations

http://www.tylervigen.com/spurious-correlations

Piracy and global warming

- Average global temperatures have risen over the last 300 years
- Number of cases of sea piracy have dropped over the same period
- Global warming and piracy are (negatively) correlated

Piracy and global warming

- Average global temperatures have risen over the last 300 years
- Number of cases of sea piracy have dropped over the same period
- Global warming and piracy are (negatively) correlated

- Global warming caused by the absence of sea pirates?
- Would an effective way of reducing global warming be to encourage piracy?

- Absence of sea pirates caused by global warming?
- Would an effective way of reducing piracy be to increase our carbon emissions?

- Association does not demonstrate causation.

Depression and memory

- Depression is associated with over-general memory.

- Depression causes memory problems?
- Memory problems cause depression?
- Both causal directions?
- Neither causal direction (e.g. both caused by childhood trauma).

- It is not possible to distinguish between these accounts on the basis of correlational data.

Longitudinal data does not solve this problem

- Use of night lights in infancy is correlated with myopia in later life (true).
- Seems causal? Causes must precede effects. The later myopia cannot cause the earlier use of night lights. So, night lights must be causing myopia?
- Ban night lights? (genuinely recommended on basis on these data).

Third factor explanations are still possible in longitudinal research

- A third factor causes both the presence of night lights and myopia.
- Developing myopia in later life has a genetic component. If your parents are myopic, this increase the chance you will become myopic.
- Myopic adults, on average, favour higher levels of illumination. This drives their decision to use night lights in their baby's room.
- The parents' myopia causes both the presence of infant night lights and later myopia.
- Ban night lights? Clearly, this would be ineffective.

Correlation does not imply causation

- Correlational research is fundamentally limited.
- It is extremely unlikely that any two variables are completely unrelated.
- Many correlations in psychology are very small e.g.
  - Extroversion explaining 2% of the variation in some other variable.
  - 2% is detectably different from no correlation
  - but not meaningful (everything likely to be related to some degree).
- Partial solution - set a minimum effect size (e.g. ignoring anything below 20% of variation).
- Still doesn't establish causation.

Determining causation through the Experimental Method

- Simplest form
  - Take two groups of people
  - Do different things to those two groups.
  - Measure something

- Independent variable - Intended difference in what we do to the two groups
- Dependent variable - The thing we measure

Example: Testing a treatment for depression

- Group 1 - 6 weeks of the new therapy
- Group 2 - Nothing.
- Take measure of depression at end (e.g. Beck Depression Inventory).
- Group 1 are less depressed than Group 2

- This has the potential to show that the therapy *causes* a reduction in depression.
- Q: Other possibilities?

Pre-existing differences

- Group 1 - 6 weeks of the new therapy
- Group 2 - Nothing.

- What if Group 1 were happier to start with?

- Approaches to this problem
  - Detection
  - Prevention

Detection

- Take pre-treatment measures
- e.g. Measure BDI of both groups before (and after) treatment period.

	Pre	Post
Therapy	25	5
Control	25	25

Prevention

- Construct groups such that we eliminate pre-existing differences.
- Matching - Take BDI measures for everyone. Allocate people to groups in such a way that the average BDI for the two groups is identical (or at least, minimized).
- Randomization - Allocate people to groups randomly.
- Matching versus Randomization - pros and cons.

Our therapy experiment

- Use large, randomized groups.
- Take pre-treatment measures
- Treatment caused the reduction in depression?

	Pre	Post
Therapy	25	5
Control	25	25

Confounding variables

- Any variable, other than the one you are attempting to study, that varies between conditions, and which could potentially have led to the effect you observe.

Attrition

- Attrition - participants dropping out before the end of the study
- If attrition rates vary between conditions, you may have a major problem.

Example

- Pre-treatement BDI scores

						Mean
Therapy	6	8	12	15	30	14.2
Control	6	8	12	15	30	14.2

- The most-depressed 20% drop out of therapy (perhaps because the therapy is quite demanding).
- There are no drop-outs in the control condition (there's not much to drop out from).
- Both therapy and control are inert (no effect) - post-treatment BDI equals pre-treatment BDI.

Example

- Pre-test BDI scores

						Mean
Therapy	6	8	12	15	30	14.2
Control	6	8	12	15	30	14.2

- Post-test BDI scores

						Mean
Therapy	6	8	12	15		10.25
Control	6	8	12	15	30	14.2

- A therapy we know to be ineffective appears to have worked, due to non-random differential attrition.

Placebo effect

- Classic example
  - Someone has a headache
  - Give them a pill with no active ingredient
  - Tell them it's a headache tablet
  - Their headache symptoms reduce
- Lesson - In order to assess drug effectiveness you need to test drug vs. placebo, NOT drug vs. nothing.
- The placebo effect typically accounts for some, but not all, of a drug's effectiveness.
- In the case of anti-depressant medication, the effect seems to be almost entirely placebo.

Placebo effect in psychological therapy

- Perhaps the therapy is inert?
- The treatment group are happier because they have the expectation that what they are receiving will work.
- Problem - a placebo pill is known to be inert; what is the equivalent in therapy?
- There is no agreement - there's someone willing to endorse the effectiveness of almost any therapy.

Placebo effect in psychological therapy

- Solution - set out to show that your new therapy works better than an existing treatment (or, as well as existing treatment, if yours is better in some practical way e.g. cheaper).
- Problem - this is seldom done.
- Worse problem - where it is done, treatments seldom differ.
- Example - Posting a pamphlet on CBT as effective as 6 weeks of 1-to-1 sessions with therapist

Demand characteristics

- Participants' responses may be affected by a desire to comply with what they think the experimenter wants to see.

Example - Evaluative conditioning

- Pairing something neutral with something people already like increases their liking of the neutral item.
- Applied in advertising
- Coke can paired with beautiful smiling people.
- Department store paired with heart-warming story of cross-species friendship

Evaluative conditioning - Experimental demonstration

- Show picture of soft-drink can.
- Pair repeatedly with something positive.
- Liking ratings go up in this treatment group...
- .. but not in a control group where the can and smiles are both presented, but in an unpaired fashion.
- Evidence for evaluative conditioning?

Evaluative conditioning - Alternative explanation

- Participant thinks - “what’s going on here? The experimenter is showing me this coke can and then smiley faces. I think they expect me to like coke more as a result. I wouldn’t want to disappoint them so sure, let’s give it a higher rating than I did last time”.

Confounding variables

- Any variable, other than the one you are attempting to study, that varies between conditions, and which could potentially have led to the effect you observe.

Confounds discussed so far...

- Pre-existing differences (address by matching or randomization)
- Differential attrition (major issue, some partial solutions)
- Hawthorne effect / Placebo effect
- Demand charcteristics

Therapy example

- Compare meditation-based therapy with relaxation training.
- Large, randomized groups.
- No pre-treatment differences in BDI
- No differential attrition
- BDI drops more for meditation than relaxation.
- Meditation is the more effective treatment.

Therapy example - Closer look

- Compare meditation-based therapy with relaxation training.
  - Meditation - Delivered by the people who developed the treatment
  - Relaxation - Delivered by people with no particular investment in relaxation therapy, who have been on a one-week training course in relaxation therapy.
- Large, randomized groups.
- No pre-treatment differences in BDI
- BDI drops more for meditation than relaxation.
- Meditation is the more effective treatment.

Therapy example - Closer look

- Compare meditation-based therapy with relaxation training.
  - Meditation - Delivered by the people who developed the treatment
  - Relaxation - Delivered by people with no particular investment in relaxation therapy, who have been on a one-week training course in relaxation therapy.
- Alternative explanation? - It's not the type of therapy that matters. It's some combination of therapist's belief in the treatment, their experience in delivering it, their general level of therapeutic expertise.

Experimenter Effects - Data analysis - Example

- Diary entries as a measure of happiness.
- Participants write about their feelings
- Experimenter rates for level of happiness.
- If experimenter knows which condition the participant is in, this may bias their assessment of happiness.

Experimenter Effects - Data analysis

- Objective measures immune?
- No! - Data analysis typically involves many decisions, all open to bias.
  - Should I exclude outliers?
  - If so, what's the cut-off?
  - Should I use a parametric or non-parametric test?
  - Are these tests multiple comparisons I should correct for, or separate analyses (for which I don't correct)?
- If the experimenter knows which condition the participants are in, this could bias their decisions.

Data analysis - Example

- My theory predicts people react more quickly to auditory than to visual alarm signals.
- I find this result if I exclude all reaction times above 3 seconds
- But not if I keep all RTs
- and not if I exclude all reaction times below 100ms.
- I choose the 3 second cut-off
- Am I sure that decision was unbiased?

Blind testing

- Single-blind testing - participant does not know which condition they are in.
  - e.g. Drug vs. placebo. Participants do not know which condition they are in.
- Double-blind testing - single-blind testing plus the experimenters do not know which condition is which until after they have completed their analysis.

Order effects

- Auditory versus visual alarm signals, within-subjects design
- Visual (300ms) $\rightarrow$ Auditory (250ms)
- Auditory faster?
- Or, practice effect?

Order effects

- Auditory versus visual alarm signals, within-subjects design
- Auditory (250ms) $\rightarrow$ Visual (300ms)
- Auditory faster?
- Or, fatigue effect?

Order effects

- Auditory versus visual alarm signals, within-subjects design
- Randomly allocate participants to the two orders
- Auditory (250ms) $\rightarrow$ Visual (300ms)
- Visual (300ms) $\rightarrow$ Auditory (250ms)
- Auditory faster - irrespective of order.
- No practice or fatigue effect (mean RT across conditions 275 ms at time 1 and time 2).

Difference versus no difference designs

- The preferred hypothesis is that people differ in the speed with which they react to auditory and visual alarm signals.
- The alternative theory against which this is compared is that there is no difference (nil hypothesis).
- Problem - Experimental control is never perfect.
- Thus - the nil hypothesis is almost certainly wrong, and detectably so if you test enough people.
- Thus - the result of the study is known before you run it.
- Thus - There was no point in running it.

Better alternatives 1

- One-tailed tests
  - The preferred theory is that auditory is faster.
  - The alternative theory against which this is compared is that there is no difference (nil hypothesis).
  - If you find visual faster, you have disproved your theory.
  - So, whatever the result, there was a point to running this experiment (because the theory was falsifiable).

Better alternatives 2

- Ordinally different theories
  - One well-established theory predicts that auditory is faster.
  - Another well-established theory predicts that visual is faster.
  - Whatever you find in this study, you've gained information (except in the unlikely case where the nil hypothesis was true).

Next week's workshops

- Bring a laptop!
- Games:
  - Horse-race game: Normal distribution
  - Exam-hall bingo: Sample size and statistical power
  - Shove ha'penny: Regression to the mean
  - Good and evil: Illusory correlation
- Worksheets and practical activities:
  - Introduction to R Studio
  - Exploring data
  - Group difference
  - Evidence
  - Analyzing your own data

Further reading/ watching

Only lecture content on these topics is examinable.

- `https://www.youtube.com/watch?v=NW2EmATcb6o`
- `http://www.youtube.com/watch?v=ZgXfWmgA9NE` (don't watch this one if you are easily upset or offended)
- `http://en.wikipedia.org/wiki/Hawthorne_effect`