Recitation 5

Rachel Lee

5 Oct 2021

Lord’s Paradox

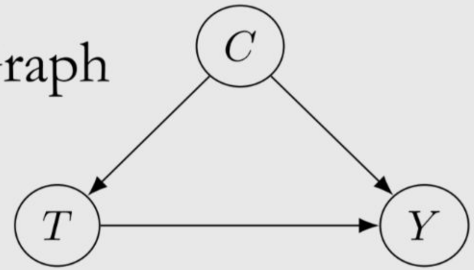
https://m-clark.github.io/docs/lord/index.html

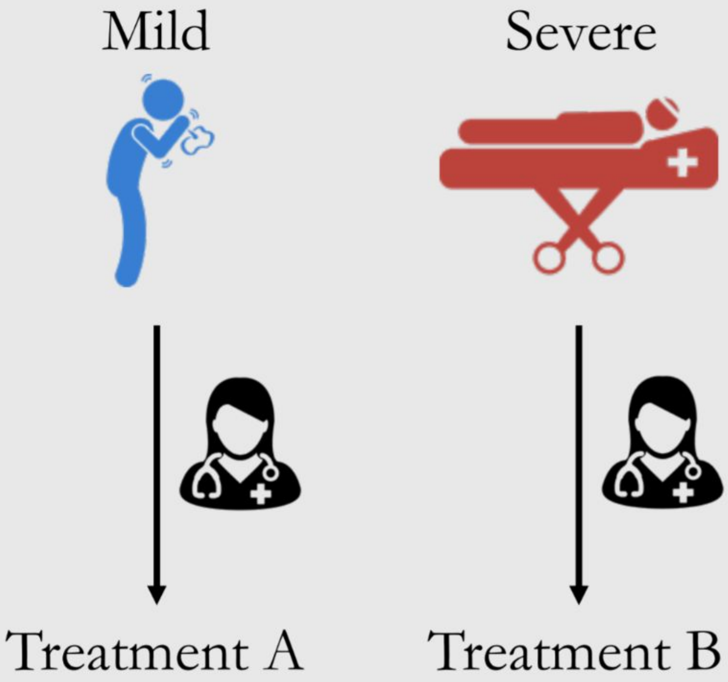
Simpson's paradox: scenario 1 (treatment B)

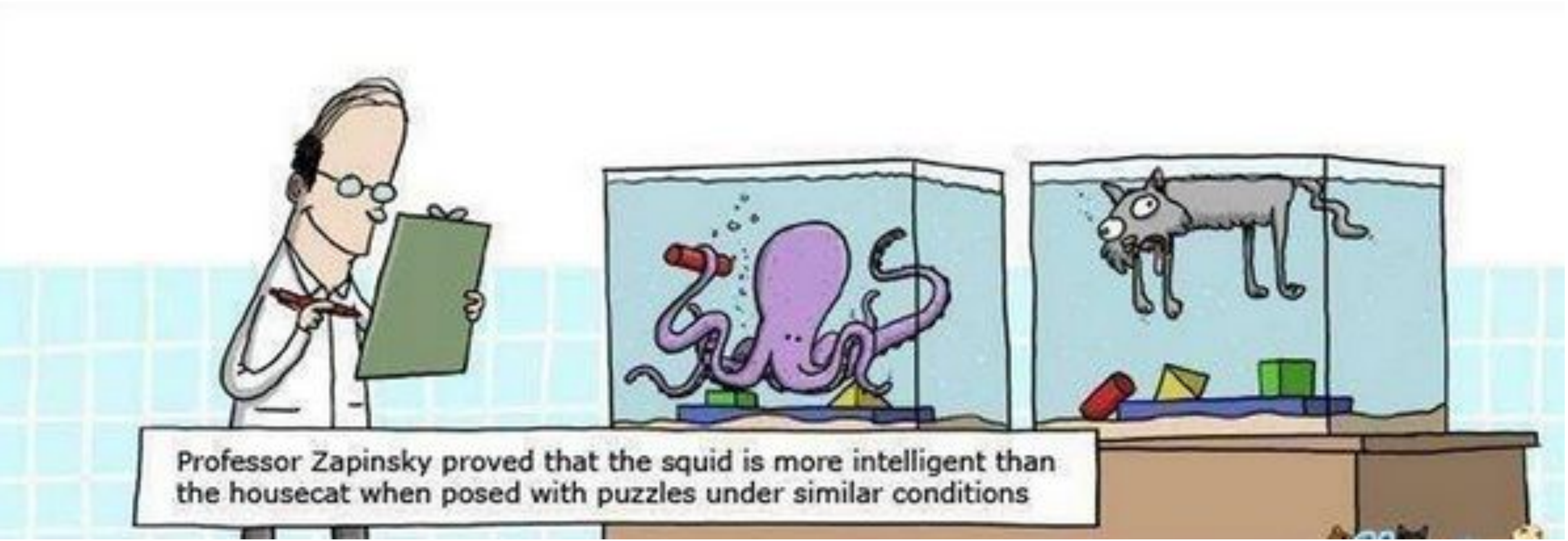
Condition

Treatment		Mild	Severe	Total
	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

Causal Graph






Professor Zapinsky proved that the squid is more intelligent than
the housecat when posed with puzzles under similar conditions

Validity of Inferences from Experiments

- In the context of experiments that aim to estimate a causal effect, "validity" refers to the correspondence between the causal estimate and the real world truth.
- In a series of books, psychologist Donald Campbell and colleagues organized potential threats to the validity of inferences from experiments [Campbell and Stanley (1963), Cook and Campbell (1979), and Shadish, Cook, and Campbell (2002)*].
- The Campbellian approach to validity is not statistical (i.e., it is not based on the mathematical assumptions required for identification and unbiased estimation of population parameters and the consequences of their violation); rather, it is based on qualitative elements of experimental design which, if violated, lead to causally ambiguous or uninterpretable results. See Shadish (2010; Psychological Methods) for further discussion.

*This book is the definitive resource on validity in experimentation.

Types of Validity

- Statistical Conclusion Validity
  - Are the assumptions of statistical methods used to make inferences satisfied?
  - Were the statistical methods implemented correctly?
- Internal Validity
  - Was the experiment designed in such a way as to allow results to be causally attributed to the treatment manipulation?
  - Or, are there other factors that confound one's ability to do that?
- Construct Validity
  - Are scales used to measure key quantities meaningful?
  - Do they reliably measure what they are supposed to measure?
- External Validity
  - Do study findings generalize across populations, settings, or time?
  - Or, do findings only apply to the particular experimental units in the particular study?

Statistical Conclusion Validity

- Bad statistical practice is a threat to SCV
  - QRPs (Questionable Research Practices), including fraud
    - **Fabricating Data**. From the NYT article on Diederik Stapel: “The experiment — and others like it — didn’t give Stapel the desired results, he said. He had the choice of abandoning the work or redoing the experiment. But he had already spent a lot of time on the research and was convinced his hypothesis was valid. “I said — you know what, I am going to create the data set,” he told me.”
    - **Manipulating Data**. From the BuzzFeed article on Brian Wansink: “Wansink also acknowledged the paper was weak as he was preparing to submit it to journals. The p-value was 0.06, just shy of the gold standard cutoff of 0.05. It was a “sticking point,” as he put it in a Jan. 7, 2012, email.“It seems to me it should be lower,” he wrote, attaching a draft. “Do you want to take a look at it and see what you think. If you can get the data, and it needs some tweeking, it would be good to get that one value below .05.”

Statistical Conclusion Validity

- Bad statistical practice, continued
  - The terms **p-hacking**, **researcher degrees of freedom**, **fishing**, and **data dredging** are similar. They refer to repeatedly testing the data until something pops up as statistically significant. A quote from an email Brian Wansink sent to a visiting researcher who would be joining his team:

    > "Second, think of all the different ways you can cut the data and analyze subsets of it so see when this relationship holds. For instance, if it works on men but not women, we have a moderator. Here are some groups you'll want to break out separately: males, females, lunch goers, dinner goers, people sitting alone, people eating with groups of 2, people eating in groups of 2+, people who order alcohol, people who order soft drinks, people who sit close to buffet, people who sit far away, and so on… Third, look at a bunch of different DVs. These might include # pieces of pizza, # trips, Fill level of plate, Did they get dessert, Did they order a drink, and so on… This is really important to try and find as many things here as possible before you come. First, it will make a good impression on people and helps you stand out a bit. Second, it will be the highest likelihood of you getting something publishable out of your visit.

  - Bad practice need not necessarily be coupled intentional deception or manipulation. For example, Gelman and Lokin discuss "Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data." (p. 2)

Statistical Conclusion Validity

- Bad statistical practice, continued
  - HARKing (Hypothesizing After the Results are Known)
  - Kerr’s summary of costs of HARKing:
    - Translating Type I errors into hard-to-eradicate theory
    - Propounding theories that cannot (pending replication) pass Popper's disconfirmability test
    - Disguising post hoc explanations as a priori explanations
    - Not communicating valuable information about what did not work
    - Taking unjustified statistical licence
    - Presenting an inaccurate model of science to students
    - Encouraging ‘fudging’ in other grey areas
    - Making us less receptive to serendipitous findings
    - Encouraging adoption of narrow, context-bound new theory
    - Encouraging retention of too-broad, disconfirmable old theory
    - Inhibiting identification of plausible alternative hypotheses
    - Implicitly violating basic ethical principles

Internal Validity

Internal Validity

- Threats to internal validity are design aspects which, if present, block the researcher's ability to attribute changes in the outcome solely to the independent (experimentally manipulated) variable.
  - Selection bias
    - When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
    - <u>Example</u>. Mode of child delivery is not randomly assigned; decisions are made by expectant mothers and their healthcare providers. At baseline, before delivery, mothers who birth via c-section have higher BMI. Three years after delivery, children born via c-section have higher rates of obesity. The effect of mode of delivery is confounded due to selection bias.
  - Attrition
    - Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
    - <u>Example</u>. In a randomized experiment looking at the efficacy of two therapy-based treatment protocols for depression, therapy A is easy to adhere to and therapy B is difficult to adhere to. Most severely depressed participants drop out of therapy B after a couple of sessions, thereby inflating the relative benefit to therapy A.

Internal Validity

- Continued
  - Testing
    - Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
    - Example. Experimenters interested in the long-term effects of memory training on recall accuracy repeatedly test participants after each training session with the same set of photos. Over time, as participants become familiar with the photos they are able to match them and recall them more efficiently. These gains, however, are not due to the trainings.
  - Regression (to the mean)
    - Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
    - Example. Participants for a study on the efficacy of a new math curriculum are selected based on their previous end-of-year standardized math scores; only those with scores at or below the 20^{th} percentile are selected. Suppose the curriculum truly has no effect, neither positive nor negative. Nevertheless, the math scores for those participants will likely improve, on average, when measures one year later due to regression.

Internal Validity

- Continued
  - Maturation
    - Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
    - <u>Example</u>. In any study of growth, either physical or emotional, with human participants, it is plausible that changes that occur during the course of the study may be due to natural maturation that is independent of any treatments administered.
  - History
    - Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.
    - <u>Example</u>. A school district carried out a three-year study of the effectiveness of a district-wide initiative to improve student health outcomes via school lunch programming and physical fitness initiatives. During the middle of the second year of the study, a federal program flooded the neighborhoods served by the school district with supplementary nutritional and physical fitness programs; many children participated. The effects of the district initiative are now confounded with the federal program.

- **Selection bias** - When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
- **Attrition** - Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
- **Testing** - Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
- **Regression (to the mean)** - Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
- **Maturation** - Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
- **History** - Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.

Internal Validity and Design Elements

- Design elements may reduce risk due to internal validity threats.
  - Selection bias; randomly assign or, if not possible control for confounding variables.
  - Attrition; follow-up with participants to minimize attrition; if unavoidable, compare baseline covariate profiles of those who dropped with those who did not and consider sensitivity analyses.
  - Testing; include a comparison group; use different scales or allow a wash out period between using the same scale twice.
  - Regression; include a comparison group; do not select experimental units based on extreme pretest cutoffs.
  - Maturation; include a comparison group.
  - History; include a comparison group.

An educational training will be investigated for increasing awareness of the symptoms and transmissibility of the coronavirus that causes COVID-19. The same assessment will be given at baseline, before the training, and posttest, after the training. After taking the baseline assessment, but before taking the post-test, a number of participants who were curious about answers looked up the answers on the internet.

- **Selection bias** - When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
- **Attrition** - Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
- **Testing** - Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
- **Regression (to the mean)** - Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
- **Maturation** - Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
- **History** - Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.

Suppose there is interest in the efficacy of a blood pressure medication on lowering blood pressure. Participants' blood pressure is measured several times throughout the study, both before and after treatment with the medication begins; there is no comparison group. Most of the participants belong to the same HMO health insurance program, which started a mindfulness meditation program for stress reduction during the course of the experiment. Many of the blood pressure study participants also participate in the mindfulness program.

- **Selection bias** - When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
- **Attrition** - Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
- **Testing** - Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
- **Regression (to the mean)** - Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
- **Maturation** - Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
- **History** - Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.

Suppose there is interest in the efficacy of a new mathematics curriculum on math achievement. Participants' math knowledge is measured twice in a pre-post design: once at the beginning of the year, before being exposed to the curriculum, and again at the end of the semester, after being exposed to the curriculum; there is no comparison group. Participants were recruited for the study only if they scored below the 25th percentile on last years' state standardized mathematics test.

- **Selection bias -** When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
- **Attrition -** Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
- **Testing -** Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
- **Regression (to the mean) -** Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
- **Maturation -** Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
- **History -** Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.

Suppose there is interest in the efficacy of a new mathematics curriculum on math achievement. Participants' math knowledge is measured once in a posttest-only design; there is a comparison group that uses the business-as-usual curriculum. At the beginning of the school year, parents were allowed to select which curriculum they wanted their children to receive.

- **Selection bias -** When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
- **Attrition** - Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
- **Testing** - Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
- **Regression (to the mean) -** Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
- **Maturation** - Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
- **History** - Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.

Suppose there is interest in the efficacy of a new mathematics curriculum on math achievement. Participants' math knowledge is measured once in a posttest-only design; there is a comparison group that uses the business-as-usual curriculum. Participants were randomly assigned to curriculums, but there was some dissatisfaction among parents of students who received the new curriculum, so a number of students changed groups early on in the school year.

- **Selection bias -** When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
- **Attrition** - Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
- **Testing** - Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
- **Regression (to the mean) -** Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
- **Maturation** - Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
- **History** - Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.

Researchers interested in studying the effect of giving grocery store vouchers to parents of children in neighborhoods with high levels food insecurity measure will measure height and weight of children in participating families at the beginning and end of the program year.

- **Selection bias** - When participants self-select (or are otherwise non-randomly allocated) into treatment conditions, differences across groups on the outcome cannot be unambiguously attributed to the treatment conditions.
- **Attrition** - Even in a randomized study, bias can result if participants drop out at different rates from treatment arms, and, especially if drop out is related to other factors.
- **Testing** - Measuring/testing may have an effect on the outcome that is independent of any effect of the manipulated treatment condition.
- **Regression (to the mean)** - Over time, those with extreme scores tend to return toward the mean when repeatedly measured. Thus, if a group is selected at baseline because of their extreme scores and then moves closer to the mean at follow-up, the cause is confounded with regression.
- **Maturation** - Changes in participants may be observed due to natural maturation as opposed to any effect of treatment. This is particularly hard to tease out without a well-designed comparison group.
- **History** - Natural events in the lives of participants that occur during the duration of a longitudinal experiment may influence outcomes apart from the treatment conditions. Again, this is hard to isolate without a good comparison group.