

MATH 1051 - Lecture 02

Variables, Sampling, and Experimental Design

Kat Clark & Wesley Burr

Trent University

Relationships Between Variables

Explanatory and Response Variables

When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other. We label these as **explanatory** and **response** variables: the variable that affects the other is the **explanatory**, and the variable that is affected is the **response**.

explains

→ the thing responding to
changes in the explanatory
variable

Explanatory and Response Variables

- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

explanatory variable ^{might affect} → response variable

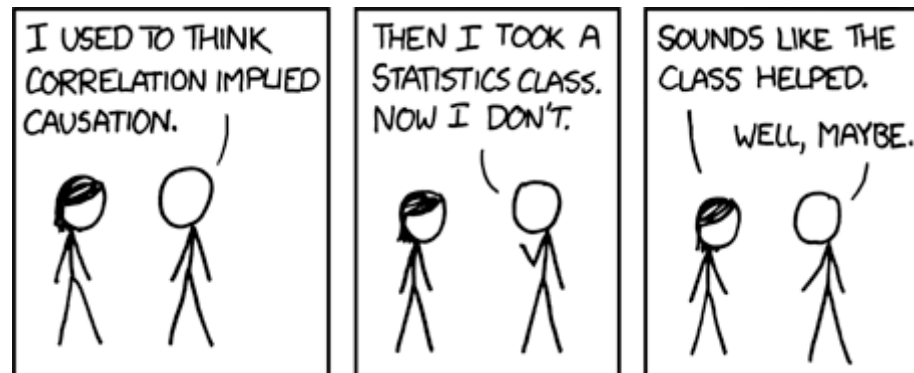
- Labelling variables as explanatory and response **does not guarantee** the relationship between the two is actually **causal**, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

Causation: Types of Study

Observational study: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely “observe”, and can only establish an association between the explanatory and response variables.

Experiment: Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.

One thing to walk away with is: “correlation does not imply causation”¹.



1. <http://xkcd.com/552/>

Observational Studies and Sampling Strategies

New study sponsored by General Mills says that eating breakfast makes girls thinner



By ALEX DOMINGUEZ, Associated Press

- Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years.
- Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute. The study received funding from the National Institutes of Health and cereal-maker General Mills. NIH
- “Not eating breakfast is the worst thing you can do, that’s really the take-home message for teenage girls,” said study author Bruce Barton, the Maryland institute’s president and CEO.

- The fiber in cereal and healthier foods that normally accompany cereal, such as milk and orange juice, may account for the lower body mass index among cereal eaters, Barton said.
- ✶• The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio and Maryland who were tracked between ages 9 and 19. Results of the study appear in the September issue of the Journal of the American Dietetic Association.
- Nearly one in three adolescent girls in the United States is overweight, according to the association. The problem is particularly troubling because research shows becoming overweight as a child can lead to a lifetime struggle with obesity.
- ✶• As part of the survey, the girls were asked once a year what they had eaten during the previous three days. The data were adjusted to compensate for factors such as differences in physical activity among the girls and normal increases in body fat during adolescence.

What type of study is this: observational or experiment?

“Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days.”

This is an **observational study** since the researchers merely observed the behaviour of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

There is an **association** between girls eating breakfast and being slimmer.

Who sponsored the study?

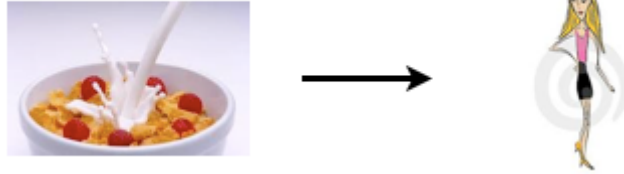
General Mills (cereal manufacturer)

observational

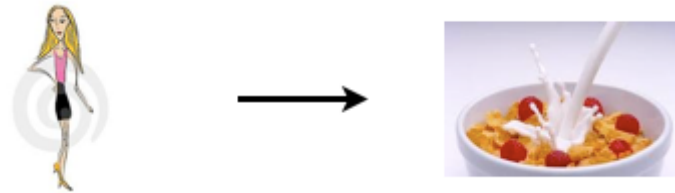
-no causation

3 Possible Explanations

1. Eating breakfast causes girls to be thinner.



2. Being thin causes girls to eat breakfast.



3 Possible Explanations

3. A third variable is responsible for both. What could it be? An extraneous variable that affects both the explanatory and the response variable and that make it seem like there is a relationship between the two is called a **confounding variable**.



Prospective versus Retrospective Studies

A **prospective study** identifies individuals and collects information as events unfold. *↳ prospect*

- **Example:** The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.

Retrospective studies collect data after events have taken place.

- **Example:** Researchers reviewing past events in medical records.

Obtaining Good Samples

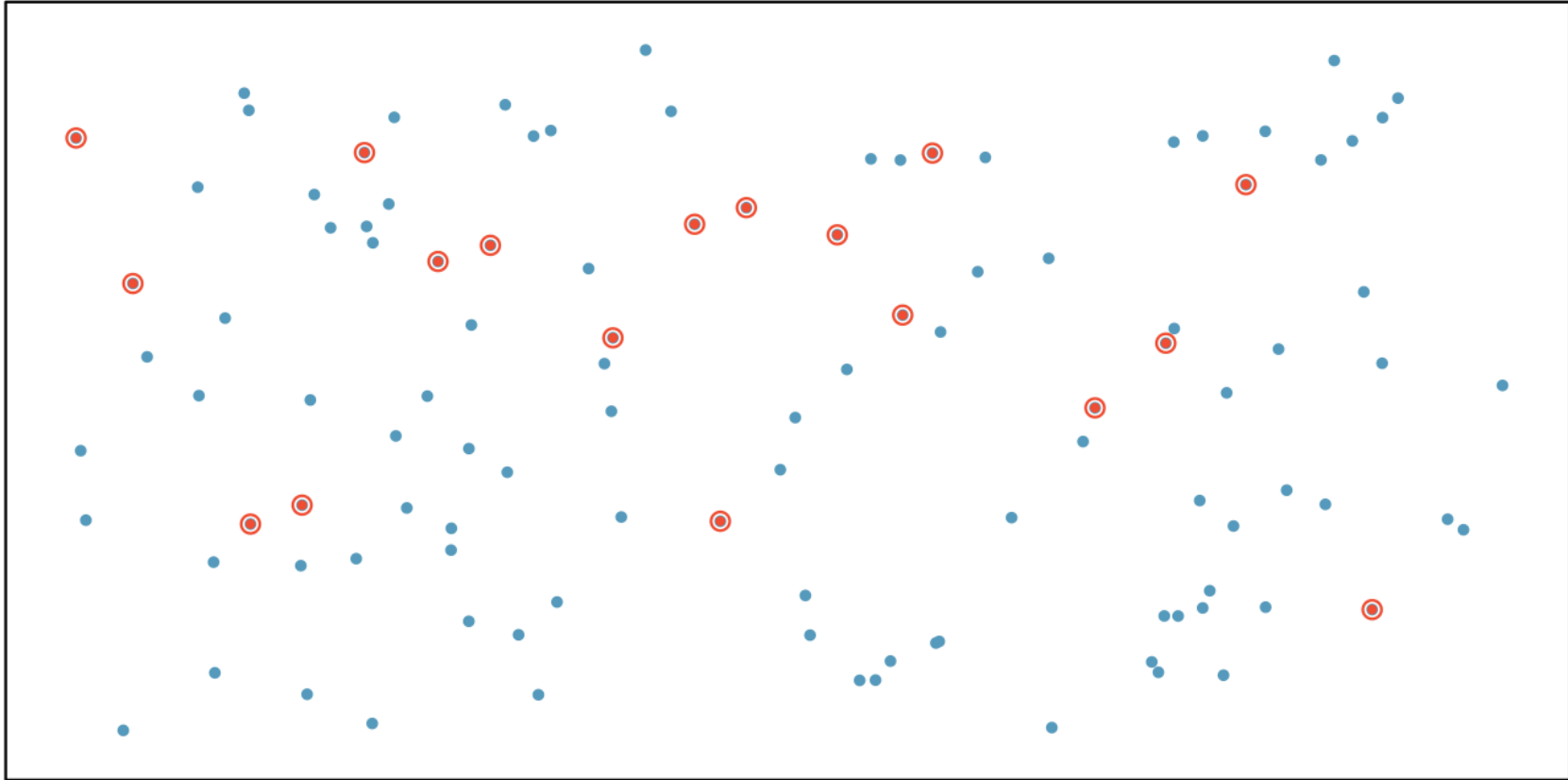
Almost all statistical methods are based on the notion of implied randomness.

If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.

Most commonly used random sampling techniques are **simple**, **stratified**, and **cluster** sampling.

Simple Random Sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



Simple Random Sampling

Advantages

- simple (it's in the name!)
- each member has an equal chance of being included

Disadvantages

- logistically difficult with larger populations (possibly)
- may miss small subpopulations
- not targeted in any way

Conclusion and Note

The only way to avoid the potential bias of samples is to select the sample **randomly**. As you deal with more complicated situations, there are a wide variety of sampling **schema** designed to help avoid potential biases. This can be complicated, but is very powerful.

Experiments

Principles of Experimental Design

Studies where researchers **assign treatments** to cases are called **experiments**. If the assignment of treatments to cases is random (e.g., using a coin flip to determine which treatment a patient receives), the study is called a **randomized experiment**.

- **treatment**: comes from agricultural experiments, where you **treat** the soil
 - some kind of intervention, where the experimenter **changes** the subject in some way
- **experiment**: any kind of study where researchers assign treatments to cases/subjects
- **randomized experiment**: an experiment where the assignment includes a random element

Randomized experiments have a series of four principles.

Principle 1: Controlling

Researchers assign treatments to cases, and do their best to **control** for other differences between groups.

Example: in a drug trial, patients may be asked to take a pill daily. Some may take the pill dry (ick!), some with just a sip of water, some with coffee, and others with juice. To **control** for the effect of accompanying liquid, a doctor may ask all patients to drink a 12 oz glass of water with the pill.

Principle 2: Randomization

Researchers **randomize patients** into treatment groups to account for variables that cannot be controlled.

Example: some patients are more susceptible to disease than others due to dietary habits. **Randomizing** patients into treatment/control groups helps even out these differences, possibly preventing accidental bias.

Principle 3: Replication

The more cases researchers observe, the more accurately they can estimate the effects of explanatory variables on response variables. In a single study, we **replicate** by collecting a sufficiently large sample. Additionally, scientists often replicate an entire study over again to verify earlier findings.

Principle 4: Blocking

Researchers sometimes know (or suspect) that variables other than the treatment influence the response. Under this situation, they may first group individuals by this variable, and then randomize cases within each block. This is known as **blocking**.

Example: If we were researching the effect of a drug on heart attacks, we might first split patients into high-risk and low-risk **blocks** (based on diet, physique, genetic screening, or some other approach), and *then* randomly assign half of each block to the control group, and the other half to the drug (treatment) group.

More on Blocking



We would like to design an experiment to investigate if energy gels make you run faster:

- **Treatment:** energy gel
- **Control:** no energy gel

It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status.

Blocking for Pro Status

- Divide the sample to pro and amateur
- Randomly assign pro athletes to treatment and control groups
- Randomly assign amateur athletes to treatment and control groups
- Pro/amateur status is equally represented in the resulting treatment and control groups

Why is this important? Other variables to block for?

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on students of different genders, so wants to make sure all genders are proportionally represented in each group. Which of the below is correct?

1. There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
2. There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)
3. There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)
4. There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on students of different genders, so wants to make sure all genders are proportionally represented in each group. Which of the below is correct?

1. There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
2. **There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)**
3. There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)
4. There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Difference between Blocking and Explanatory Variables

Factors are conditions we can impose on the experimental units.

Blocking variables are characteristics that the experimental units come with, that we would like to control for.

More Experimental Design Terminology

Placebo: fake treatment, often used as the control group for medical studies

Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment

Blinding: when experimental units do not know whether they are in the control or treatment group

Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Double Blind Experiments

- If members of a control group have knowledge (or suspicion) that they are not being given an actual treatment, then the effect of the treatment cannot be accurately assessed (depending).
- A **double-blind** experiment sets up the protocol for the experiment so neither the experimenters nor the subjects are aware of the group status of the subjects (e.g., treatment versus control).
- This typically reduces experimental bias. It can also be useful for non-medical trials where any level of experimental subjectivity is possible.

Practice

What is the main difference between observational studies and experiments?

- ~~1.~~ Experiments take place in a lab while observational studies do not need to.
- ~~2.~~ In an observational study we only look at what happened in the past *retro but prospecti*
- 3. Most experiments use random assignment while observational studies do not.
- 4. Observational studies are completely useless since no causal inference can be made based on their findings.

Practice

What is the main difference between observational studies and experiments?

1. Experiments take place in a lab while observational studies do not need to.
2. In an observational study we only look at what happened in the past.
3. **Most experiments use random assignment while observational studies do not.**
4. Observational studies are completely useless since no causal inference can be made based on their findings.

Random Assignment versus Random Sampling

<i>ideal experiment</i>	<p>Random assignment</p>	<p>No random assignment</p>	<i>most observational studies</i>
<p>Random sampling</p>	<p>Causal conclusion, generalized to the whole population.</p>	<p>No causal conclusion, correlation statement generalized to the whole population.</p>	<p>Generalizability</p>
<p>No random sampling</p>	<p>Causal conclusion, only for the sample.</p>	<p>No causal conclusion, correlation statement only for the sample.</p>	<p>No generalizability</p>
<i>most experiments</i>	<p>Causation</p>	<p>Correlation</p>	<p>bad observational studies</p>

Handwritten annotations: Two red arrows point down to the top row. Two orange arrows point from the top row to the middle row. Two orange arrows point from the middle row to the bottom row. A red 'bad' stamp is over the word 'bad' in the bottom right cell.