Announcements

- All RStudio accounts should be up and running.
- HW1 due next Thursday (Feb 2):
  - Questions from the book
  - On your own answers for Lab 1 (next Monday’s lab)
- Practice online quiz experience: 87% took the quiz. If you’re among the 13% who didn’t, make sure you review what you missed.
Practice online quiz - commonly missed questions

Questions 1 and 3:
Are height and weight continuous or discrete numerical variables? Hint: Are they measured or counted?
(a) Males go home more than females since the max for males is higher.
(b) The # of times males go home is more variable than females.
(c) Something is wrong with the males plot since half of the box is missing.
(d) Median and the 75th percentile for the male distribution are equal.
1. Observational studies and sampling strategies
   - Observational studies
   - Three sampling methods
   - Recap

2. Experiments
New study sponsored by General Mills says that eating breakfast makes girls **thinner**

Study: Breakfast Helps Girls Stay Slim
I love these studies....and finding out who sponsored them!

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.

"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.

A girl who reported eating breakfast on all three days had, on average, a body mass index 0.7 units lower than a girl who did not eat breakfast at all. If the breakfast included cereal, the average was 1.65 units lower, the researchers found.
What type of study is this, observational study or an 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.”

What is the conclusion of the study?

Who sponsored the study?
3 possible explanations:
3 possible explanations:

1. Eating breakfast causes girls to be thinner.
3 possible explanations:

1. Eating breakfast causes girls to be thinner.

2. Being thin causes girls to eat breakfast.
3 possible explanations:

1. Eating breakfast causes girls to be thinner.

2. Being thin causes girls to eat breakfast.

3. A third variable is responsible for both. What could it be?
Lurking variables

- 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 are called lurking or confounding variables.
- Observational studies do not control for lurking variables hence they can’t be used for establishing causality but they are useful for discovering trends and possible relationships.
Types of observational studies

- A *prospective study* identifies individuals and collects information as events unfold.
  - Ex: Nurses’ Health Studies for investigating factors that influence women’s health (http://www.channing.harvard.edu/nhs).

- *Retrospective studies* collect data after events have taken place.
  - Ex: Use medical records from past patients to estimate the optimum time to start anticoagulant therapy after surgery.
Three sampling methods

1. *Simple random sampling:* Each case in the population has equal chance of being selected.
Three sampling methods

1. **Simple random sampling**: Each case in the population has equal chance of being selected.

2. **Stratified sampling**: First divide population into homogenous strata, then sample cases randomly from within each strata.
Three sampling methods

1. **Simple random sampling**: Each case in the population has equal chance of being selected.

2. **Stratified sampling**: First divide population into homogenous strata, then sample cases randomly from within each strata.

3. **Cluster sampling**: First divide population into clusters, then randomly sample a few clusters, then randomly sample cases from within those clusters.
   - Clusters don’t need to be homogenous.
   - Usually clusters are formed in a way that makes sampling easier (more economical), such as based on geography.
Three sampling methods

1. **Simple random sampling**: Each case in the population has equal chance of being selected.

2. **Stratified sampling**: First divide population into homogenous strata, then sample cases randomly from within each strata.

3. **Cluster sampling**: First divide population into clusters, then randomly sample a few clusters, then randomly sample cases from within those clusters.
   - Clusters don’t need to be homogenous.
   - Usually clusters are formed in a way that makes sampling easier (more economical), such as based on geography.

Describe a situation where cluster sampling would be more efficient than simple random or stratified sampling.
Why random sampling?

- Random sampling yields a sample that is representative of the population.
- Conclusions based on studies using randomly sampled data can be extended to the population at large.
- If a study is based on a non-random (non-representative) sample, conclusions are only true for that particular sample and may not be true for the population at large.
1. Observational studies and sampling strategies

2. Experiments
   - Principles of experimental design
   - Reducing bias in human experiments
   - Recap
Experiments

- In an experiment researchers randomly assign subjects (also called experimental units) to treatment and control groups.

- The objective of random assignment (or randomization) is to ensure that the treatment and control groups are similar in all characteristics, except for the treatment being investigated, so that any observed difference between the two groups is due to the treatment.

- While observational studies can be used to only infer association, experiments allow researchers to make causal statements.
Principles of experimental design

1. **Control**: Compare treatment of interest to a control group.
Principles of experimental design

1. **Control**: Compare treatment of interest to a control group.
2. **Randomize**: Randomly assign subjects to treatments.
Principles of experimental design

1. **Control**: Compare treatment of interest to a control group.
2. **Randomize**: Randomly assign subjects to treatments.
3. **Replicate**: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
Principles of experimental design

1. **Control:** Compare treatment of interest to a control group.

2. **Randomize:** Randomly assign subjects to treatments.

3. **Replicate:** Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.

4. **Block:** If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.
We would like to design an experiment to investigate if energy gels makes you run faster:
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 professional and amateur athletes differently:

- Block for pro status
- Divide the sample to professional and amateur athletes
- Randomly assign professional athletes to treatment and control groups
- Randomly assign amateur athletes to treatment and control groups

Can you think of any other variables we should block for?
We would like to design an experiment to investigate if energy gels makes you run faster:

- Treatment: energy gel
- Control: no energy gel

It is suspected that energy gels might affect pro and amateur athletes differently:
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:
  - Block 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
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:
  - Block 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

Can you think of any other variables we should block for?
Clicker question

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 males and females, so wants to make sure both genders are represented equally under different conditions. Which of the below is correct?

(a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(b) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

(c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)
Reducing bias in human experiments

- Randomized experiments are the gold standard, but they do not ensure an unbiased perspective into the cause and effect relationships in all cases.
- Bias can unintentionally arise especially in human studies.
- Using a control for the treatment is one way to reduce this bias.
- Keeping the patients uninformed about their treatment, \textit{blinding}, can also help reduce bias. This is usually done through the use of a \textit{placebo}.
- If both the patients and the researchers are uninformed about which treatment which patient is getting, this is called a \textit{double-blind} study.
Clicker question

What is the main difference between observational studies and experiments?

(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) Most experiments use random assignment while observational studies do not.

(d) Observational studies are completely useless since no causal inference can be made based on their findings.
Random assignment vs. random sampling

<table>
<thead>
<tr>
<th>Random assignment</th>
<th>No random assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sampling</td>
<td>Causal inference, generalized to the whole population.</td>
</tr>
<tr>
<td></td>
<td>No causal inference, correlation statement generalized to the whole population.</td>
</tr>
<tr>
<td>No random sampling</td>
<td>Causal inference, only for the sample.</td>
</tr>
<tr>
<td></td>
<td>No causal inference, correlation statement only for the sample.</td>
</tr>
<tr>
<td></td>
<td>Conclusions not generalized to population.</td>
</tr>
</tbody>
</table>

Causation

Correlation