An Overview of Topics for Exam 1

Ryan Miller

Data

The work of statisticians is intrinsically tied to data

1) We want to be able to *identify* and *describe* important trends
2) We want to be able to *explain* those trends

Describing Data

- **Cases** - the units of observation that our data describe (typically the rows of the data spreadsheet)
  - Different colleges in the US, different babies in an experiment, etc.

Describing Data

- **Cases** - the units of observation that our data describe (typically the rows of the data spreadsheet)
  - Different colleges in the US, different babies in an experiment, etc.
- **Variables** - any characteristic that is recorded for each case (generally stored in a *column*)
  - *categorical* variables place the cases into groups (ie: region, treatment/control, etc.)
  - *quantitative* variables (numeric variables) record a numeric measurement for each case (ie: enrollment, height, etc.)

Describing Data

- **Cases** - the units of observation that our data describe (typically the rows of the data spreadsheet)
  - Different colleges in the US, different babies in an experiment, etc.
- **Variables** - any characteristic that is recorded for each case (generally stored in a *column*)
  - *categorical* variables place the cases into groups (ie: region, treatment/control, etc.)
  - *quantitative* variables (numeric variables) record a numeric measurement for each case (ie: enrollment, height, etc.)
- Being able to distinguish different types of variables allows us to determine the proper graphs and summary measures to use!
  - 1 categorical variable -> summarize with proportions; graph with bar charts
  - 1 quantitative variable -> summarize with mean, median, standard deviation, Q1, Q3, IQR; graph with histograms or box plots

Relationships Between Variables

- The most interesting aspects of a data set are the relationships between variables

Relationships Between Variables

- The most interesting aspects of a data set are the relationships between variables
- Two variables are **associated** if the distribution of one variable depends upon the other variable
  - Sex (categorical) and Height (quantitative) are associated if the distribution of male heights differs from the distribution of female heights

Relationships Between Variables

- The most interesting aspects of a data set are the relationships between variables
- Two variables are **associated** if the distribution of one variable depends upon the other variable
  - Sex (categorical) and Height (quantitative) are associated if the distribution of male heights differs from the distribution of female heights
  - Victims Race (categorical) and Death Penalty Verdict (categorical) are associated if the distribution of death penalty outcomes differs for white and black offenders

Relationships Between Variables

- The most interesting aspects of a data set are the relationships between variables
- Two variables are **associated** if the distribution of one variable depends upon the other variable
  - Sex (categorical) and Height (quantitative) are associated if the distribution of male heights differs from the distribution of female heights
  - Victims Race (categorical) and Death Penalty Verdict (categorical) are associated if the distribution of death penalty outcomes differs for white and black offenders
  - Tuition (quantitative) and Average Faculty Salary (quantitative) are associated if higher tuition corresponds with higher average faculty salaries

Describing Associations

How we describe and display an association is also linked to the types of variables involved:

- Two categorical variables
  - Summarize using a **contingency table** (usually the reporting **conditional proportions**)
  - Visualize using a *stacked or conditional bar chart*

Describing Associations

How we describe and display an association is also linked to the types of variables involved:

- Two categorical variables
  - Summarize using a **contingency table** (usually the reporting **conditional proportions**)
  - Visualize using a *stacked or conditional bar chart*
- Two quantitative variables
  - Summarize using the **correlation coefficient** or the *slope* of the **regression line**
  - Visualize using a *scatter plot*

Describing Associations

How we describe and display an association is also linked to the types of variables involved:

- Two categorical variables
  - Summarize using a **contingency table** (usually the reporting **conditional proportions**)
  - Visualize using a *stacked or conditional bar chart*
- Two quantitative variables
  - Summarize using the **correlation coefficient** or the *slope* of the **regression line**
  - Visualize using a *scatter plot*
- One categorical and one quantitative variable
  - Summarize by comparing *means, medians, Q1, Q3*, etc. across groups
  - Visualize using *side-by-side boxplots*

Explaining Associations

- Once we've found an association, the next step is to try and rule out possible explanations for why it exists

Explaining Associations

- Once we've found an association, the next step is to try and rule out possible explanations for why it exists
- Possible explanations we've looked at include:
  - The presence of a confounding variable (race of the victim)

Explaining Associations

- Once we've found an association, the next step is to try and rule out possible explanations for why it exists
- Possible explanations we've looked at include:
  - The presence of a confounding variable (race of the victim)
  - Sampling bias (a systematic favoring of longer words)

Explaining Associations

- Once we've found an association, the next step is to try and rule out possible explanations for why it exists
- Possible explanations we've looked at include:
  - The presence of a confounding variable (race of the victim)
  - Sampling bias (a systematic favoring of longer words)
  - Other types of bias (the placebo effect, leading questions, measurement bias)

Explaining Associations

- Once we've found an association, the next step is to try and rule out possible explanations for why it exists
- Possible explanations we've looked at include:
    - The presence of a confounding variable (race of the victim)
    - Sampling bias (a systematic favoring of longer words)
    - Other types of bias (the placebo effect, leading questions, measurement bias)
    - Random chance
    - A real causal relationship

Ruling out Possible Explanations

- We've covered how careful planning, study design, and statistical analysis can rule out each of these explanations:
  - Confounding variables -> balance groups via **random assignment** or **stratification**

Ruling out Possible Explanations

- We've covered how careful planning, study design, and statistical analysis can rule out each of these explanations:
  - Confounding variables -> balance groups via **random assignment** or **stratification**
  - Sampling bias -> collect a **representative sample** using **simple random sampling**

Ruling out Possible Explanations

- We've covered how careful planning, study design, and statistical analysis can rule out each of these explanations:
  - Confounding variables -> balance groups via **random assignment** or **stratification**
  - Sampling bias -> collect a **representative sample** using **simple random sampling**
  - Other biases -> using a **placebo**, **blinding**, and *careful measurement*

Ruling out Possible Explanations

- We've covered how careful planning, study design, and statistical analysis can rule out each of these explanations:
  - Confounding variables -> balance groups via **random assignment** or **stratification**
  - Sampling bias -> collect a **representative sample** using **simple random sampling**
  - Other biases -> using a **placebo**, **blinding**, and *careful measurement*
  - Random chance -> we'll explore this in greater detail soon, but think about coin flips in the infant toy choice example
  - A real causal relationship -> this is what we want!

Putting it all Together

In this class I'd like you to be able to carry out the following data analysis steps:

1) Recognize the cases and variables in a dataset (this helps prevent things like the *ecological fallacy*)

Putting it all Together

In this class I'd like you to be able to carry out the following data analysis steps:

1) Recognize the cases and variables in a dataset (this helps prevent things like the *ecological fallacy*)
2) Understand the distributions and nuances of each variable (ie: studying univariate graphs and descriptive summaries)

Putting it all Together

In this class I'd like you to be able to carry out the following data analysis steps:

1) Recognize the cases and variables in a dataset (this helps prevent things like the *ecological fallacy*)
2) Understand the distributions and nuances of each variable (ie: studying univariate graphs and descriptive summaries)
3) Identify possible associations between variables (either finding them yourself by studying bivariate graphs and descriptive summaries, or identifying explanatory and response variables based upon a pre-defined research question)

Putting it all Together

In this class I'd like you to be able to carry out the following data analysis steps:

1) Recognize the cases and variables in a dataset (this helps prevent things like the *ecological fallacy*)
2) Understand the distributions and nuances of each variable (ie: studying univariate graphs and descriptive summaries)
3) Identify possible associations between variables (either finding them yourself by studying bivariate graphs and descriptive summaries, or identifying explanatory and response variables based upon a pre-defined research question)
4) Evaluate the plausibility of these associations (by considering whether the explanations on the prior slide can be ruled out or not)