CMPT 733 – Big Data Programming II

Statistics (I)

Instructor: Steven Bergner
Course website: https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages/
Slides by: Jiannan Wang & Steven Bergner
Why Should You Care?

There are three kinds of lies: lies, damned lies, and statistics

1. The Sample with the Built-in Bias  13
2. The Well-Chosen Average  29
3. The Little Figures That Are Not There  39
4. Much Ado about Practically Nothing  55
5. The Gee-Whiz Graph  62
6. The One-Dimensional Picture  68
7. The Semiattached Figure  76
8. Post Hoc Rides Again  89
9. How to Statisticulate  102
10. How to Talk Back to a Statistic  124
Simpson's paradox

Is UC Berkeley gender biased?

<table>
<thead>
<tr>
<th></th>
<th>Applicants</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8442</td>
<td>44%</td>
</tr>
<tr>
<td>Women</td>
<td>4321</td>
<td>35%</td>
</tr>
</tbody>
</table>
Simpson's paradox

Is UC Berkeley gender biased?

<table>
<thead>
<tr>
<th>Department</th>
<th>Men</th>
<th></th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
<td>Applicants</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
<td>108</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
<td>593</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
<td>375</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
<td>393</td>
</tr>
<tr>
<td>F</td>
<td>373</td>
<td>6%</td>
<td>341</td>
</tr>
</tbody>
</table>

NO!

Women tended to apply to competitive departments with low rates of admission.
Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics
Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics
Statistical Thinking

1. Data is just a sample

2. You goal is to infer a population

3. Think about how to go “backwards” from the sample to the population
Example 1. Image Classification

Is it a dog or a cat?

Dataset: 1000 images collected from the Web
Without Statistical Thinking

Treat the 1000 images as the population

> Train a model on the data
> Evaluate a model on the same data
> Model accuracy: 95%
With Statistical Thinking

What is the population?
- All the images in the Web

What is your dataset?
- A sample of 1000 images drawn from the Web

What should you do?
- Split the dataset into a training dataset and a test dataset
- Train the model on the training dataset
- Evaluate the model on the test dataset
Example 2. Market Trend Analysis

What will be the market share of electric vehicles by 2025?

Dataset: Analysis of 5 years of sales data from the automotive industry
Without Statistical Thinking

Misinterpreting a Small Sample as the Entire Market

> Count the number of people who intend to buy an electric vehicle, e.g., 60
> Count the number of people who intend to buy a gasoline vehicle, e.g., 40
> Incorrect Conclusion: Electric vehicles will represent 60% of all car sales
With Statistical Thinking

Understanding Market Predictions

What is the population?
- All the consumers in the market for new vehicles

What is your dataset?
- A sample of 1000 potential car buyers surveyed before a major auto show

Analysis result

Electric Vehicles: 60% ± 5%
Gasoline Vehicles: 40% ± 5%

Assumption: Consumer preferences remain consistent with the survey results until the auto show.
Summary

Statistical Thinking
- Sample, Population and Their Connection
- With vs. Without Statistical Thinking

Descriptive Statistics

Inferential Statistics
Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics
Descriptive vs. Inferential Statistics

Descriptive Statistics: e.g., Median
- **Why?** Aim to understand the data
- **How?** Data summarization, data visualization, etc.

Inferential Statistics: e.g., A/B Testing
- **Why?** Aim to use the data (i.e., sample) to learn about a population
- **How?** Estimation, confidence intervals, hypotheses testing, etc.
Exploratory Data Analysis (EDA)

Understand data and discover insights via data visualization, data summarization, etc.

Understand “Age” column

<table>
<thead>
<tr>
<th>Minimum</th>
<th>0.42</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-th Percentile</td>
<td>4</td>
</tr>
<tr>
<td>Q1</td>
<td>20.125</td>
</tr>
<tr>
<td>Median</td>
<td>28</td>
</tr>
<tr>
<td>Q3</td>
<td>38</td>
</tr>
<tr>
<td>95-th Percentile</td>
<td>56</td>
</tr>
<tr>
<td>Maximum</td>
<td>80</td>
</tr>
<tr>
<td>Range</td>
<td>79.58</td>
</tr>
<tr>
<td>IQR</td>
<td>17.875</td>
</tr>
</tbody>
</table>
Current EDA Solutions in Python

Solution 1: Pandas + Matplotlib

- Hard to Use
  - **Beginner:** Need to know how to write plotting code
  - **Expert:** Need to write lengthy and repetitive code

Understand “Age” column

<table>
<thead>
<tr>
<th>Age</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.42</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>17.875</td>
<td>1</td>
</tr>
</tbody>
</table>

- Write Code
- Write Code
- Write Code
Current EDA Solutions in Python

- Solution 2: Pandas-profiling
  - Slow
  - Hard to Customize

```python
profile = ProfileReport(df, title="Pandas Profiling Report")
```
Correlation Analysis

Correlation

- It is a measure of relationship between two variables

Why is correlation analysis useful?

- For understanding data better
- For making predictions better
Case Study: How to do correlation analysis

Height and weight are correlated

<table>
<thead>
<tr>
<th></th>
<th>height</th>
<th>weight</th>
<th>age</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>151.765</td>
<td>47.8256065</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>139.7</td>
<td>36.4858065</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>136.525</td>
<td>31.864838</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>156.845</td>
<td>53.0419145</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>145.415</td>
<td>41.276872</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>163.83</td>
<td>62.992589</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>149.225</td>
<td>38.2434755</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: Think Stats -- Exploratory Data Analysis in Python
Idea 1. Visualization
## Scatter Plot

<table>
<thead>
<tr>
<th></th>
<th>height</th>
<th>weight</th>
<th>age</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>151.765</td>
<td>47.8256065</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>139.7</td>
<td>36.4858065</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>136.525</td>
<td>31.864838</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>156.845</td>
<td>53.0419145</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>145.415</td>
<td>41.276872</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>163.83</td>
<td>62.992589</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>149.225</td>
<td>38.2434755</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>

![Scatter plot diagram](image)
Scatter Plot (with transparency)

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
<th>Age</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>151.765</td>
<td>47.8256065</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>139.7</td>
<td>36.4858065</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>136.525</td>
<td>31.864838</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>156.845</td>
<td>53.0419145</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>145.415</td>
<td>41.276872</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>163.83</td>
<td>62.992589</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>149.225</td>
<td>38.2434755</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>
# Hexbin Plot

![Graph showing a hexbin plot with data points distributed across the chart.](image)

<table>
<thead>
<tr>
<th>1</th>
<th>height</th>
<th>weight</th>
<th>age</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>151.765</td>
<td>47.8256065</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>139.7</td>
<td>36.4858065</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>136.525</td>
<td>31.864838</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>156.845</td>
<td>53.0419145</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>145.415</td>
<td>41.276872</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>163.83</td>
<td>62.992589</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>149.225</td>
<td>38.2434755</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>
Characterizing relationships

<table>
<thead>
<tr>
<th></th>
<th>height</th>
<th>weight</th>
<th>age</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>151.765</td>
<td>47.8256065</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>139.7</td>
<td>36.4858065</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>136.525</td>
<td>31.864838</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>156.845</td>
<td>53.0419145</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>145.415</td>
<td>41.276872</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>163.83</td>
<td>62.992589</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>149.225</td>
<td>38.2434755</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>
Idea 2. Correlation Coefficient
Covariance

Covariance is a measure of the tendency of two variables to vary together.

\[
\text{cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]
\]

\[
\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]
\]

Hard to interpret

113 kilogram-centimeters
Pearson’s correlation

Pearson’s correlation is a measure of the linear relationship between two variables

\[ \rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \]

Easy to Interpret
- \([-1, 0) \rightarrow \text{Negative Correlated}\]
- \((0, +1] \rightarrow \text{Positive Correlated}\]
- \(-1 \text{ or } +1 \rightarrow \text{Perfectly Correlated}\]

What about non-linear relationship?
Spearman’s rank correlation is a measure of monotonic relationship between two variables

\[ r_s = \rho_{r_X, r_Y} = \frac{\text{cov}(r_X, r_Y)}{\sigma_{r_X} \sigma_{r_Y}} \]

Advantages
- Mitigate the effect of outliers
- Mitigate the effect of skewed distributions
Summary

Statistical Thinking

Descriptive Statistics

- Descriptive vs. Inferential Statistics
- Exploratory Data Analysis with DataPrep
- Correlation Analysis

Inferential Statistics
Outline

Statistical Thinking

Descriptive Statistics

Inferential Statistics
  ◦ Estimation
Estimation

Problem statement
- Estimate a numerical value associated with a population

Examples
- Estimate the percentage of the people in the US who will vote for Biden
- Estimate the median annual income of all households in the US
Example: Median Annual Income

How to estimate the median annual income of all households in the US?

• Randomly select 10,000 households from the US
• Report their median annual income: 50,000 USD

• BUT, we need to report something like

  50,000 ± 500 USD
A Naïve Solution

• Randomly select 10,000 households from the US
• Report their median annual income

Repeat this process for 100 times

50,000  49,600  50,200  ⋮  49,200

You have to survey 1,000,000 million households in total!
A Smart Solution: Bootstrapping

Key Idea: Resampling
- Sample with replacement from the original data sample

Population: 1, 1, 8, 2, ... 3, 3
Sample: 3, 8, 1, 8, 3
Resample: 8, 3, 3, 3, 1
A Smart Solution: Bootstrapping

- Randomly select 10,000 households in Canada
- Draw a resample from the 10,000 households
- Report the median annual income of the resample

Repeat this process for 100 times

You do NOT need to survey any new household. 😊
Notes on Bootstrapping

• Start with a large random sample (at least 30)

• Replicate the resampling procedure as many times as possible (more than 1000 times)

• Does not work for min/max
Conclusion

Statistical Thinking
- Sample, Population and Their Connection
- With vs. Without Statistical Thinking

Descriptive Statistics
- Descriptive vs. Inferential Statistics
- EDA with DataPrep.eda
- Correlation Analysis

Inferential Statistics
- Estimation and Bootstrapping