ST1131

AY21/22 Sem 2

github.com/jasonqiu212

01. Exploratory Data Analysis

Types of Variables

- **Quantitative**: Discrete vs. Continuous
- **Categorical**: Ordinal vs. Nominal
- How to tell difference: Is distance between 2 points meaningful?

1 Variable

Frequency Table - Categorical

- **Proportion** - aka relative frequency. $\frac{\text{\# of obs. in 1 cat.}}{Total\#ofobs.}$
- **Modal Frequency** - Category with highest frequency
- Summarizing: Modal category and its proportion

Bar Plots - Categorical, Visual

- Summarizing: Modal category and its proportion, Categories with high/low proportions, Mention trends if ordinal

Histogram - Quantitative

- Summarizing: Gaps/Outliers, Unimodal/Bimodal/Multimodal, Symmetric/Skewed
- Skewed left: Left tail is longer. Skewed right: Right tail is longer.

Describing Center

- **Mean** - $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Linear Transformation: $\hat{Y} = b\hat{X} + a$
  - Sensitive to outliers, unlike median
- **Median** - $X_{(0.5)}$
- If $\bar{X} > X_{(0.5)}$, skewed right. If $\bar{X} < X_{(0.5)}$, skewed left

Describing Variability

- **Range** - Sensitive to outliers
- **Variance** - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Standard deviation** - $sd = \sqrt{S^2}$
  - Linear Transformation: $S_y^2 = b^2 s_x^2$ $S_y = |b| s_x$
- **Inter-quartile Range (IQR)** - $Q_3 - Q_1$
  - **Quantile** - (q_p) Value such that p of observations are below
  - Lower quartile (Q_1), Median (Q_2), Upper quartile (Q_3)
- If symmetric, mean and variance. If skewed, median and IQR.

Boxplot - Variability

- Includes: Min, Q1, Q2, Q3, Max
- **Outliers** - $< Q_1 - 1.5IQR$ or $> Q_3 + 1.5IQR$
- **Max/min Whisker Reach** - Boundary of outliers
- **Upper/lower Whisker** - Min/max obs. excluding outliers
- Does not show features of distribution. If unimodal, can show skewness.
- Summarizing: Median, Outliers, Compare medians and IQR if > 1 boxplots

2 Variables (Response Variable vs. Explanatory Variable)

2 Categorical Variables

Bar Plots

Contingency Table

- **Conditional Percentage** - % out of total
- **Join Percentage** - % out of some group. Use explanatory as group.
- Be careful of phrasing (Eg. Ppl w/o cancer of PMH users vs. PMH users of those w/o cancer)
- **Relative Risk** - Ratio of 2 percentages. (Eg. % of cancer in PMH users is 1.24 times the % of cancer in non-PMH users)

1 Categorical and 1 Quantitative

2 Boxplots - Split by categories

2 Quantitative Variables

Scatter Plot

- Summarizing: Pos./neg. association, Linear, Constant variability, Outliers

Correlation - $r \in [-1, 1]$

- 2 variables have same correalation, no matter $x \sim y$ or $y \sim x$
- Correlation is linear, when $r = \pm 1$

02. Data Collection

- **Confounding Variable** - Related to exp. and resp. variable. Confounds their association. Observed.
- **Lurking Variable** - Unobserved
- **Experimental Study** - Assign subjects to treatments and observe response variable
  - Pros: Control over lurking variables
  - Cons: Costly, Unethical
- **Observational Study** - Explanatory and response variable observed for subjects. No treatments.

Sample Survey

1. Identify population
2. Compile **sampling frame** - Where sample is from
3. **Sampling design** - How to choose subjects from sampling frame
   - **Simple Random Sample** - Each sample has same chance of being chosen

Sources of Bias in Sample Survey:

- **Sampling Bias** - Sample not random or undercoverage
- **Non-response Bias** - No response from subject
- **Response Bias** - Subject does not answer truthfully

Elements of Good Experimental Study:

- Control comparison group
- Randomization: Eliminate lurking variables
- Blinding the study: Placebo

03. Probabililty

Axioms of Probability

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. If A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$. $P(A \cap B) = 0$
4. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

For any events A and B:

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) = P(A \cap B) + P(A \cap B^c)$
- A and B are **independent** if $P(A \cap B) = P(A)P(B)$

Conditional Probability

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Law of Total Probabililty

$P(A) = P(A \cap B_1) + ... + P(A \cap B_n)$

Bayes' Theorem

$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1)...P(A|B_n)P(B_n)}$

Epidemiological Terms

- **Sensitivity** - Given person has disease, prob. of positive test
- **Specificty** - Given person has no disease, prob. of negative test
- **Prevalence** - No. of people with disease / Total population

04. Random Variables

05. Sampling Distribution

06. Confidence Intervals

07. Hypothesis Testing

08. Linear Regression

09. R Code

```
test(2)
```