Descriptive Statistics

Javier Tasso

University of Pennsylvania

Introduction

Basic Concepts

- Descriptive vs Inferential Statistics.
- Population vs Sample.
- Unit of analysis.
- Variable. Types:
    - Qualitative.
    - Quantitative.

Describing One Variable

Frequency Distributions and Graphs

- Example.
- (Empirical) Cumulative Distribution Function (cdf).
- Histograms.
- Polygons.
- Boxplots.

Measures of Center

- Sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + \cdots + X_n}{n}$$

- Trimmed mean: trim 10% of the data to reduce the impact of outliers.
- Sample median: the value in the $(\frac{n+1}{2})^{\text{th}}$ position of sorted data.
- Quartiles: the values in the $(\frac{n+1}{4})^{\text{th}}$, $(2 \cdot \frac{n+1}{4})^{\text{th}}$, and $(3 \cdot \frac{n+1}{4})^{\text{th}}$ positions. First, second, and third quartile.

Measures of Variability

- Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

  - Average the squared deviations from the sample mean.
  - Why squared deviations?
  - Shortcut formula: $S^2 = \frac{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}{n-1}$
  - Units.
  - Sum of squares: $S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2$
  - $\bar{X}$ minimizes sum of squared deviations.
- Standard deviation: $S = \sqrt{S^2}$.
- Interquartile range: $\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$.
- Outliers: $X < \mathrm{Q}_1 - 1.5\mathrm{IQR}$ or $X > \mathrm{Q}_3 + 1.5\mathrm{IQR}$.

Measures of Shape

- Z-scores:

$$Z_i = \frac{X_i - \bar{X}}{S}$$

- Mean and variance of Z_i are 0 and 1.
- Skewness:

$$\text{Skew} = \frac{\sum_{i=1}^{n} Z_i^3}{n} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{nS^3}$$

- Kurtosis:

$$\text{Kurt} = \frac{\sum_{i=1}^{n} Z_i^4}{n} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{nS^4}$$

Describing Two Variables

Scatter-plots

- Describe each individual variable. Variables X and Y.
- Analyze their association (if any) by constructing an scatter-plot.
    - Positive correlation.
    - Negative correlation.
    - No correlation.
    - Non-linear associations.
- Example.

Covariance and Correlation

- Covariance:

$$\mathrm{COV}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

  - Covariance of X with itself: $\mathrm{COV}(X,Y) = S_X^2$.
  - Shortcut formula $\mathrm{COV}(X,Y) = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{n-1}$
  - Positive, zero, or negative depending on the type linear association between the variables.

- Correlation:

$$r = \frac{\mathrm{COV}(X,Y)}{S_x S_y}$$

  - $-1 \leq r \leq 1$
  - Closer to 1 (or -1) when the linear association is strong.

Linear Fit

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Fit a linear function between the two variables according to some criteria.
- Most common criteria: minimize the discrepancy between $\hat{Y}_i$ and observed Y_i.

$$\hat{\beta}_1 = \frac{\mathrm{COV}(X, Y)}{S_X^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- Relationship between r and $\hat{\beta}_1$: $\hat{\beta}_1 = \frac{S_X}{S_Y} \cdot r$.