Lecture 2: Introduction to Statistical Methodology

Taeyong Park

Carnegie Mellon University in Qatar

September 4, 2018

Today

- Introduction to statistical methodology.
- Install R and RStudio.
- Introduction to R.

Introduction to Statistical Methodology

- *Statistics* consists of a body of **methods for obtaining and analyzing data**.

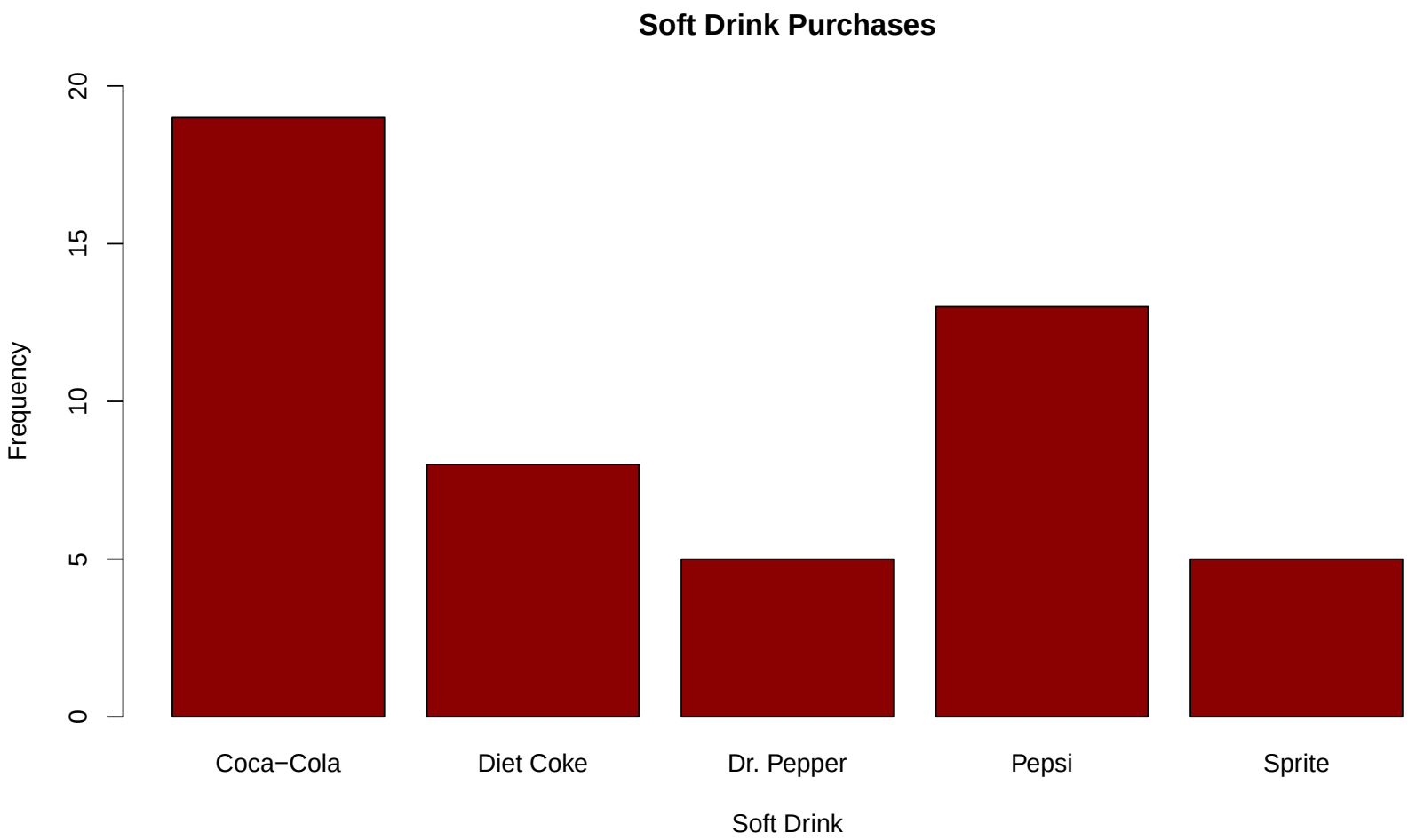
Introduction to Statistical Methodology

- *Statistics* consists of a body of **methods for obtaining and analyzing data**.
  - Data: the observations gathered on the characteristics of interest
    - ★ Survey data: e.g. Q: "How do you feel about the economy in your country?" A: "1) Getting better; 2) Stay about the same; 3) Getting worse."
    - ★ Visa Card's payment transactions (6,800 transactions per second).
    - ★ Time series data of crude oil prices; Time series data of macroeconomic indicators.
    - ★ Text data from Twitter mentions.
    - ★ ...

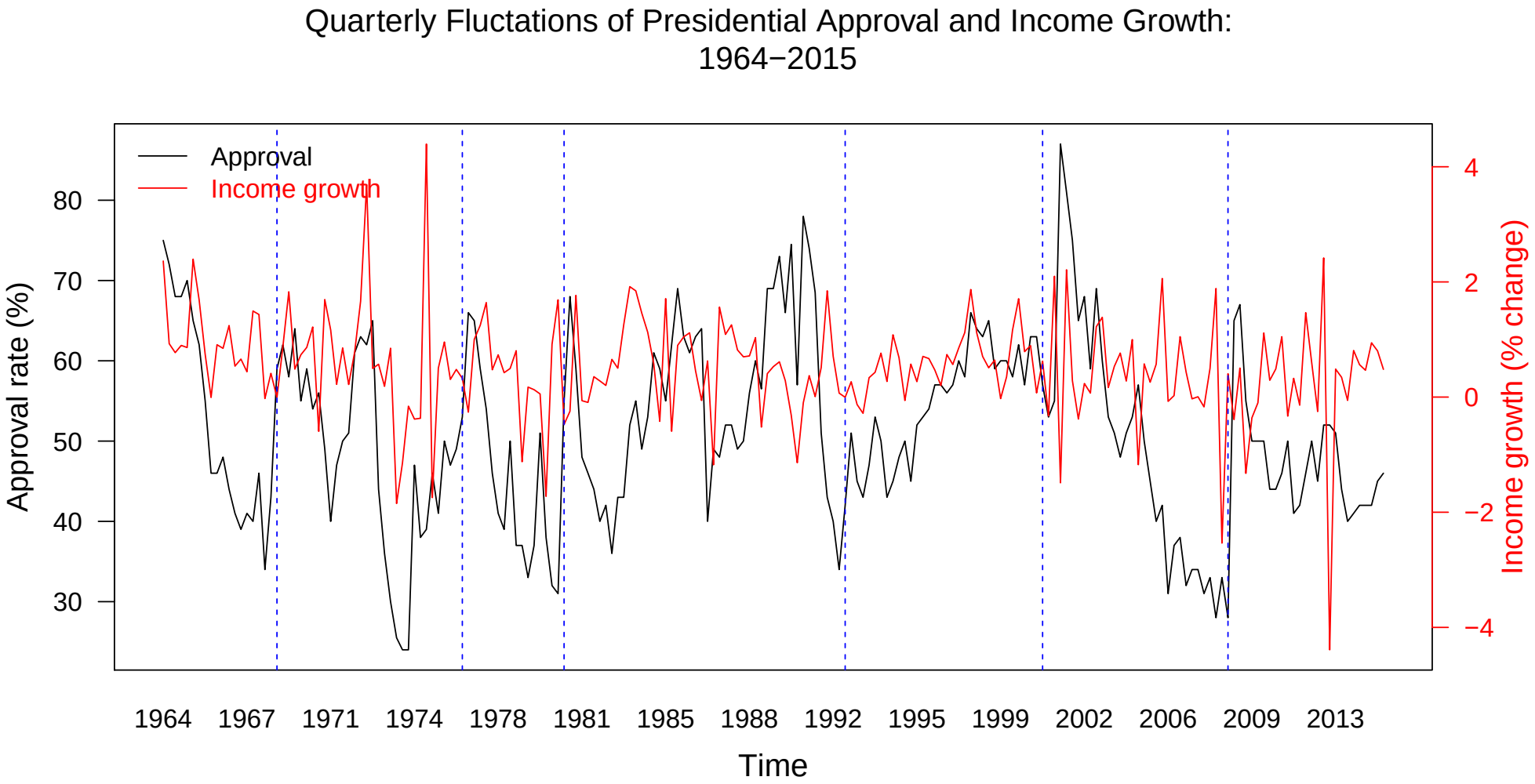
Data: Example

Soft Drink	Frequency	Proportion	Percentage
Coca-Cola	19	.38	38
Diet Coke	8	.16	16
Dr. Pepper	5	.10	10
Pepsi	13	.26	26
Sprite	5	.10	10
Total	50	1.00	100

Data: Example



Data: Example



Introduction to Statistical Methodology

- *Statistics* consists of a body of methods for obtaining and analyzing data.
  - Methods for gathering data for research (Design).
  - Methods for summarizing the data (Description).
  - Methods for making predictions based on the data (Inference).

Introduction to Statistical Methodology

- *Statistics* consists of a body of methods for obtaining and analyzing data.
  - Methods for gathering data for research (Design).
  - Methods for summarizing the data (Description).
  - Methods for making predictions based on the data (Inference).

Methods for analyzing data: Description and Inference

- Descriptive statistics and inferential statistics
  - **Descriptive statistics** summarize the information in a collection of data.
    - ★ Graphs, tables, and numbers such as averages and percentages.
  - **Inferential statistics** provide predictions about a population based on data from a sample of that population.
    - ★ **Population**: The total set of subjects of interest in a study.
    - ★ **Sample**: The subset of the population on which the study collects data.

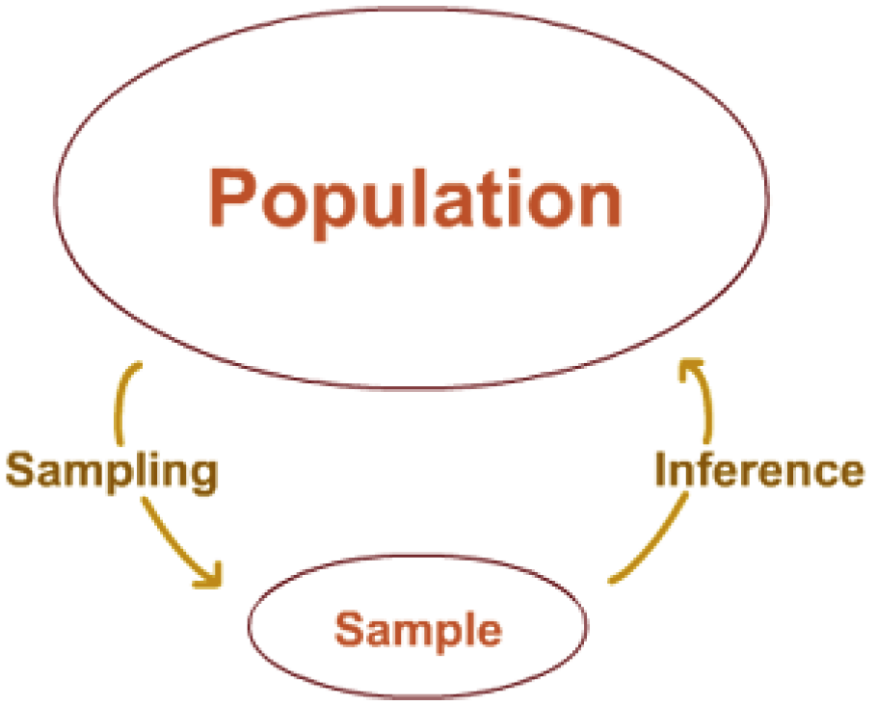
Inferential Statistics

- Statistical Methodology focuses on inferential statistics.
  - How to learn and/or make predictions about a **population** based on a **sample** data set.

- Purpose: Estimate the parameter of interest.
  - **Parameter**: Numerical summary of the population.
  - **Statistic**: Numerical summary of the sample data.

- Example: Using survey data containing 1,000 respondents of Qatar residents to draw inference for the average age of the Qatar population.
  - **Parameter**: The average age of the Qatar population.
  - **Statistic**: The average age of the 1,000 respondents.

Inferential Statistics



Is Statistical Inference Accurate (Unbiased)?

- We need to avoid biased samples.
  - What if we draw a sample in Education City for the sample of 1,000 residents?

Is Statistical Inference Accurate (Unbiased)?

- We need to avoid biased samples.
  - What if we draw a sample in Education City for the sample of 1,000 residents?
- Subjects be chosen by **randomization**. $\Rightarrow$ Good sample representation.

Random sampling

A method of sampling for which every possible sample has **equal chance of selection**.

Why Random Sampling?

Random sampling

A method of sampling for which every possible sample has **equal chance of selection**.

- True parameter for the population - Inference based on a sample = **Bias**
- Random samples A, B, ..., Z $\subset$ *Pop*. :
  - Inference based on sample A = based on sample B = ... = based on Z = based on Pop.
- True parameter - Inference based on sample A = **Bias** $\approx$ **0**

How to Sample Randomly?

Sample 2,000 residents in Qatar to conduct a survey (1,000 / 2 mil.)

- In theory, assign every resident a unique number from 1 to 2 mil.
- Choose 1,000 numbers from a box. Or, generate 1,000 random numbers in R.
- The probability that each resident is sampled $= \frac{1,000}{2,000,000}$.
- Samples A, B, ..., Z $\subset$ *Pop*. have **equal chance of selection**.

Random Sampling and Potential Problems

In practice?

- Not possible to give everyone a unique number to select random respondents.
- A typical method of sampling in surveys: Telephone interviews obtain the sample with random digit dialing.
  - A sample obtained might not be representative of the population and yield bias. Why?

Random Sampling and Potential Problems

- An error may occur whenever we use a sample to predict the population parameter $\Rightarrow$ **Sampling error**.

Random Sampling and Potential Problems

- An error may occur whenever we use a sample to predict the population parameter $\Rightarrow$ **Sampling error**.
- Land-line telephone users vs. Cell-phone only users $\Rightarrow$ **Sampling bias**.

Random Sampling and Potential Problems

- An error may occur whenever we use a sample to predict the population parameter $\Rightarrow$ **Sampling error**.
- Land-line telephone users vs. Cell-phone only users $\Rightarrow$ **Sampling bias**.
- Incorrect answers; Shy Trump supporters $\Rightarrow$ **Response bias**.

Random Sampling and Potential Problems

- An error may occur whenever we use a sample to predict the population parameter $\Rightarrow$ **Sampling error**.
- Land-line telephone users vs. Cell-phone only users $\Rightarrow$ **Sampling bias**.
- Incorrect answers; Shy Trump supporters $\Rightarrow$ **Response bias**.
- Those who respond vs. those who refuse to respond $\Rightarrow$ **Nonresponse bias**.