

# 1 What is Statistics?

## 1.1 Introduction

- Example: Past presidential election- who would win?
- The study of statistics is concerned with the design of experiments or sample surveys to obtain a specified quantity of information at minimum cost and the optimum use of this information in making an inference about a population.
- The objective of statistics is to make an inference about a population based on information contained in a sample from that population and to provide an associated measure of goodness for the inference

## 1.2 Characterizing a set of measurements: graphical methods

- The response or dependent variable. E.g. the yield of a stock in a year.
- Independent variables. E.g. various features of a company like P/E ratio, market cap etc.
- Use the information in the sample to infer the relationship between the two. Want to determine optimum condition for maximizing profit (“which stock to buy?”)
- The population is represented by a distribution of the profit measurements, with the form of the distribution depending on specific values of the independent variables.
- An individual population or any set of measurements can be characterized by a relative frequency distribution, which can be represented by a relative frequency histogram.
  - A few rule of thumbs for creating histogram by hand
    - Spanning the range of the data with 5~20 intervals and using the larger number of intervals for larger quantities of data
  - Or use R, minitab, matlab, etc.
  - “Area under the curve” – a naïve concept of probability
- Relative frequency distribution of profit for a conceptual population of profit responses at a given settings of the independent variables
  - “Area under the curve”

## 1.3 Characterizing a set of measurements: numerical methods

- Numerical descriptive measures of a set of data
- Two types:
  - Measures of central tendency
  - Measures of dispersion or variation
- The arithmetic mean (Measures of central tendency)
  - The mean of a sample of  $n$  measured responses  $y_1, \dots, y_n$  is...
  - The corresponding population mean is denoted  $\mu$  (unknown)
- Measures of dispersion or variation
  - The variance of a sample of measurements  $y_1, \dots, y_n$  is the sum of the square of the differences between the measurements and their mean, divided by  $n-1$ .
  - The corresponding population variance is denoted by  $\sigma^2$ .
- The standard deviation of a sample of measurements is the positive square root of the variance.
  - The corresponding population standard deviation is denoted  $\sigma$ .
  - Has the same scale as the original measurements.
- Many distributions of data in real life are mound- or bell-shaped. They can be approximated by a bell-shaped frequency distribution known as a normal curve. They have definite characteristics of variation (“empirical rule”)

- The interval with endpoints
- $\mu \pm \sigma$  contains  $\sim 68\%$  of the measurements
- $\mu \pm 2\sigma$  contains  $\sim 95\%$  of the measurements
- $\mu \pm 3\sigma$  contains almost all of the measurements
- E.g. scores on a test  $\sim N(64, 10^2)$
- Approximately 68% of the scores are between ...
- Approximately 95% of the scores are between ...
- Almost all of the scores are between ...
- If a single high school student is randomly selected from the students, what's the probability that his score will be between 54 and 74?

## 1.4 How inferences are made

Presidential election example again.

20 samples and all 20 favor Kerry. What do we conclude?? What's the thought process?

Impossible vs improbable.

Importance of probability theory as a mechanism for making statistical inferences.

Going from sample to probability.

Theory and Reality