STAT206

Andy Zhang

Fall 2014

Contents

Chapter 1

Introduction

1.1 Statistics

1.1.1 Definitions

Statistics Collection, organization, analysis, interpretation and presentation of data. It is also defined as the quantification of uncertainty.

Unit A single element, usually a person or object, whose characteristics are of interest. Ex: A student enrolled in the course.

Population The set of all units which are of interest. Ex: All students enrolled in the course

Variable A measurement of the characteristic of interest from a unit. Ex: Number of Canadian provinces visited by a student

Sample A subset of units from the population for which measurements of the desired variable are actually made. Ex: 29 students chosen from the class

Descriptive Statistics Summarize the data in the sample, both graphically and numerically

Inferential statistics USe the sample data to estimate an attribute of the population. Include a quantification of uncertainty

Sampling Error An error which occurs due to the uncertainty in randomly selecting a sample.

Study error A systematic error which occurs because the sample does not accurately represent the population

1.1.2 Process

Identify the problem of interest

- Who or what do you want to learn about?
  - Define the **population** of interest

  - Individual elements of the population are called **units**
- What research question would you like answered?
  - Define your **hypothesis**

Plan the data collection

- How will you select a subset of **units** from the **population** to be in your **sample**?
  - How large will the **sample** be?
- What is (are) the **variable (s)** of interest?
  - How will you measure it (them)?

Analyze the data

- Graph the data — histogram, scatter-plot, etc
- Compute **Descritive statistics** — e.g. sample mean, sample variance, etc.
- Compute **Inferential statistics** — e.g. confidence intervals, hypothesis tests about population **parameters**
  - Inferential statistics include a quantification of the sampling error

Draw conclusions

- Use the results of your analysis to address the original research question
- Address limitations of the study, especially any potential systematic **study errors**

1.1.3 Data Types

Categorical Variable A qualitative measure. Each unit belongs to **one of K** possible classes.

Discrete variable A quantitative measure. Each unit's measurement can take on one of a **countable** number of possible values

Continuous variable A quantitative measure. Each unit's measurement can take on an **uncountable** number of possible values, usually some interval of real numbers

1.1.4 (Grouped) Frequency Tables

- Display the number of units which are in each class
- Discrete / Continuous variables are grouped into classes
- In the case of numerical variables, there is a loss of information

See more: `http://en.wikipedia.org/wiki/Stem-and-leaf_display`

1.1.5 Stem and Leaf Plot

- A **stem-and-leaf plot** is a way to summarize a relatively **small** data set, without the loss of information that occurs with a frequency table
- Left is possible **first** digits, right is remaining digits in ascending order

See more: `http://en.wikipedia.org/wiki/Stem-and-leaf_display`

1.1.6 Bar Chart

- Bar charts are used to graphically display information from categorical variables

See more: `http://en.wikipedia.org/wiki/Bar_chart`

1.1.7 Histogram

- A histogram is similar to a bar chart, but it's for numerical data
- The range is divided in distinct classes, and each observation is assigned to exactly one class
- Histogram shows frequency of observations in each class

See more: `http://en.wikipedia.org/wiki/Histogram`

- If class ranges are not same length, we can use density histogram instead
- When interpreting a density histogram, it is the area that is meaningful
- Height is $height = \frac{relative frequency}{width} = \frac{frequency}{width*n}$

See more `http://en.wikipedia.org/wiki/Histogram`

1.1.8 Measures of Centrality

- The **sample mean** of a set of n values, $x_1, x_2, x_3, \ldots, x_n$ denoted by $\bar{x}$ is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- The **median** is the number x^* such that half of the observed values are below x^* and half are above
- If after writing our values in ascending order, we donte the i^{th} value as $x_{(i)}$, then

$$x^* = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})} & \text{if } n \text{ is even} \end{cases}$$

1.1.9 Measures of Variability

Measures of variability

- The **sample variance** of a set of values $x_1, x_2, x_3, \ldots, x_n$ denoted by s^2 is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- The **sample standard deviation** denoted s, is the square root of the sample variance
- The **range** of the set is the difference between the maximum and minimum value

$$range = x_{(n)} - x_{(1)}$$

1.1.10 Box Plot

- The box indicates the middle 50% of the observations, i.e. the second and third quartiles
- The line through the box indicates the median observation
- The whiskers indicate the highest and lowest observations

See more: `http://en.wikipedia.org/wiki/Box_plot`