Lecture 5

Descriptive Statistics III

Text: Chapter III

STAT 8010 Statistical Methods I
August 30, 2019

Whitney Huang
Clemson University

Notes

Agenda

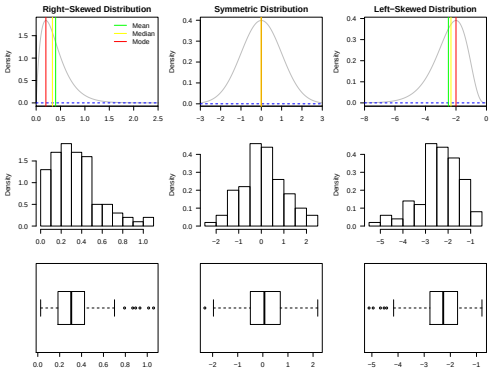
Notes

Last Lecture

- Graphical Summaries of **Quantitative Variables**
  - Dot plot
  - Stem-and-Leaf Plot
  - Histogram
  - Boxplot
- Numerical Summaries of **Quantitative Variables**
  - Mean $\Rightarrow$ A measure of **central tendency**

Notes

Shapes of Distributions



5.4

Notes

Measures of Center

- A **measure of center** attempts to report a “typical" value for the variable
- When a measure of center is calculated with **sample data** it is a **statistic**
- When a measure of center is calculated with popular (e.g., census data) it is a **parameter**
- **Measures:** Mean, Median, Mode

5.5

Notes

Mean

- The **population mean**, denoted by μ_X, is the sum of all the population values ($\{X_i, \cdots, X_N\}$) divided by the total number (N) of population values. That is,

$$\mu_X = \frac{\sum_{i=1}^{N} X_i}{N}$$

- The **sample mean**, denoted by $\bar{X}$ is the sum of all the sample values ($\{X_1, \cdots, X_n\}$) divided by the total number of sample values (n). That is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

5.6

Notes

Median

The median is the value separating the higher half from the lower half of a data sample

How to compute the median: Order the n observations in a data set from smallest to largest, then

$$\text{Median} = \begin{cases} \text{the single middle value,} & \text{n odd} \\ \text{the average of the middle two values,} & \text{n even} \end{cases}$$

Notes

Mode

The mode is the value of the observation that appears most frequently

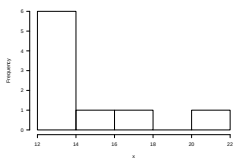
How to compute the mode(s): Order the observations in a data set from smallest to largest, then find the number that is repeated more often than any other

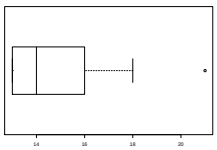
Notes

Example

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Plot this “data set" and describe the shape of the distribution





Notes

Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13+18+13+14+13+16+14+21+13}{9} = 15$$

- Find the sample median
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
  2. Compute the sample size n and identify (or compute) the median value
  3. $n = 9 \Rightarrow$ the median is the 5th number, which is 14

Notes

Example cont'd

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
  2. We have 3 13 and 2 14 $\Rightarrow$ 13 is the mode

Notes

Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^{9} \frac{13+18+13+14+13+16+14+210+13}{9} = 36$$

- Find the sample median
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  2. Compute the sample size n and identify (or compute) the median value
  3. $n = 9 \Rightarrow$ the median is the 5th number, which is (still) 14

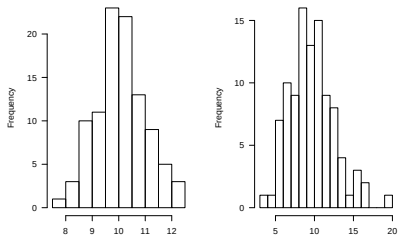
Notes

Example cont'd

- Find the mode
  1. Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  2. We have 3 13 and 2 14 $\Rightarrow$ 13 is (still) the mode

Notes

Measures of Spread



- **Measures:** Range, Variance(Standard Deviation), Interquartile range (IQR)

Notes

Range

The range of a dataset is the difference between the largest and smallest values

$$\text{Range} = \text{Largest Value} - \text{Smallest Value}$$

- Compute the range of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute the range of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

Question: Is Range a robust statistic?

Notes

Standard Deviation (Variance)

- The sample standard deviation (variance), denoted by s (s^2), can be used to estimate the population standard deviation (varaince), denoted by σ (σ^2).
- s is calculated in the following way:
  1. Calculate the sample mean $\bar{x}$
  2. Calculate the deviation (from the sample mean) for each observation (i.e., $x_i - \bar{x}, \quad i = 1, , \cdots, n$)
  3. Square each deviation and add them (i.e., $\sum_{i=1}^{n}(x_i - \bar{x})^2$)
  4. Divide by $n-1$ and take the square root (i.e., $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$)

Notes

Example

- Compute s of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute s of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

Question: Is standard deviation a robust statistic?

Notes

Interquartile range (IQR)

- IQR $= Q_3 - Q_1$, where Q_1 is the Lower Quartile (the median of the lower half of the data) and Q_3 is the Upper Quartile (the median of the upper half of the data)
- Compute the IQR of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute the IQR of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

Question: Is IQR a robust statistic?

Notes