1. MEASURES OF CENTER

Prof. Ryung Kim
ryung.kim@einstein.yu.edu

Algebraic Notation

- *n* observations of a variable *y*:

$$y_1, y_2, \cdots, y_n$$

- Ordered observations:

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$$

Mean

- Definition

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Read it 'y-bar'
- Mean is not '**robust**' to outliers

Example: Numerical Summary of Blood Sugar Reading

- Eight blood Sugar Readings (mg/dL)

120
140
120
118
125
140
90
89

Mean

= (89+90+118+120+120+125+140+140)/8

= 117.75 (mg/db)

Robust?

Median

- Definition

$$\text{median}(y) = \begin{cases} y_{\left(\frac{n+1}{2}\right)} & \text{if n is an odd number} \\ \frac{1}{2}\left(y_{(n/2)} + y_{(n/2+1)}\right) & \text{if n is an even number} \end{cases}$$

- Median is '**robust**' to outliers

Example: Numerical Summary of Blood Sugar Reading

- Eight blood Sugar Readings (mg/dL)

120
140
120
118
125
140
90
89

89, 90, 118, 120, 120, 125, 140, 140

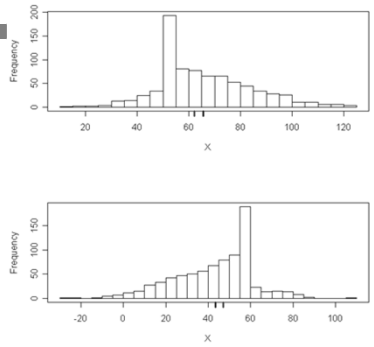
Median =

Robust?

7

Find mean and median.

"Mean follows the tail."



2. MEASURES OF VARIATION (SPREAD)

Prof. Ryung Kim
ryung.kim@einstein.yu.edu

Range

9

- Eight blood Sugar Readings (mg/dL)

120
140
120
118
125
140
90
89

Range = maximum-minimum

= 140-89 = 51 (mg/db)

Standard Deviation and Variance

10

- Variance:

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

- Standard Deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Example: Blood Sugar Reading (cont.)

11

- Eight blood Sugar Readings (mg/dL)

120
140
120
118
125
140
90
89

Variance = 378.5 (mg/db)^2

Standard Deviation = 19.455 (mg/db)

Intuition for Standard deviation

12

- To roughly estimate the standard deviation from a symmetric data, use

  s is close to range/4

  - This is a very rough approximation.

Percentile & Quartiles

13

- 25^{th} percentile
  - separates the smaller 25% of the sorted values from the bigger 75%
- 50^{th} percentile = **median (M)**
  - separates the smaller 50% of the sorted values from the bigger 50%
- 75^{th} percentile
  - separates the smaller 75% of the sorted values from the bigger 25%

This can only be a rough definition because 25% of the number of all observations may not be a natural number.

k^{th} percentile - More accurate definition

14

1. $p_0 = Y_{(1)}$, $p_{100} = Y_{(n)}$
2. $p_k = (Y_{(nk/100)} + Y_{(nk/100+1)})/2$ *if nk/100 is an integer*
   $= Y_{(m)}$ *if nk/100 is not an integer.*
   - *m is the smallest integer that is greater than nk/100*

- There are different versions of definitions.

Example: Blood Sugar Reading (cont.)

15

- Eight blood Sugar Readings (mg/dL)

120
140
120
118
125
140
90
89

89, 90, 118, 120, 120, 125, 140, 140

25^{th} percentile = 104.0 (mg/db)

30^{th} percentile = 118.0 (mg/db)

Median = 120.0 (mg/db)

75^{th} percentile = 132.5 (mg/db)

Interquartile Range (IQR)

16

- Definition
  - The **interquartile range (IQR)** is the difference between 75^{th} percentile and 25^{th} percentile
    IQR = Q3-Q1
- The IQR is the range the middle 50% of the data.
- In previous example,
  IQR is 132.5-104.0 = 28.5 (mg/db)