

Lecture 2 Location, Dispersion and Boxplot

Notation: x_1, x_2, \dots, x_n

Example 1 (continued):

63.0, 64.1, 65.2, 66.1, 64.3, 66.2, 67.1, 62.5,
68.5, 66.3, 66.7, 70.1, 66.6, 69.4, 67.1

Mean (sample mean): arithmetic average.

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Interpretations:

Drawback:

Alternative measures?

Median (sample median): middle point, denoted by \tilde{x}

Steps to derive \tilde{x} :

1. ordering the observations from smallest to largest.

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

2. if n is odd, \tilde{x} = the single middle value

if n is even, \tilde{x} = the average of the two middle values

Median divides the sample into two subsamples:

if n is even:

$$\{x_{(1)}, x_{(2)}, \dots, x_{(\frac{n}{2})}\} \text{ and } \{x_{(\frac{n}{2})+1}, \dots, x_{(n)}\}$$

if n is odd:

$$\{x_{(1)}, x_{(2)}, \dots, x_{(\frac{n+1}{2})}\} \text{ and } \{x_{(\frac{n+1}{2})}, \dots, x_{(n)}\}$$

Lower subsample and upper subsample

Other measures of location:

Quartiles:

first quartile (lower fourth), Q_1 :

median of lower subsample

second quartile, Q_2 :

median of the whole sample

third quartile (upper fourth), Q_3 :

median of upper subsample

InterQuartile Range (IQR) (fourth spread):

$$\text{IQR} = Q_3 - Q_1$$

Percentiles: be discussed later

Measures of dispersion

An extreme example:

sample 1: -100, -80, -40, -20, 0, 20, 40, 80, 100

sample 2: -10, -8, -4, -2, 0, 2, 4, 8, 10

sample 3: -1, -.8, -.4, -.2, 0, .2, .4, .8, 1

Conclusion:

Sample: x_1, x_2, \dots, x_n ; sample mean: \bar{x}

Deviations:

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$$

Sample Variance s^2 :

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

Sample standard deviation s :

$$s = \sqrt{s^2}$$

Why $n - 1$ instead of n ?

A computing formula for s^2 :

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Proposition:

Let x_1, x_2, \dots, x_n be a sample and c be any nonzero constant.

1. if $y_1 = x_1 + c, y_2 = x_2 + c, \dots, y_n = x_n + c$, then

$$s_y^2 = s_x^2$$

2. if $y_1 = cx_1, y_2 = cx_2, \dots, y_n = cx_n$, then

$$s_y^2 = c^2 s_x^2$$

where s_x^2 is the sample variance for the x 's and s_y^2 is the sample variance for the y 's.

Steps to construct boxplot:

1. Q_1 , Q_2 (median), Q_3 and IQR.
2. smallest and largest observations in $[Q_1 - 1.5\text{IQR}, Q_3 + 1.5\text{IQR}]$

Suppose they are x_l and x_u .

3. mild outliers, i.e., observations in $[Q_1 - 3\text{IQR}, Q_3 - 1.5\text{IQR}]$ or $[Q_1 + 1.5\text{IQR}, Q_3 + 3\text{IQR}]$.
4. extreme outliers, i.e. observations less than $Q_1 - 3\text{IQR}$ or larger than $Q_3 + 3\text{IQR}$.
5. construct a rectangle above a horizontal axis with left edge (Q_1), right edge (Q_3), a inside vertical line (median).
6. Draw a whisker from Q_1 to x_l , and a whisker from Q_3 to x_u
7. Plot mild outliers with solid dots, extreme outliers with circles.