Measures of Location

Given sample data $\{x_1, x_2, \ldots, x_n\}$

Want: a "number" that best locates points in the dataset.

① Sample Mean

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

= "arithmetic mean/average".

Note: The "true mean" for the population is denoted by μ.

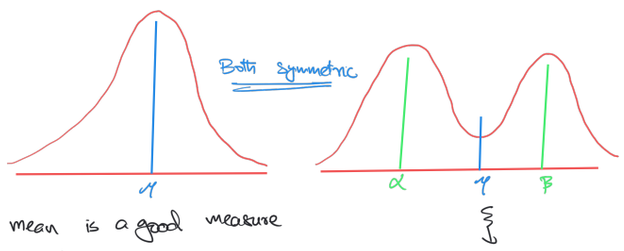
μ is fixed and usually unknown

want to use $\bar{x}$ to estimate μ.

might not always be a good idea!!

Applicability of the Mean

① mean (sample/population) is a good measure of location if data is symmetric and unimodal.



mean is a good measure of location

mean is not a good measure of location.

② sample mean is sensitive to individual data points and outliers.

mean is "pulled" in the direction of outlying points

not a good measure of location when data has "outliers".

② Percentiles

Let $p \in (0,1)$ ⇝ the $100 \cdot p^{th}$ percentile for the sample data $\{x_1, x_2, \ldots, x_n\}$ is a number satisfying

$(100 \cdot p)\%$ of data less than this number.

Important Percentiles:

① Median ⇝ 50^{th} percentile.

Given a sample $\{x_1, x_2, x_3, \ldots, x_n\}$

to calculate sample median $\tilde{x}$

a) order the dataset in increasing order

b) $\tilde{x} = \begin{cases} \left(\frac{n+1}{2}\right)^{th} \text{ ordered value if } n \text{ is odd.} \\ \text{avg of } \left(\frac{n}{2}\right)^{th} \text{ and } \left(\frac{n}{2}+1\right)^{th} \text{ ordered values if } n \text{ is even.} \end{cases}$

The true population median is denoted as $\tilde{\mu}$

② Quartiles:

Q1 ⇝ 1st Quartile ⇝ 25^{th} percentile

Q2 ⇝ 2nd Quartile ⇝ 50^{th} percentile

Q3 ⇝ 3rd Quartile ⇝ 75^{th} percentile.

Note: a) $\tilde{x} = Q3$

b) the quartiles are robust to minor changes in data values.

c) robust to outliers.

d) have poor arithmetic properties.

③ Trimmed Means

Combines arithmetic properties of the mean and the robustness of percentiles

Let $k \in (0, 50)$.

k% trimmed mean := $\bar{x}_k$:= avg of data values between k^{th} and $(100-k)^{th}$ percentiles.

Note: $\bar{x}_0 = \bar{x}$ = sample mean

$\bar{x}_{50} = \tilde{x}$ = sample median.