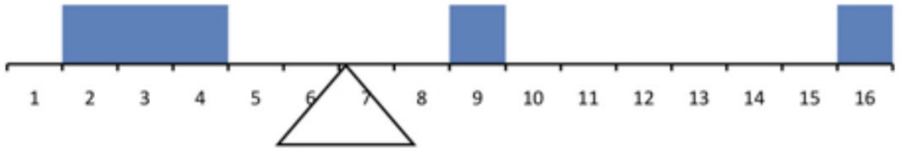
Ch 3. Summarizing Distributions

- Central Tendency: mean, median & mode
- Other Measures of Central Tendency
- Comparing Central Tendency
- Measures of Variability: Range, IQR, Variance
- Linear Transformation of variable
- Variance Sum Law I

Central Tendency

What is a "central" location of a distribution?

- Idea 1: Find the "balance point" of the distribution



What is the location of the triangle that "balances" unit masses at points: X={2,3,4,9,16}?

- Let μ be the location.

- We need: $$\sum_{X_i<\mu}(\mu - X_i) = \sum_{X_i \geq \mu}(X_i - \mu)$$

$$N\mu = \sum X_i$$

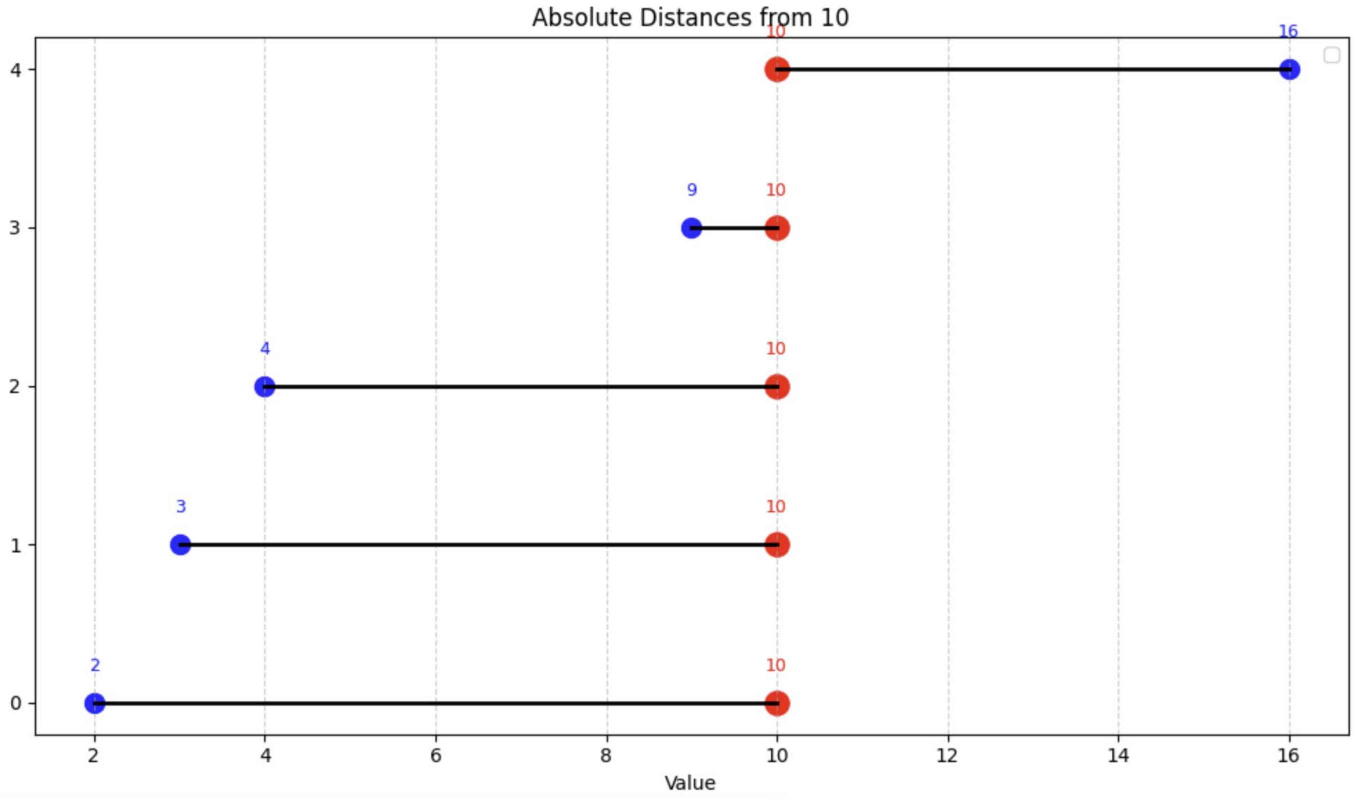
$$\boxed{\mu = \frac{1}{N}\sum X_i}$$

Answer: distribution mean

Central Tendency

What is a "central" location of a distribution?

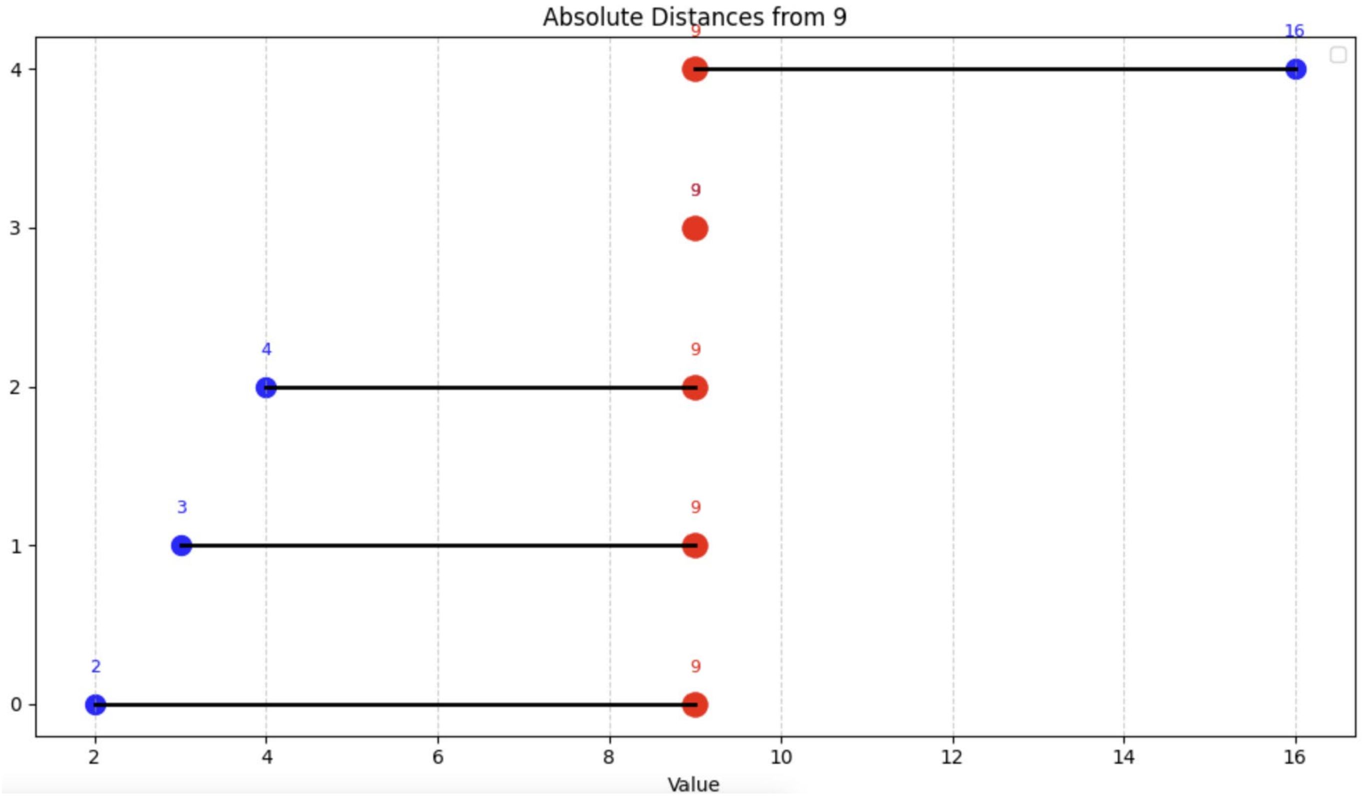
- Idea 2: Find the point of smallest absolute deviation (from it).



▪ Central Tendency

What is a "central" location of a distribution?

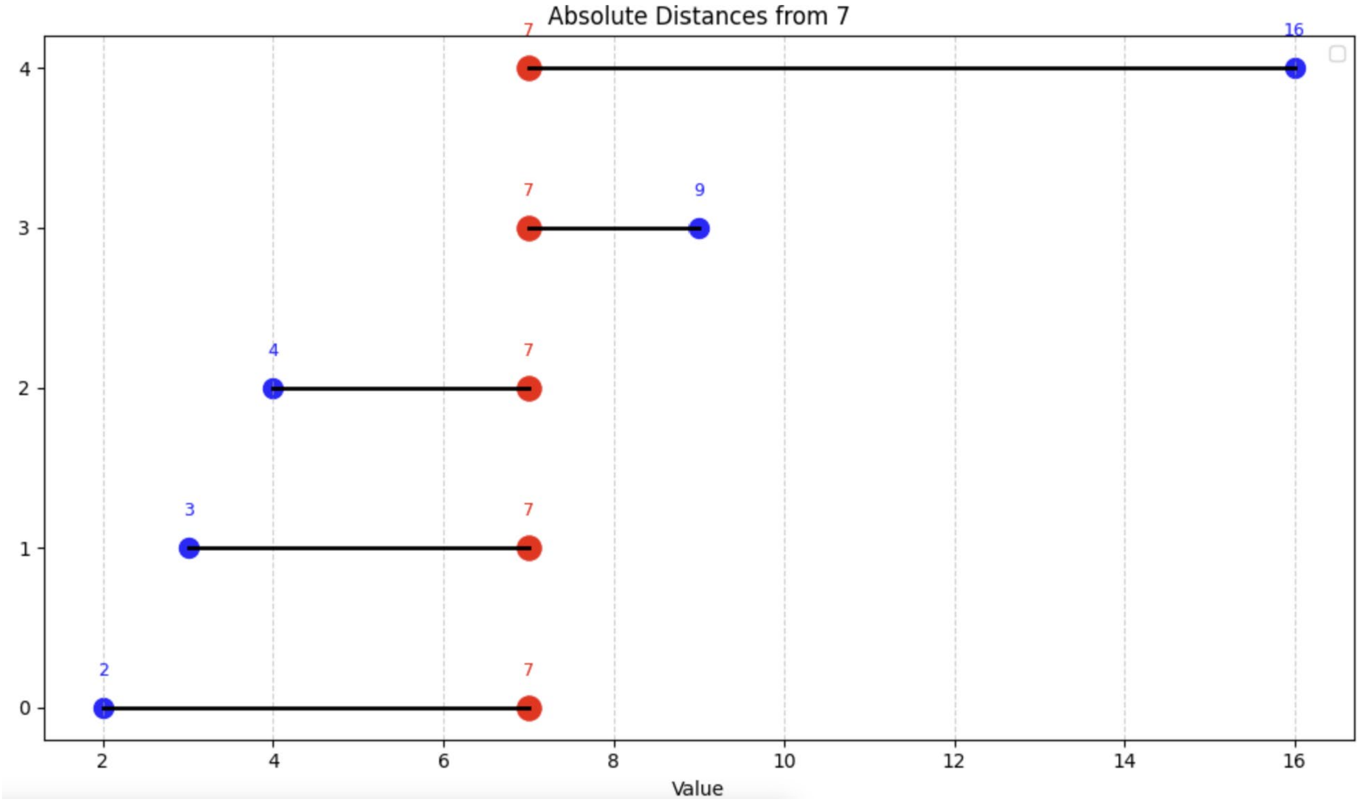
- Idea 2: Find the point of smallest absolute deviation (from it).



Central Tendency

What is a "central" location of a distribution?

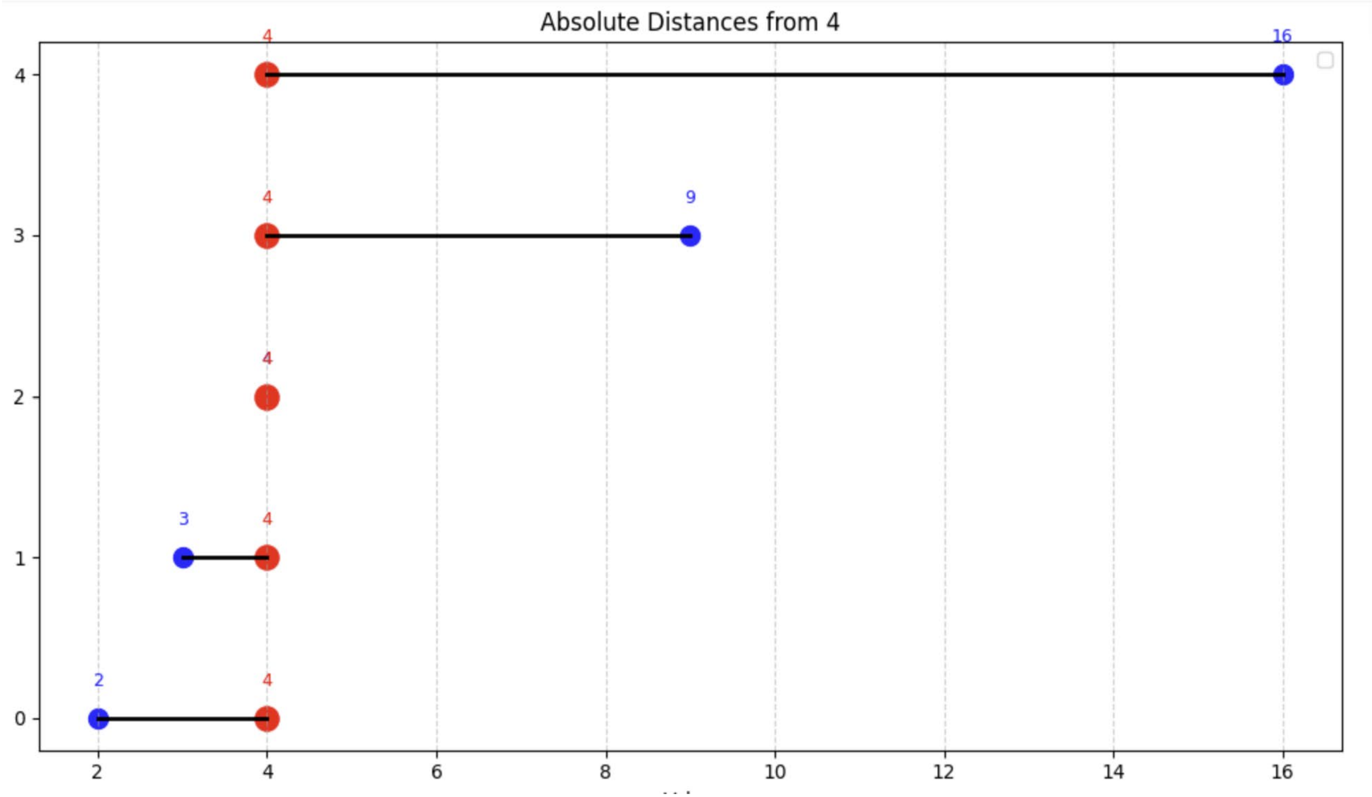
- Idea 2: Find the point of smallest absolute deviation (from it).



Central Tendency

What is a "central" location of a distribution?

- Idea 2: Find the point of smallest absolute deviation (from it).



Answer: Median of the distribution

Central Tendency

What is a "central" location of a distribution?

- Idea 3: Find the point of smallest squared deviation (from it).

We want to minimize: $$\sum_i (X_i - y)^2$$

Equivalently minimize:

$$\mathbb{E}[(X-y)^2] = \frac{1}{N}\sum_i (X_i - y)^2$$

$$= \frac{1}{N}\sum X^2 - 2y\frac{1}{N}\sum X + y^2$$

$$\mu = \frac{1}{N}\sum X$$

$$= \mathbb{E}[X^2] + \mu^2 - 2y\cdot\mu + y^2 - \mu^2$$

is the mean

$$= \mathbb{E}[X^2] + (\mu - y)^2 - \mu^2$$

Minimized at: $y = \mu$

Central Tendency

What is a "central" location of a distribution?

- Idea 3: Find the point of smallest squared deviation (from it).

We want to minimize: $\sum_i (X_i - y)^2$

Equivalently minimize:

$$\mathbb{E}[(X-y)^2] = \frac{1}{N}\sum_i (X_i - y)^2$$

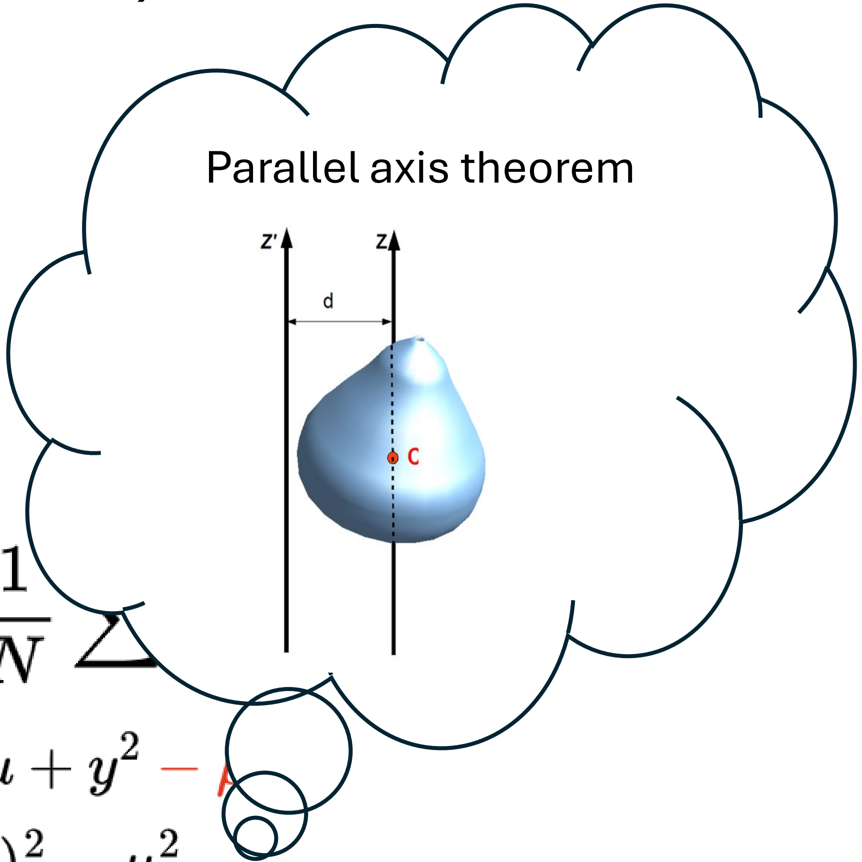
$$= \frac{1}{N}\sum X^2 - 2y\frac{1}{N}\sum$$

$$= \mathbb{E}[X^2] + \mu^2 - 2y\cdot\mu + y^2 - $$

$$= \mathbb{E}[X^2] + (\mu - y)^2 - \mu^2$$

$$\mu = \frac{1}{N}\sum X$$

is the mean



Minimized at: $y = \mu$

- Central Tendency

Summarizes a distribution by its central location

Different ways to define central tendency:

- Mean $\mu = \frac{\sum X}{N}$
- Median = 50th Percentile
- Mode = the value with the highest frequency
- Trimean = $\frac{25^{\text{th}}\text{ Percentile} + 2*\text{Median} + 75^{\text{th}}\text{ Percentile}}{4}$
- Geometric mean $= (\prod X)^{\frac{1}{N}}$, where □ means to multiply

  Eg: $\prod_{i=1}^{5} X_i = X_1 * X_2 * X_3 * X_4 * X_5$
- Trimmed mean = mean for data with some higher and lower values removed

- Central Tendency:

Eg: Given the following data set, compute the mean, the median, the mode, the trimean, the geometric mean and the mean trimmed 18.2%.

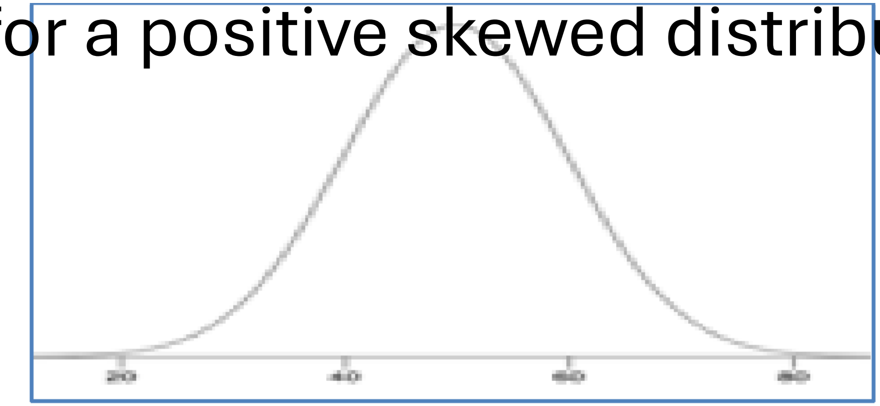
1	3	4	4	4	5	5	7	8	9	31

Measure of central tendency	Value
mean	81/11=7.36
median	5
mode	4
trimean	(4+2*5+7.5)/4=21.5/4=5.375
Geometric mean	5.2
Mean trimmed 18.2%	49/9=5.44

▪ Central Tendency – comparing various measures

For symmetric distributions:
Mean = Median= Trimean = Trimmed mean = Mode (except bimodal distr)

For skewed distributions:
Differences among the measures.
Example – the mean is typically higher than the median for a positive skewed distribution



Symmetric

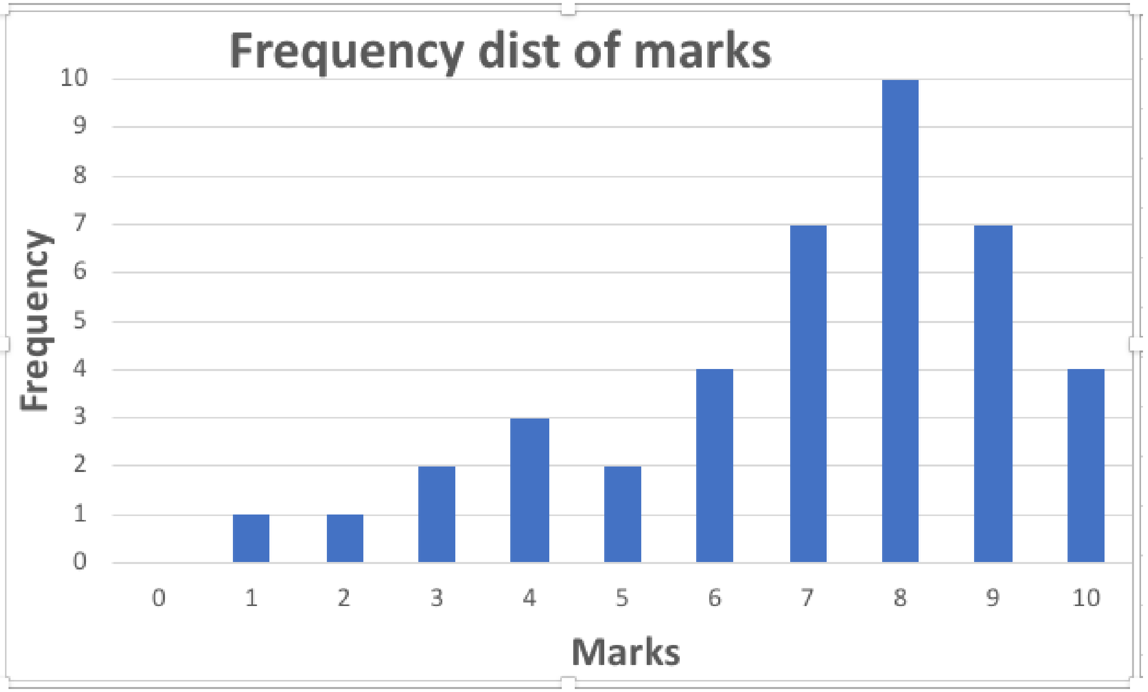
- Central Tendency – comparing various measures

Example:

Negative skewed

Calculate the following:

1. Mode
2. median
3. trimean
4. mean



N=41

- Central Tendency – comparing various measures

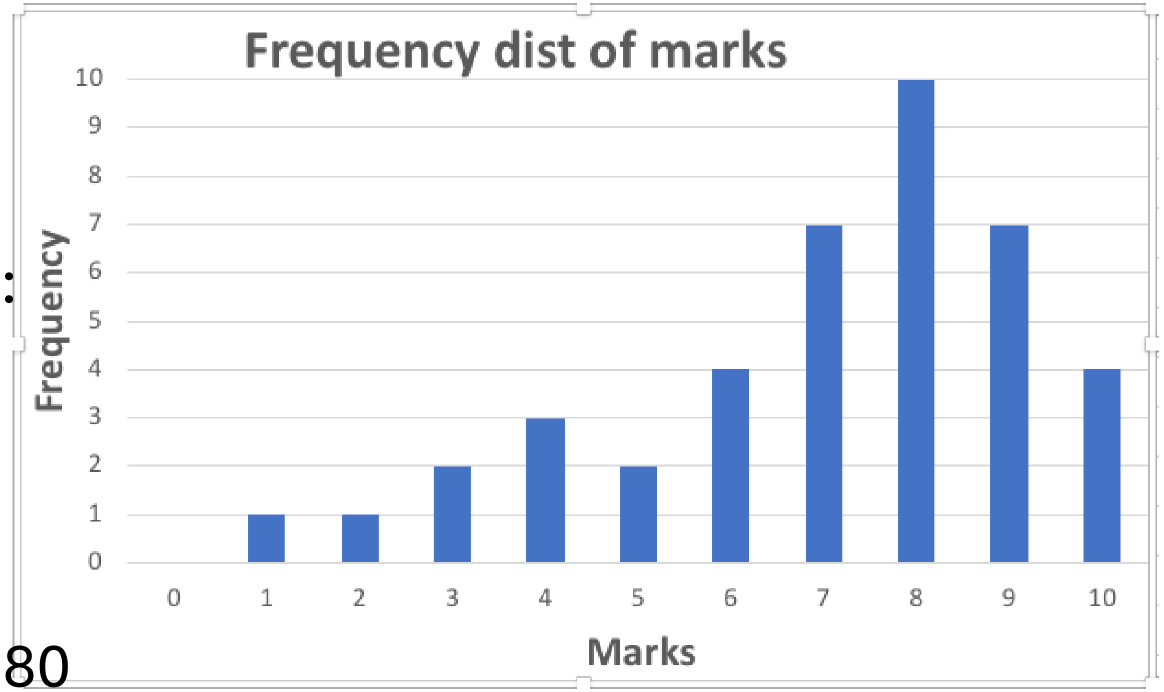
Example:

Negative skewed

Calculate the following:

1. Mode (=8)
2. Median (=8)
3. Trimean (=7.75)
4. Mean

[1+2+6+12+10+24+49+80+63+40]/41= 287/41=7



N=41

Measures of Variability

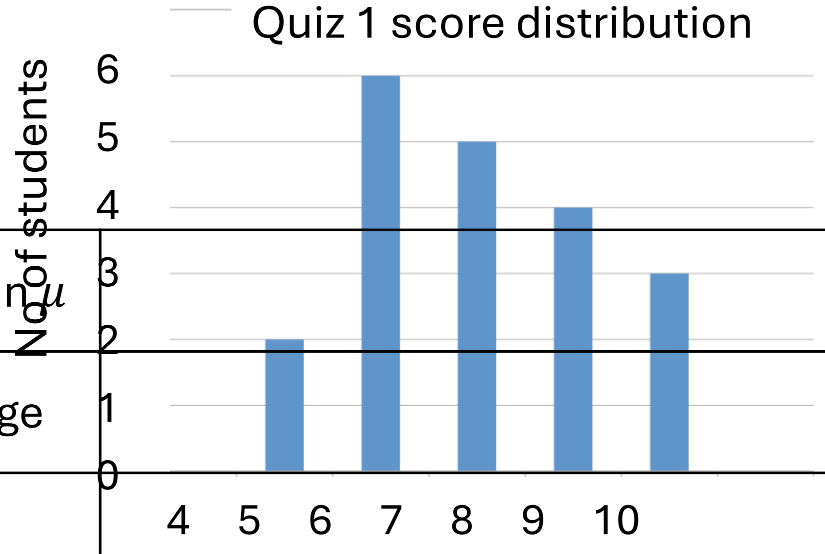
An indication of how spread out is the distribution

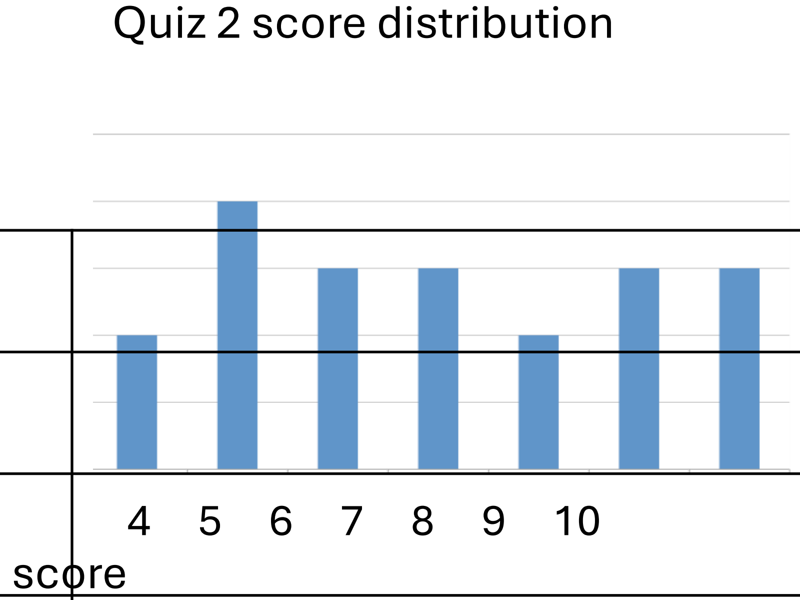
Frequently used measures of variability:

- Range = Highest value – Lowest value
- Interquartile Range IQR = 75^{th} – 25^{th} Percentile
- Variance $\sigma^2 = \dfrac{\sum(X-\mu)^2}{N}$
- Standard deviation = $\sqrt{\text{Variance}}$

- ## Variability:

Eg: Given the following 2 data set, compute the mean, the range, the IQR and the variance.





Mean μ		
Range		
IQR		
Var σ^2		

True or False questions:

Compute mean and variance from the **population** of size *N* :

Population Mean $\mu = E[X] = \frac{\sum X}{N}$

$\sigma^2 = E[X^2] - \mu^2$

Population Variance $\sigma^2 = E[(X-\mu)^2] = \frac{\sum (X-\mu)^2}{N}$ or $\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$

Estimate mean and variance from a **sample** of size *n* :

Sample Mean $\bar{x} = \frac{\sum X}{n}$

Sample Variance $s^2 = \frac{\sum (X-\bar{x})^2}{n-1}$ or $\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$

Why n-1 ?

Suppose the denominator of s^2 is n, instead of $(n - 1)$:

$$s^2 = \frac{\sum(X - \bar{x})^2}{n} = \frac{1}{n}\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)$$

For unbiased estimate, we expect the mean of s^2 to be equal to σ^2:

$$E[s^2] = E\left[\frac{1}{n}\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\right]$$

$$= \frac{1}{n}\left(\sum E[X^2] - \frac{E[(\sum X)^2]}{n}\right)$$

$\sigma^2 + \mu^2$ from the previous slide

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{E[(\sum X)^2]}{n}\right)$$

$$E[s^2] = \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{E[(\sum X)^2]}{n}\right)$$

Let $Y = \sum X$

$E[Y^2] = \sigma_Y^2 + \mu_Y^2$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{Var[\sum X] + (E[\sum X])^2}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{\sum Var[X] + (\sum E[X])^2}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{n\,\sigma^2 + (n\,\mu)^2}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \sigma^2 - n\,\mu^2\right) \quad = \frac{1}{n}(n - 1)\,\sigma^2$$

If the denominator is $(n - 1)$, then

the mean of s^2 is equal to σ^2, i.e. $$E[s^2] = \sigma^2$$