Probability and Statistics

DR. AHMED TAYEL

Department of Engineering Mathematics and Physics, Faculty of Engineering, Alexandria University

ahmed.tayel@alexu.edu.eg

Statistics

- Gathering and analyzing information about a random phenomenon.
- This how you decide that a certain random variable follows a certain distribution → STATISTICS

Definitions

- **Population** → what you are testing "all observations".
- **Sample** → a subset of the population, which should be enough to represent the entire population.

sample size "large enough relative to the population size" + Representative + Indep.

Statistics - Classification

Today's lecture

وصفي

120, 118, 90, - - -

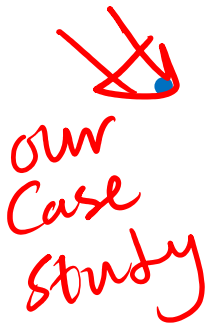
- **Descriptive statistics**
  - Describes the sample through numbers
    "aka numerical summaries" → what you actually measured.

استدلالي / استنتاجي

- **Inferential statistics**
  - Drawing conclusions about the population based on the sample.
  - Statistical Inference.

Types of Data

- ***Qualitative:*** described by **words**
  - **Example:** color, type, blood group, etc.



***Quantitative*:** described by **numbers**

- **Example:** number of customers, length of a metal rod, time until sth occurs, ets.

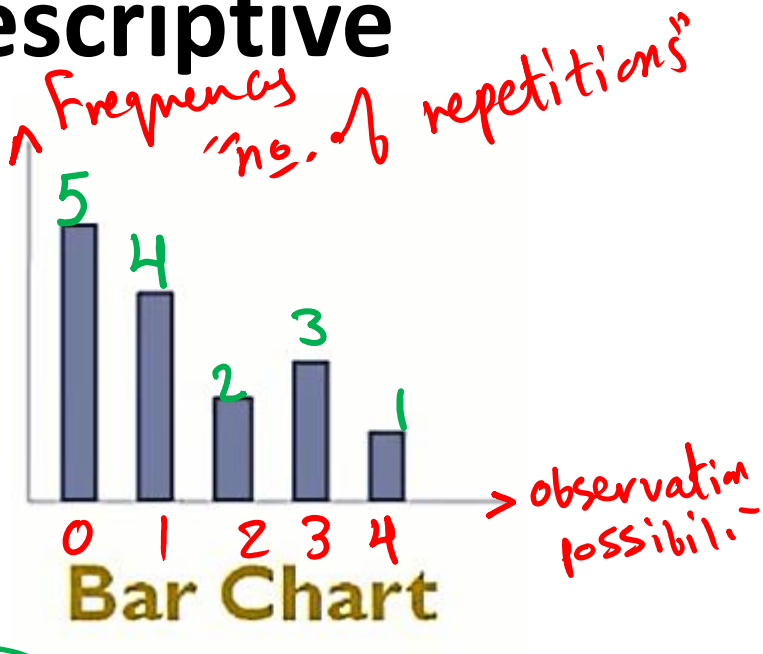
- ***Ordinal:*** *neither qualitative nor quantitative,* this is an "in-between" case.
  - Observations are not numbers but they **can be ordered**
  - **Example:** much improved, improved, same, worse, much worse.

How to describe data graphically (Descriptive statistics) – Quantitative data

Frequency is the same concept as the probability

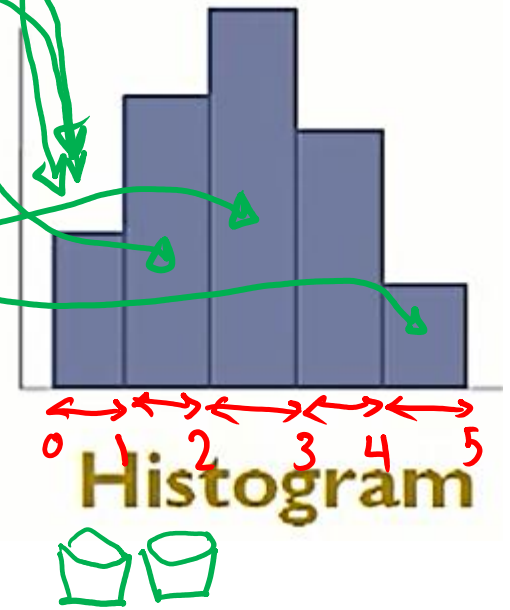
- ***Discrete data:***
  - use **frequency table** and **bar chart.**

0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 3, 3, 3, 4



- ***Continuous data*:**
  - use a **histogram**

0.8, 0.7, 0.94, 2.3, 4.5, 1.5



→ Can also be used for discrete data

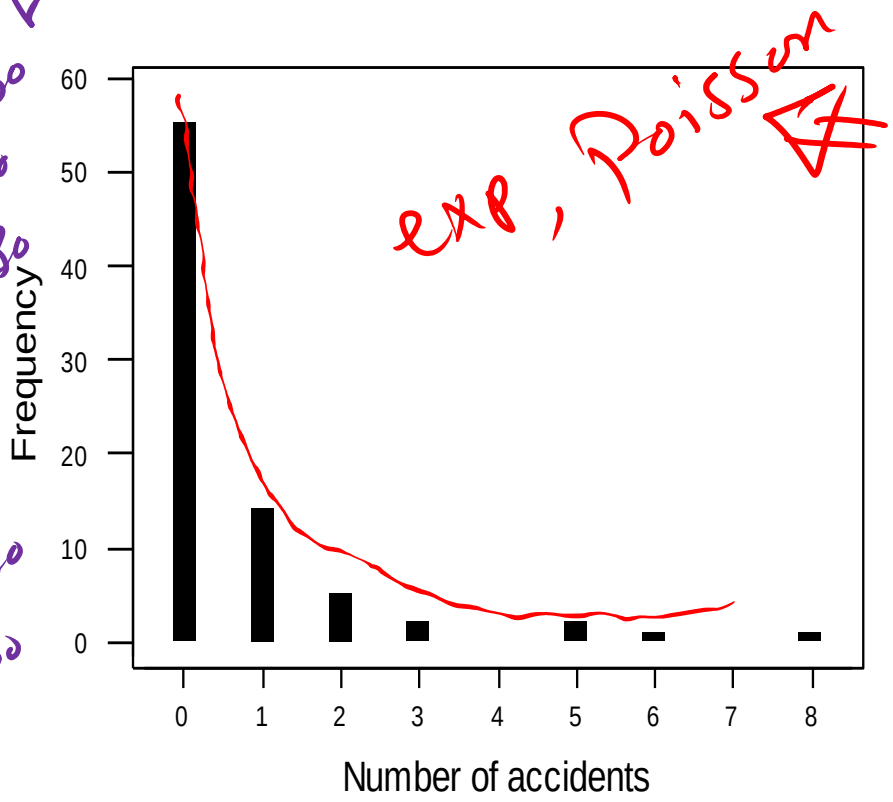
Ex: discrete dat 1 → 100, each possible observation happens with freq 1 or 2

Discrete data – Bar charts

Data about the number of car accidents in a town is collected over 80 days

Number of accidents	Frequency
0	55
1	14
2	5
3	2
4	0
5	2
6	1
7	0
8	1

Frequency table

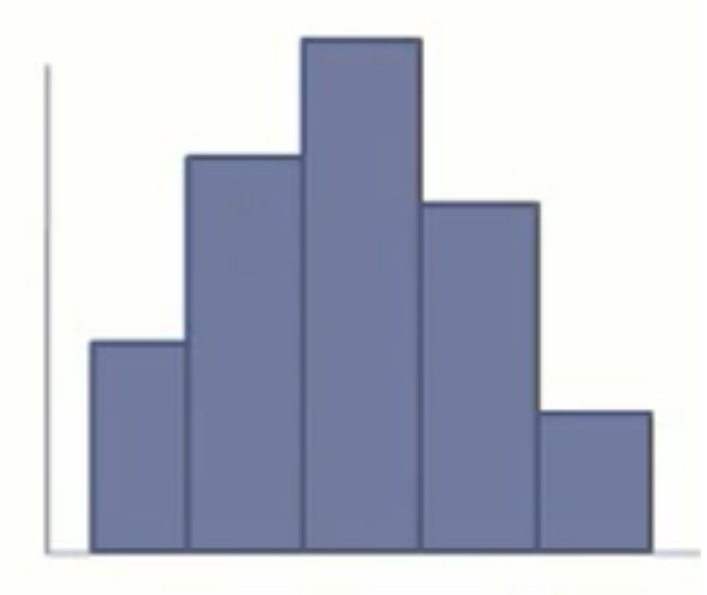


Bar chart

Continuous data – Histogram

Range = Max − Min

- For **continuous data** sets
- The **range** are **divided** into **intervals** (or "**bins**")
- If we have $\boldsymbol{n}$ sample points, then the number of bins is usually taken to be $\sqrt{\boldsymbol{n}}$.
  - Number of bins could be stated in the problem statement



Steps

* Determine the Range = Max − Min
* Extend the Range a bit larger
  - Extended range divisible by # of bins
  - Sample is much smaller the the population
    - $\min_{pop} < \min_{sample}$ also $\max_{pop} > \max_{sample}$
* Calculate the difference in Range

Diff = Extended range − orig. range

Continuous data – Histogram

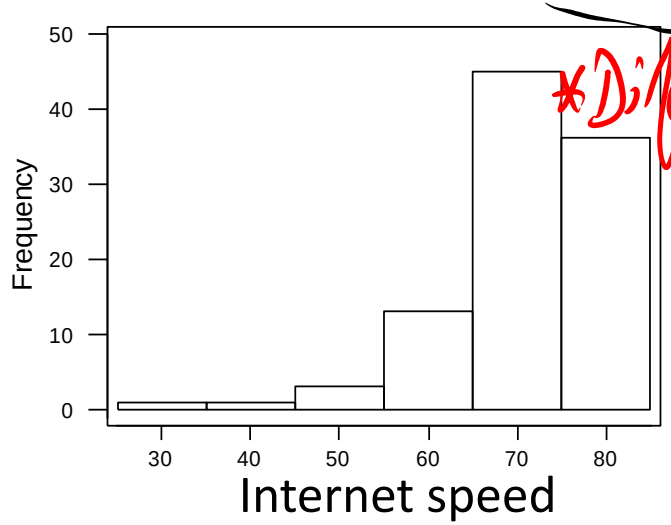
Example

The following data are the internet speed taken at different 99 time instant in MBps.

Construct a frequency table and **histogram of 6 bins**.

62	64	63	70	63	69	65	74	67	77	65	72	65
77	71	79	75	78	64	78	72	32	78	78	80	69
69	65	76	53	74	78	59	79	77	76	72	76	70
76	76	74	67	65	79	63	71	70	84	65	78	66
72	55	74	79	75	64	73	71	80	66	50	48	57
70	68	71	81	74	74	79	79	73	77	80	69	78
73	78	78	66	70	36	79	75	73	72	57	69	82
70	62	64	69	74	78	70	76					

Speed interval	Frequency
24.5 - 34.5	1
34.5 - 44.5	1
44.5 - 54.5	3
54.5 - 64.5	13
64.5 - 74.5	45
74.5 - 84.5	36



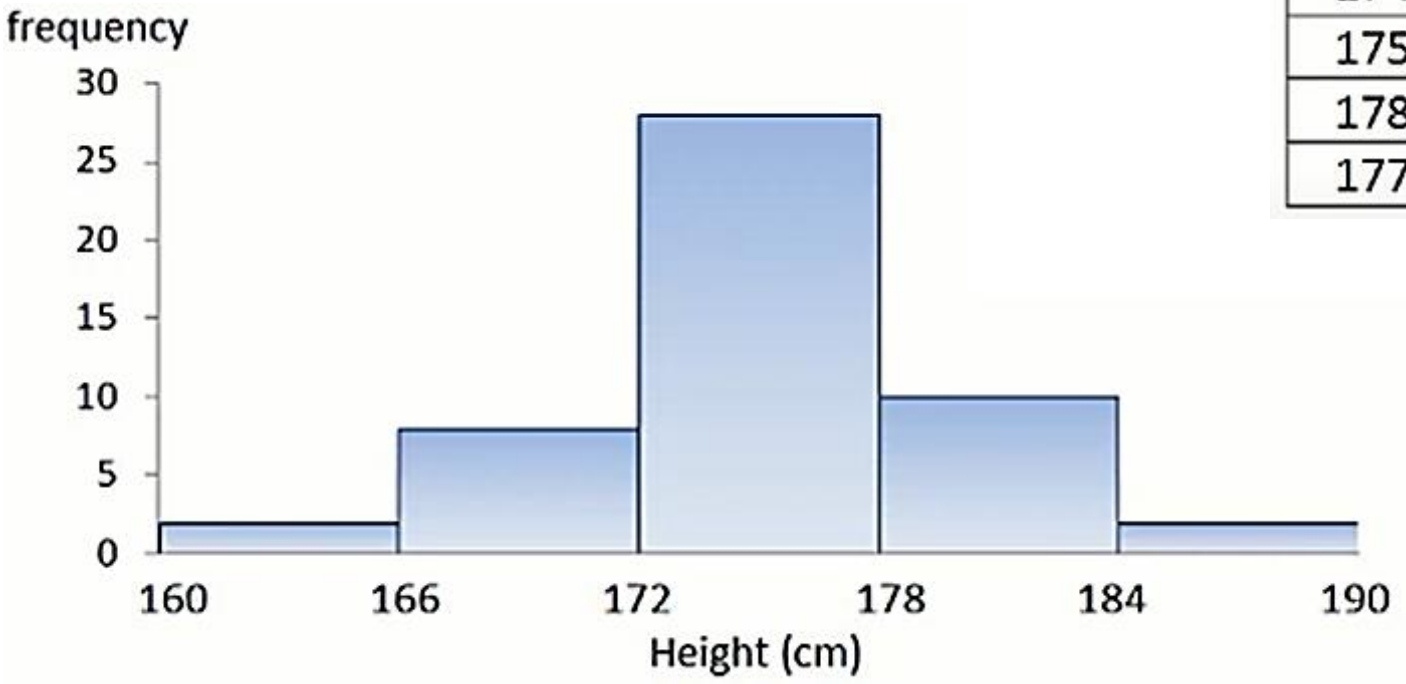
Continuous data – Histogram

Example

Data heights of 50 students.

Plot **histogram of 5 bins**.

174	179	170	173	175
165	181	175	173	181
167	180	176	176	173
171	184	176	178	174
171	186	176	177	178
173	175	174	179	170
174	181	166	180	176
175	173	168	180	168
178	174	169	184	169
177	178	171	185	171



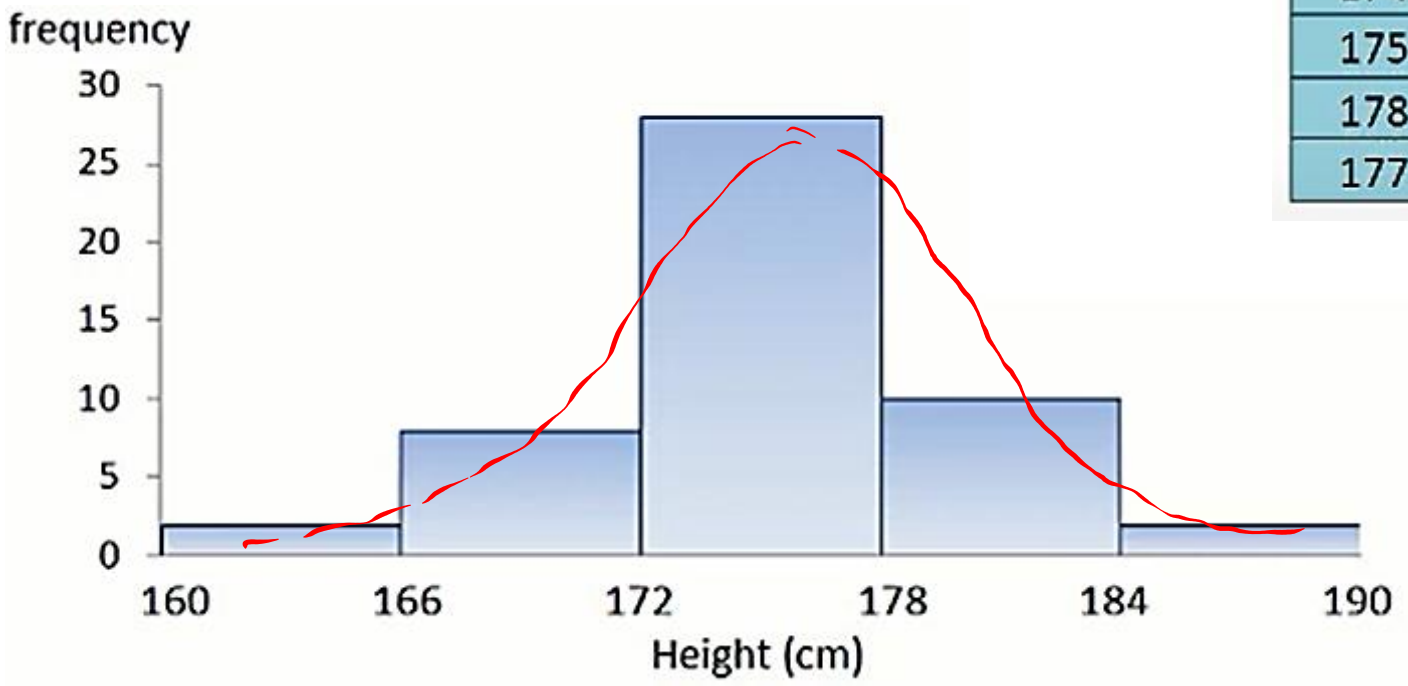
Continuous data – Histogram

Example

Data heights of 50 students.

Plot **histogram of 5 bins**.

174	179	170	173	175
165	181	175	173	181
167	180	176	176	173
171	184	176	178	174
171	186	176	177	178
173	175	174	179	170
174	181	166	180	176
175	173	168	180	168
178	174	169	184	169
177	178	171	185	171



Looks like a Normal distribution

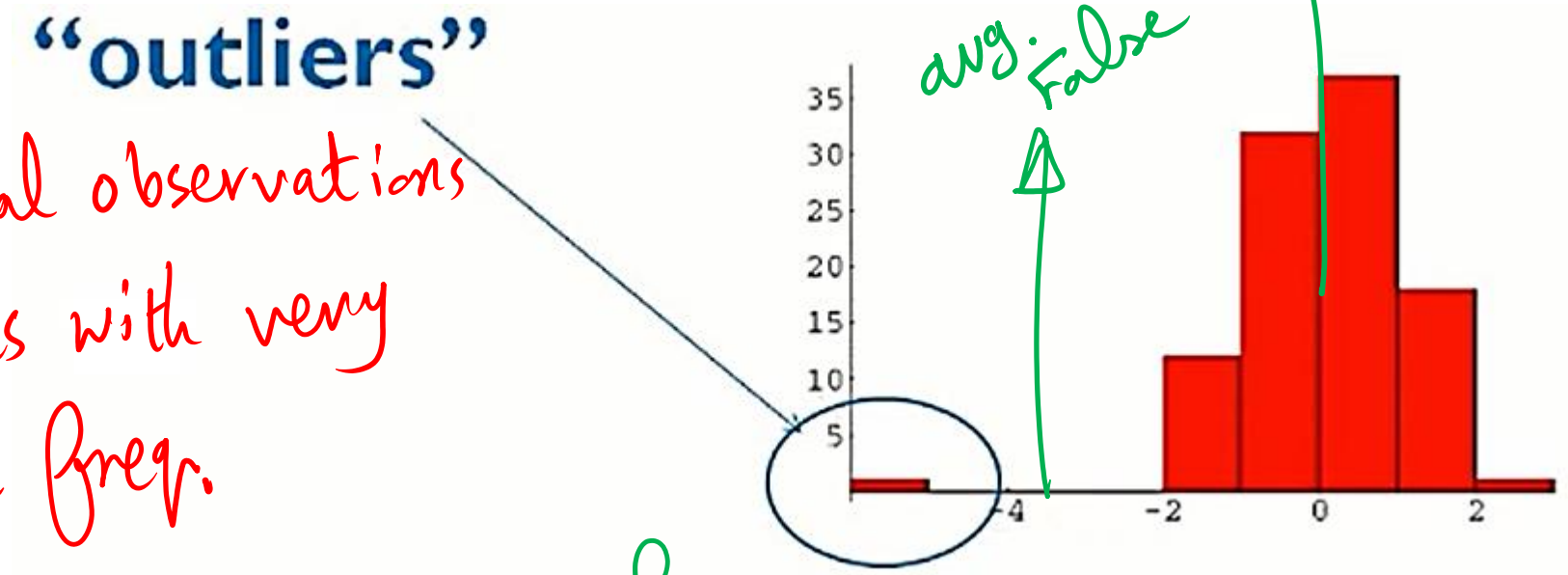
Continuous data – Histogram

- Sometimes, you find **observations** that lie at an **abnormal distance** from the rest of the observations.
- These observations are called

"outliers"

– or extreme abnormal observations

– happens with very small freq.



Numerical summaries

Measures of Central Tendency

- Sample mean
- Median
- Mode

Measures of Variability

- Sample Variance
- Sample Standard Deviation
- Coefficient of Variation

Other measures

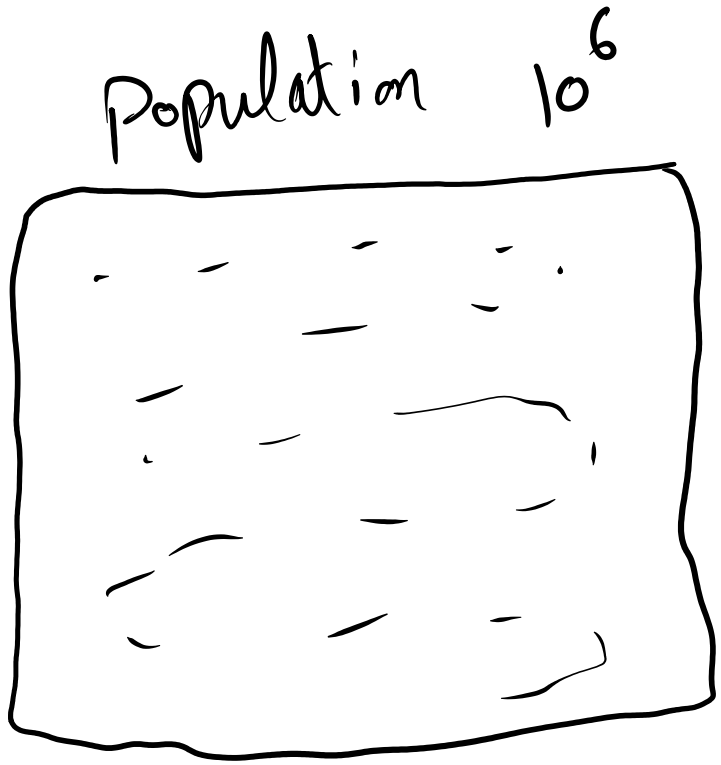
- Range
- Percentile
- Skewness

Difference betn.. mean / Variance & sample mean / Variance

Related to the entire population → Deterministic μ, σ^2

Related to the sample → Random variable $\bar{x}, s^2$

			Sample Size	Comment
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	10^3	very different
$\bar{x}_4$	$\bar{x}_5$	$\bar{x}_6$	10^4	less different
$\bar{x}_7$	$\bar{x}_8$	$\bar{x}_9$	10^5	closer ✓
$\bar{x}$	$\bar{x}$	$\bar{x}$	10^6 entire population	$\bar{x} = \mu$



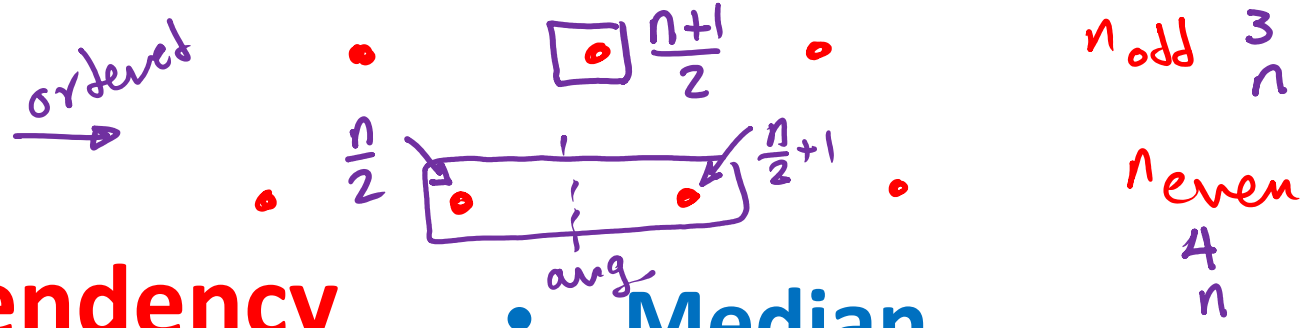
Numerical summaries

Measures of Central Tendency • Sample mean

- If we have n observations $x_1, x_2, \ldots, x_n$
- The average or sample mean is given by

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

Numerical summaries



Measures of Central Tendency

- **Median**

- The median **divides the data** into **two equal halves**.
- To find it, we first **arrange** the given data in **ascending** order.
- Given n observations $\{x_1, x_2, \ldots, x_n\}$, then the sorted version in ascending order will be denoted as $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$.
- Using this notation, the sample median is calculated as

$$\hat{x}_{0.5}(n) = \begin{cases} x_{(n+1)/2}, & \text{if n odd} \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{if n even} \end{cases}$$

Numerical summaries

Measures of Central Tendency

- **Median**

- The **median** for a probability distribution, whose cumulative distribution function is **F(x)** is given by

$$F(x_{0.5}) = 0.5$$

$$x_{0.5} = F^{-1}(0.5)$$

Numerical summaries

Measures of Central Tendency

- **Mode**

- The **mode** is the **most frequently** occurring value in the given data

Example: Data = {1, 2, 3, 2, 5, 4, 1, 2}

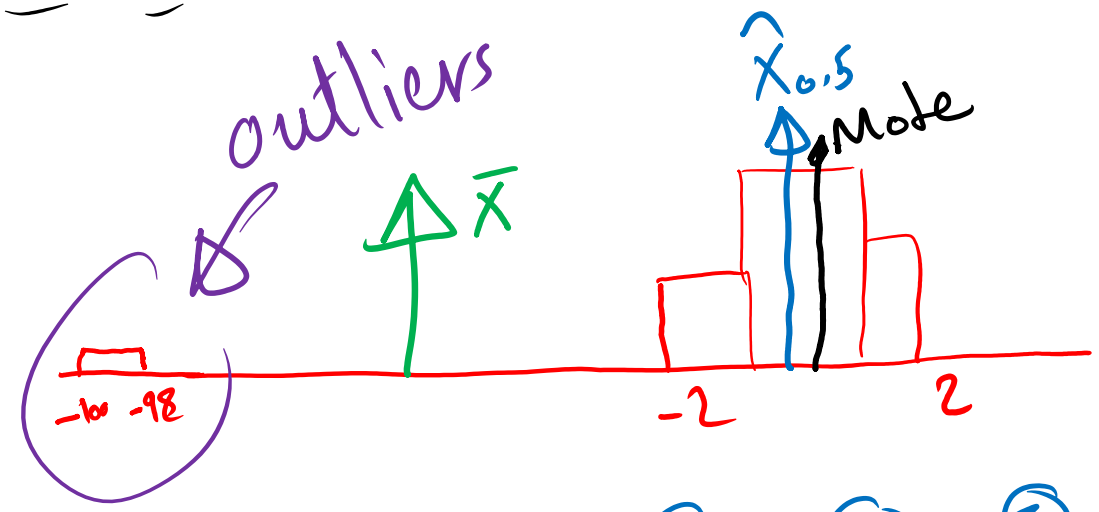
Mode = 2

$-100, -98, -2, -1.8, -1.5, -1, -0.7, -0.1, 0.2, 0.4, 0.8,$

$1.4, 1.7, 2, \circ, - - -$

$\bar{X}$ = Sample mean = $\dfrac{-100 - 98 - 2 - 1.8 - \cdots}{n}$

- very sensitive to outliers
- Outliers are evaluated by their value



(1) (2) (3) (4) (5) (6) (7) (8)

$-100, -98, -2, -1.8, -1.5, -1, \underbrace{-0.7, -0.1}_{\text{avg.}},$

(9) (10) (11) (12) (13) (14) ← even

$0.2, 0.4, 0.8, 1.4, 1.7, 2$

* Median = $\hat{X}_{0.5}$ "data ordered"
  - less sensitive to outliers
  - outliers are evaluated by their order not the value

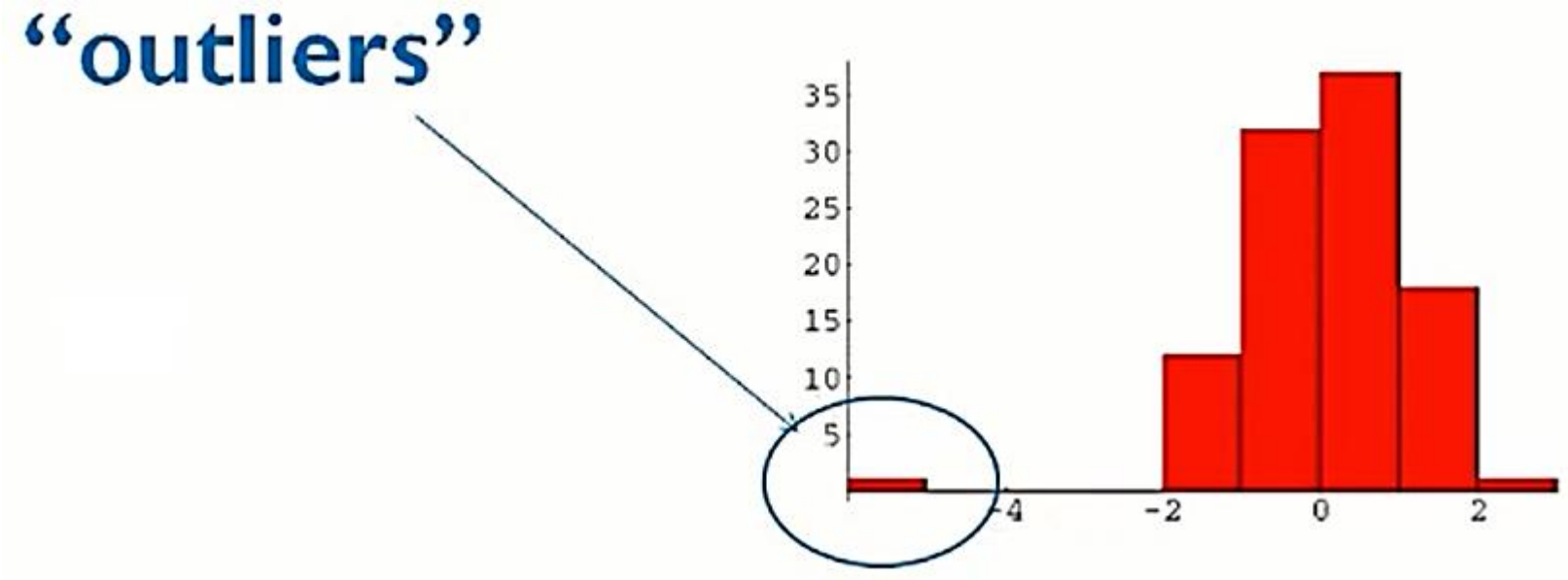
* Mode : most freq. data pt. --> 0 → Not sensitive at all to outliers

Numerical summaries

Measures of Central Tendency

Notes:

- The **median** is **less sensitive** than the **mean** to **extreme observations.**
- The **mode** is **not affected** by **extreme observations.**



Numerical summaries

Measures of Variability / dispersion around the mean

- **Sample Variance**

observation, sample mean

$$s^2(n) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$ why

Computational formula

$$s^2(n) = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\,\bar{x}^2}{n-1}$$

- **Sample Standard Deviation**

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

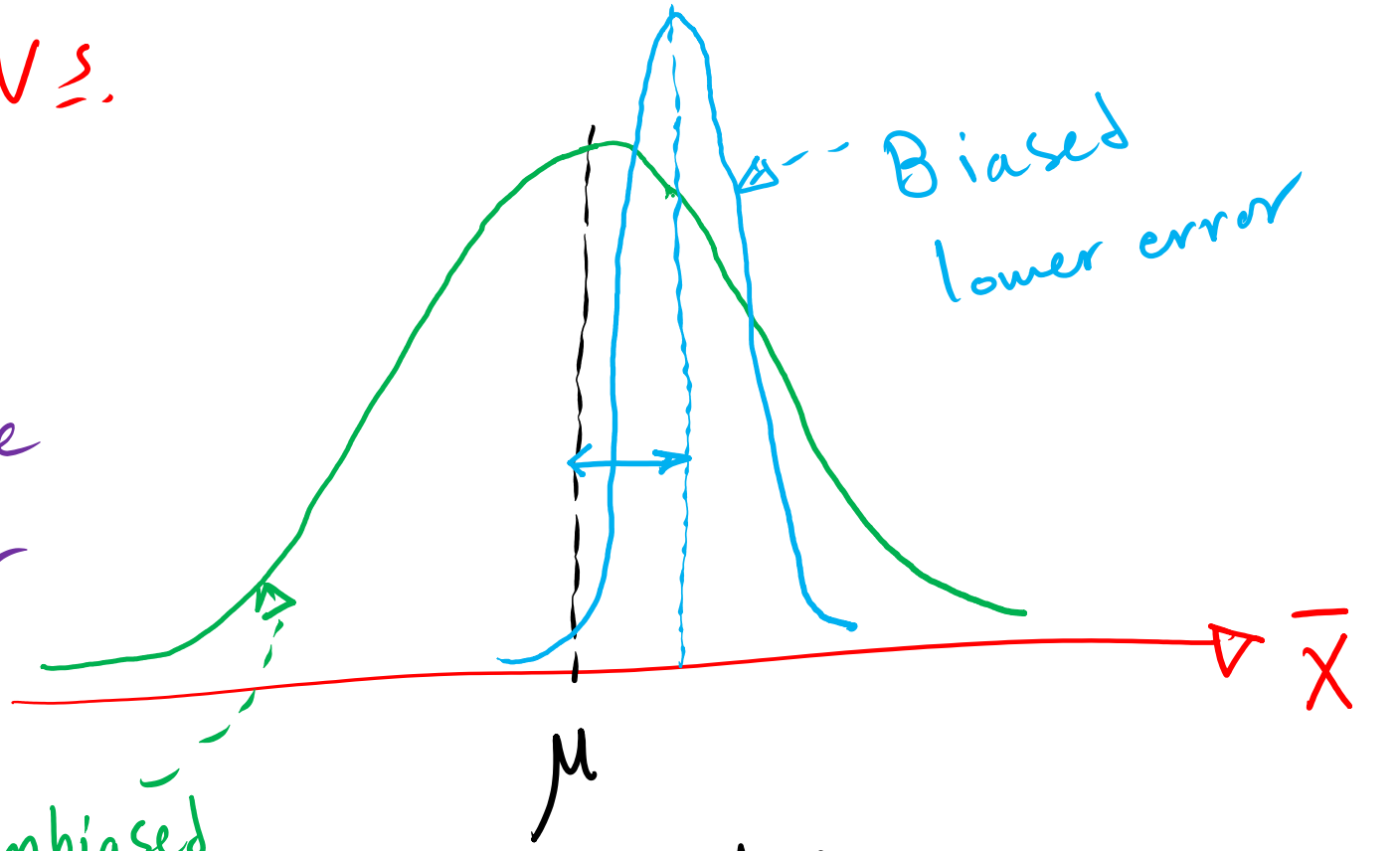
"has the same units as $\bar{x}$ or any other observation"

$\bar{X}$ & s^2 are R.V's.

has a distn.

mean → Bias

Variance → Error



will be used in our calculations

$$s^2 = \frac{\Sigma(\quad)^2}{n-1}$$

unbiased, large error

$$\delta^2 = \frac{\Sigma(\quad)^2}{n}$$

Exist in some reference

Biased, small error

Numerical summaries

data set 1
$s = 100$
$\bar{x} = 10^6$

data set 2
$s = 1$
$\bar{x} = 2$

Measures of Variability

Relative

- **Coefficient of Variation**
  - **Ratio** of the sample **standard deviation** to the sample **mean**.

The closer to zero the better

$$cv = \frac{s(n)}{\bar{x}(n)}$$

data set 1

$\frac{100}{10^6}$

10^{-4}

data set 2

$\frac{1}{2}$

0.5

$s = 0.5\bar{x}$

Numerical summaries

Example

A sample of three random batteries had lifetimes of 2, 6 and 4 hours.

What is the sample variance of the lifetime?

$$\bar{x} = (2+6+4)/3 = 4$$

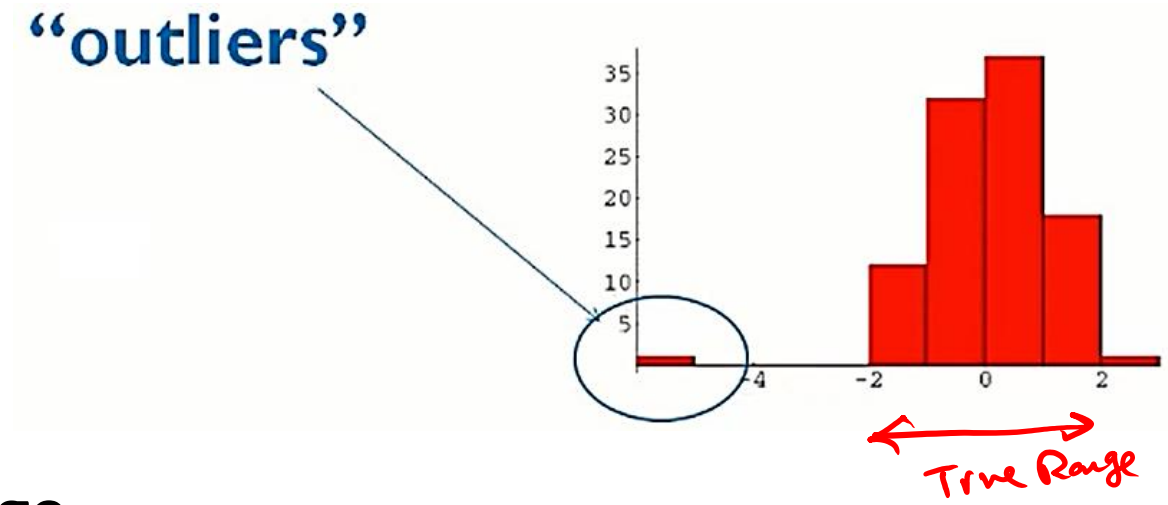
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \Rightarrow s^2 = \frac{(2-4)^2 + (6-4)^2 + (4-4)^2}{3-1} = 4$$

"n-1"

$$s^2 = \frac{\left(\sum_{i=1}^{n}(x_i^2)\right) - n(\bar{x})^2}{n-1} \Rightarrow s^2 = \frac{(2^2+6^2+4^2) - 3\times(4)^2}{3-1} = 4$$

Percentiles and the Interquartile range

The **sample range** is an easy to calculate measure of variability, yet **sensitive to outliers**



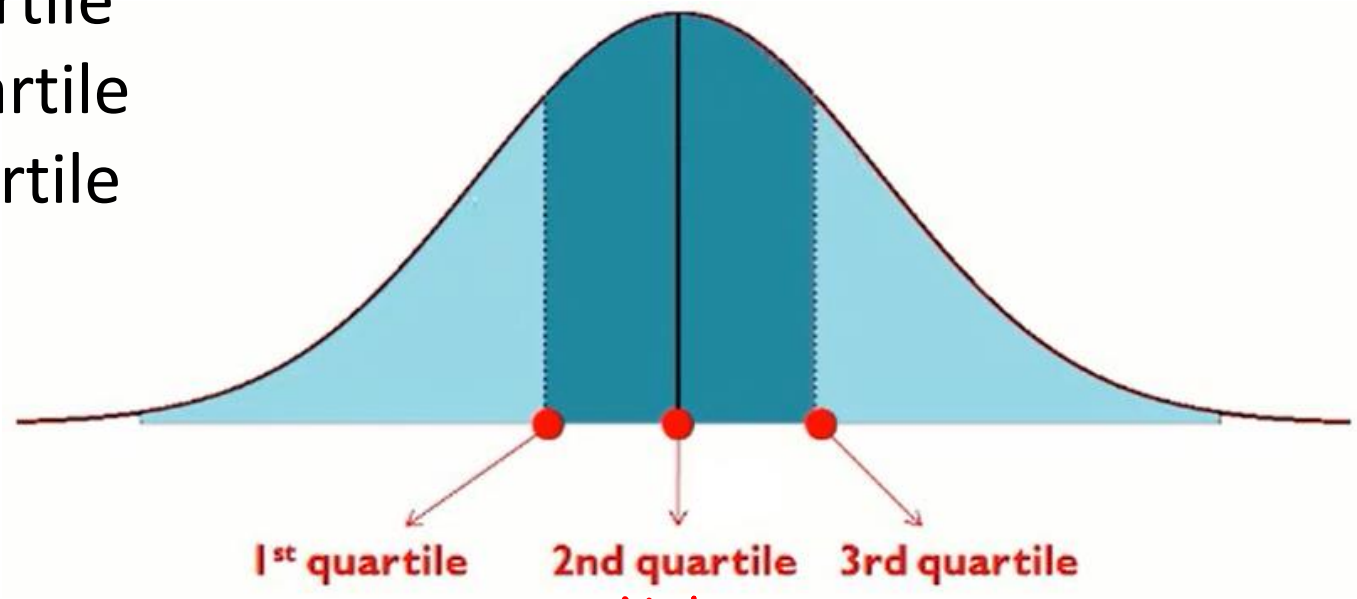
Percentiles and the interquartile range

- The ***k*th percentile** is the value corresponding to cumulative frequency of $k/100$
- Looks like the **CDF** function

Percentiles and the Interquartile range

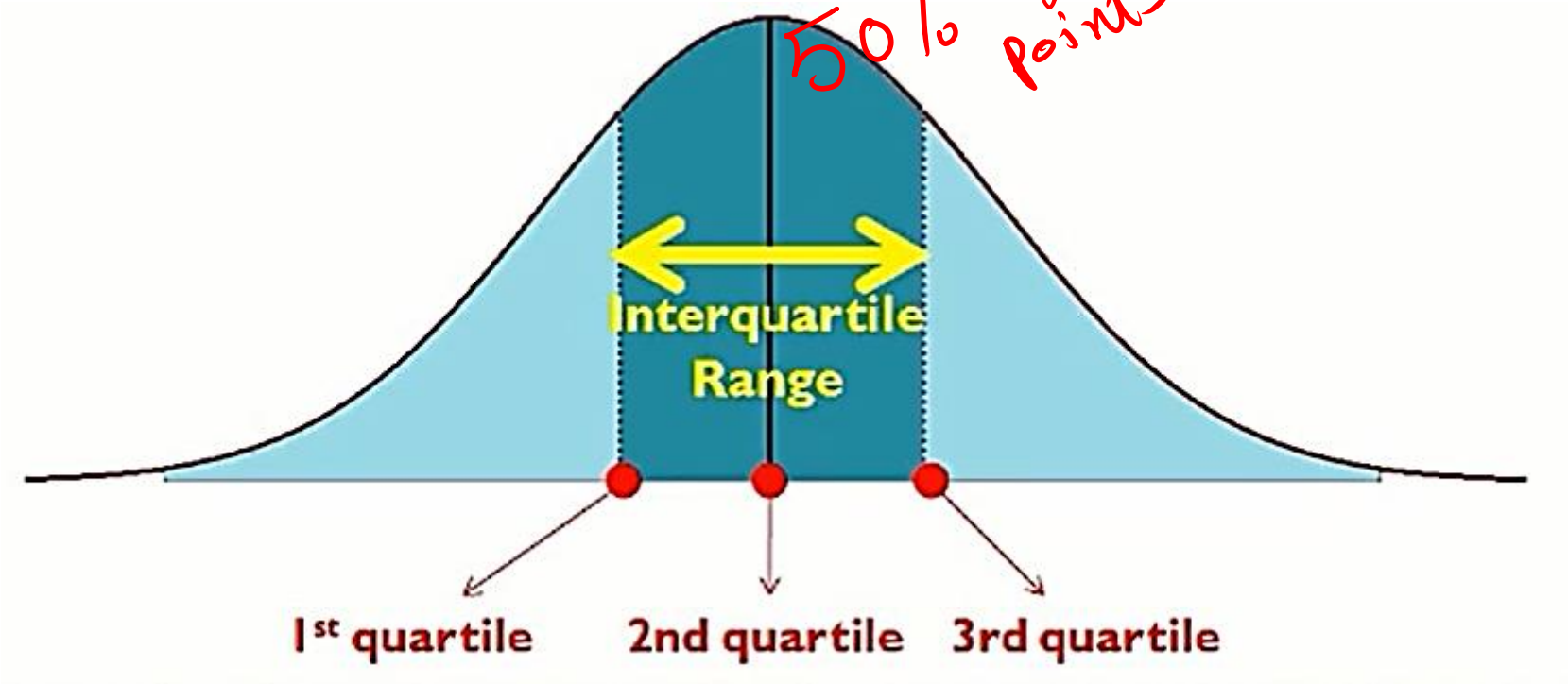
Percentiles and the interquartile range

- The ***k*th percentile** is the value corresponding to cumulative frequency of $k/100$
- Quartiles
  - 25^{th} percentile → 1^{st} quartile
  - 50^{th} percentile → 2^{nd} quartile
  - 75^{th} percentile → 3^{rd} quartile



Percentiles and the Interquartile range

Interquartile range = 3rd quartile – 1st quartile

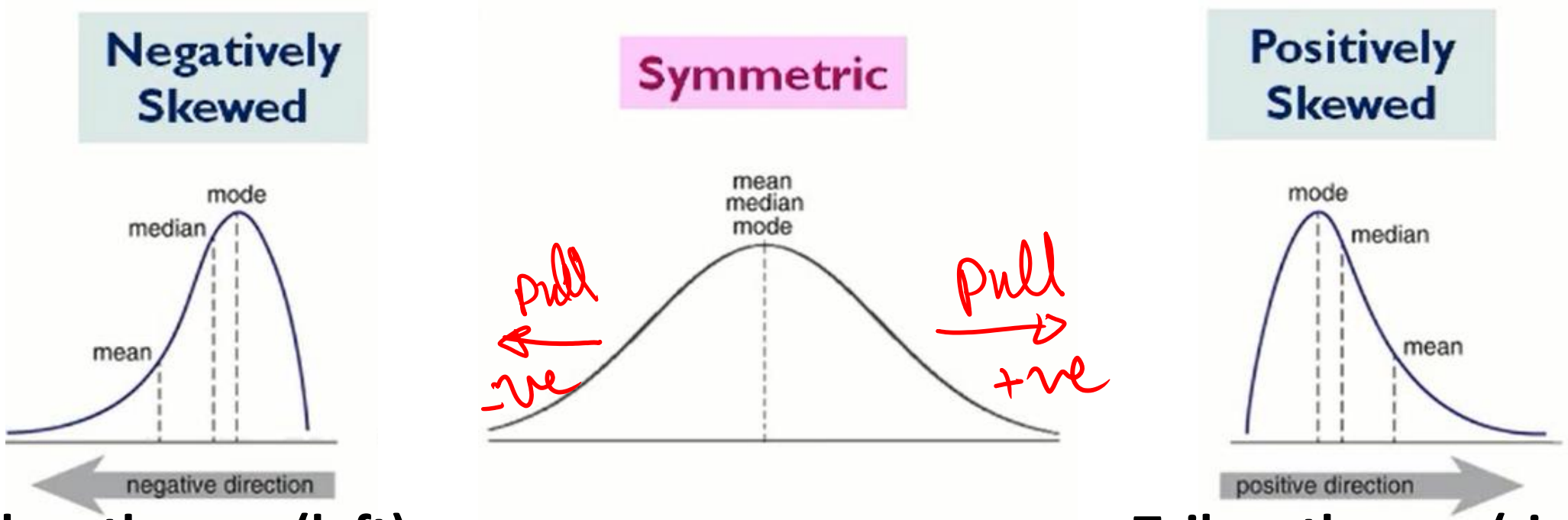


It is the **range** within which the "**middle half**" of the data lie, and so is a **measure of spread** which is **not too sensitive to one or two outliers**.

Numerical summaries

Measures of Symmetry • Skewness

The mean, median, and mode will be approximately equal



Tail on the - ve (left) side of the peak

Tail on the +ve (right) side of the peak

Example

A semiconductor manufacturer produces devices used as central processing units in personal computers. The speed of devices (in megahertz) is important because it determines the price that can be charged for the devices. The following table contains measurements on 12 devices. Compute the sample mean, sample variance, the sample median, and mode. Construct a frequency distribution table and a histogram for the given data.

649	681
652	681
662	681
669	683
677	700
680	717

Notic: Data is ordered as in exam

Example

orig. Range = 717 − 649 = 68 } Diff = 4
Ext. Range = 72
divisible by 4

649	681
652	681
662	681
669	683
677	700
680	717

- $n = 12$
- Number of bins $= \sqrt{12} = 3.46$
- So let's take **4 bins** (Ceil)
- Range:
  - Min = 649 , Max = 717
  - Range= 647 → 719 = 719-647= 72 "dividable by the 4 bins" (−2, +2)
  - Bin width = Range/number of bins = 72/4 = 18

Bin	647 ≤ x < 665	665 ≤ x < 683	683 ≤ x < 701	701 ≤ x < 719
Frequency	3	6	2	1
Relative Frequency	0.25	0.5	0.1667	0.0833

+18 +18 +18 +18

÷12 ÷12 ÷12 ÷12

Total 12

Example

Bin	$647 \le x < 665$	$665 \le x < 683$	$683 \le x < 701$	$701 \le x < 719$
Frequency	3	6	2	1
Relative Frequency	0.25	0.5	0.1667	0.0833

Skewness ?! +ve ly skewed

mean, mode, median will be different

Example

$CV = \frac{s}{\bar{x}} = \frac{18.8647}{677.667} = 0.0278$ "small dispersion"

649	681
652	681
662	681
669	683
677	700
680	717

- Sample Mean

$$\bar{x}(n) = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{649 + 652 + \cdots + 717}{12} = 677.667$$

Using calculator:

Mode → 3:STAT → 1:1-VAR → Enter the data "separated by =" → AC → SHIFT+1 → 4:Var → 2: $\bar{x}$ or 4:sx

- Sample Variance

$$s^2(n) = \frac{(\sum_{i=1}^{n} x_i^2) - n\ \bar{x}^2}{n-1} = \frac{(649^2 + 652^2 + \cdots + 717^2) - 12\ (677.667)^2}{12-1} = 355.878$$

- Sample Standard Deviation $s = \sqrt{s^2} = \sqrt{355.878} = 18.8647$

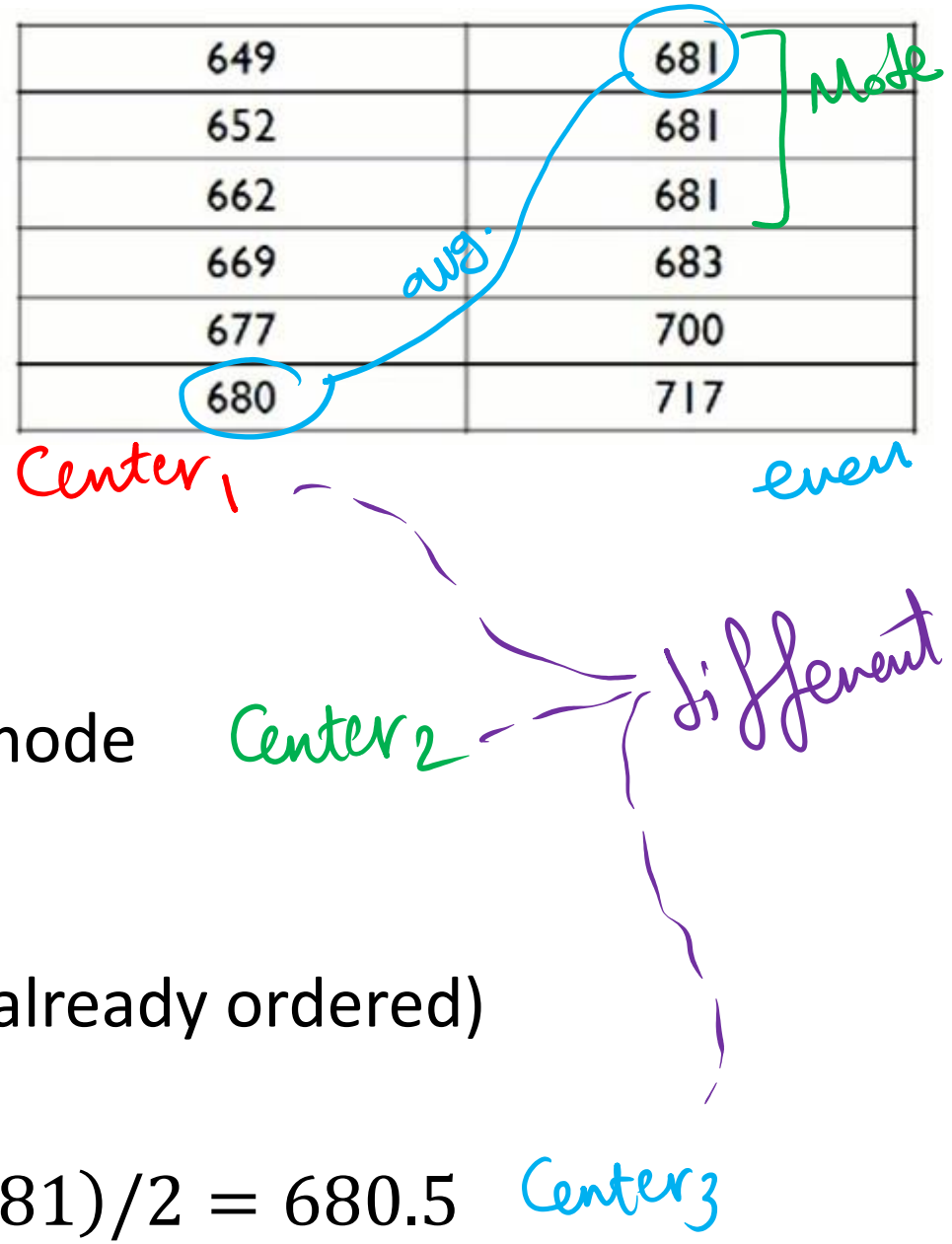
Example

- Sample Mean

$$\bar{x}(n) = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{649 + 652 + \cdots + 717}{12} = 677.667$$

649	681
652	681
662	681
669	683
677	700
680	717



- Mode

  The most repeated number is 681 → mode

- Median

  We have an even number of samples (already ordered)

  Median = $\left(x_{(12/2)} + x_{(12/2+1)}\right)/2$

  $= \left(x_{(6)} + x_{(7)}\right)/2 = (680 + 681)/2 = 680.5$