Statistics

Hello everyone!

In our lesson today, we will discuss descriptive statistics, which requires real-life data. To help illustrate these concepts, please provide me with some information about yourself such as a nickname, your current age, and your weight. This information will only be used within the context of this class to demonstrate measures of central tendency and variation.

Here's an example:

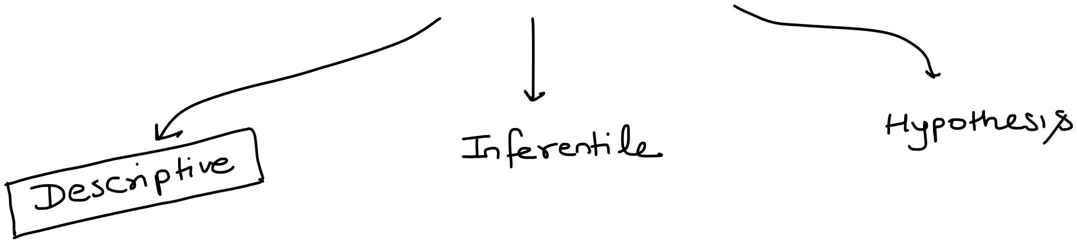
- Nickname: Mitra
- Age: 28
- Weight: 79

Please feel free to share similar details about yourself so that we can make the most out of this learning experience!

Link: https://forms.gle/LtCs7H3nevw9nuAZ8

Available in the chat section pinned comment

Types of Statistics



Descriptive

Inferentile

Hypothesis

Descriptive Statistics

name	gender	age	english.grade	math.grade
Kiana Lor	F	22	3.5	3.7
Joshua Lonaker	M	22	2.9	3.2
Dakota Blanco	F	22	3.9	3.8
Natasha Yarusso	F	20	3.3	2.8
Brooke Cazares	F	21	3.7	2.6
Rochelle Johnson	F	21	3.4	3.1
Joey Abreu	M	22	3.7	3.9
Preston Suarez	M	22	3.8	3.7
Lee Dong	F	24	3.9	3.6
Maa'iz al-Dia	M	22	2.4	2.8
Maja Nicholson	F	23	3.4	3.5

→ info

metrics

measures of central tendency

measures of variation

Using measures of central tendency & variation we can summarise the entire data into few metrics/ numbers which can help an individual to get a sense about the data without going through it.

=

overall

Data consolidation

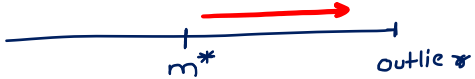
Measures Of Central Tendency

outlier

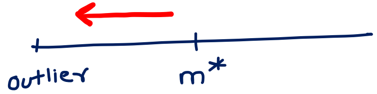
$x = [5 \quad 100 \quad 112 \quad 115 \quad 111]$

$$\text{avg}(x) = \frac{5+100+112+115+111}{5}$$

$$= 88.6$$



m* is actual mean without outlier



1. mean/average
2. median
3. mode

$x =$

in the calculation of avg we are using actual data

$$\text{mean} = \frac{\sum_{1} x_i}{n}$$

$x = [1, 2, 3, 4, 5]$

$$= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

outliers

data points which are very different from majority data

$X = [1, 2, 3, 4, 5, 100]$

$$\text{mean/avg} = \frac{1+2+3+4+5+100}{6} = \frac{115}{6} = 19.16$$

- mean/avg dependent on the data
- mean/avg is influenced by the presence of outliers in the data
- avg/mean shifts towards the outlier

2.5

$x = [1, 2, 3, 4]$

$\frac{2+3}{2} = 2.5$

$x = [1, \boxed{2}, 3]$

Median

[to calculate median, we don't use the actual data]

$x = [\cancel{7}, \cancel{9}, \cancel{11}, \cancel{2}, \cancel{18}, \cancel{5}, \cancel{3}, \cancel{3}, \cancel{9}, \cancel{11}, \cancel{18}, \cancel{9}, \cancel{7}, \cancel{100}]$

① Arrange the data in ascending order

$$\begin{array}{cccccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ [2, & 3, & 3, & 5, & 7, & 7, & \underline{9}, & \underline{9}, & 9, & 11, & 11, & 18, & 18, & 100] \end{array}$$

$\boxed{n = 14}$

② If the no.of data points in the data are odd → median value is the value at $(n/2)^{th}$ position

If the no.of data points in the data are even → median value is the avg of the value at $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ position

$\left(\frac{n}{2}\right)^{th} \rightarrow 9$

$\frac{14}{2}$

$\boxed{7}$

$\left(\frac{n+1}{2}\right)^{th} = \frac{15}{2} = 7.5 \simeq 8$

$\downarrow$

9

$\text{median} = \frac{9+9}{2} = \boxed{9}$

$x = [5, 3, \overset{\downarrow}{7}, 9, 11]$ $n = 5$

ascending order

$= 3, 5, \boxed{7}, 9, 11$

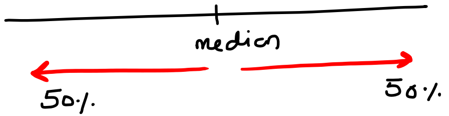
$\left(\frac{n+1}{2}\right)^{th}$ position

$\frac{5+1}{2} = 3^{th}$

median $= 7$

median — ① It's not dependent on the data

② It represents the actual middle value



③ Not impacted by outliers

Mode

observation in the data with maximum occurance

$x = 1, 2, 5, 5, 3, 5, 7, 8, 5, 3, 9$

$1 \longrightarrow 1$

$2 \longrightarrow 1$

$3 \longrightarrow 2$

$\boxed{5 \longrightarrow 4}$

$7 \longrightarrow 1$

$8 - 1$

$9 - 1$

5 = mode of the data

① If there is no clear majority, we don't have any mode

$1, 1, 2, 1, 2, 2$

$\left.\begin{matrix} 1 \longrightarrow 3 \\ 2 - 3 \end{matrix}\right]$ No mode

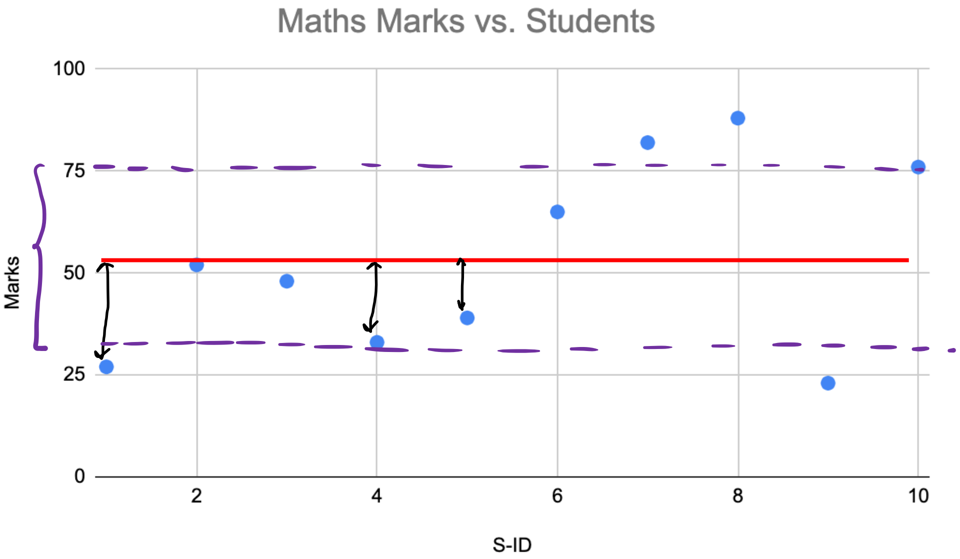
Measures Of Variance

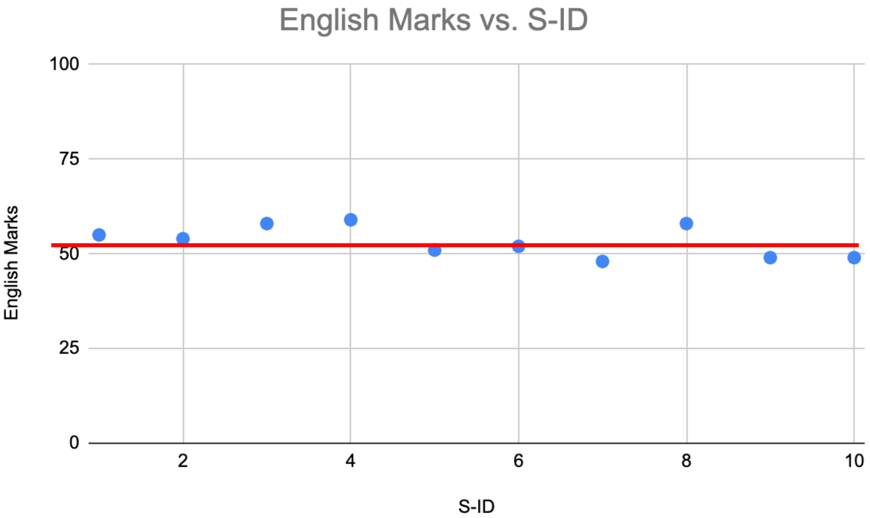
S-ID	Maths Marks	English Marks
1	27	55
2	52	54
3	48	58
4	33	59
5	39	51
6	65	52
7	82	48
8	88	58
9	23	49
10	76	49
Average	53.3	53.3

Quantify the spreadness in the data

How scattered the data is around avg-value

Measures Of Variance





on an avg for maths subject
a students has scored
22 marks above mean or
22 marks below mean

$||d||$ ←

Measures Of Variance

S-ID	Maths Marks	distance	d	d^2
1	27	27 - 53	-26	676
2	52	52 - 53	-1	1
3	48	48 - 53	-5	25
4	33	33 - 53	-20	400
5	39	39 - 53	-14	196
6	65	65 - 53	12	144
7	82	82 - 53	29	841
8	88	88 - 53	35	1225
9	23	23 - 53	-30	900
10	76	76 - 53	23	529
Average	53.3			

avg squared distance of a point from the mean value

$$\frac{\sum d^2}{n}$$

493.7

$\boxed{\text{variance} = \sigma^2}$

= 493.7

If we directly take mean or sum of the distances, the positive distance will cancel negative distance and we might get a value close to 0

avg dist. of a point from the mean value

$\sigma^2 = 493.7$

$\sigma = \sqrt{493.7}$

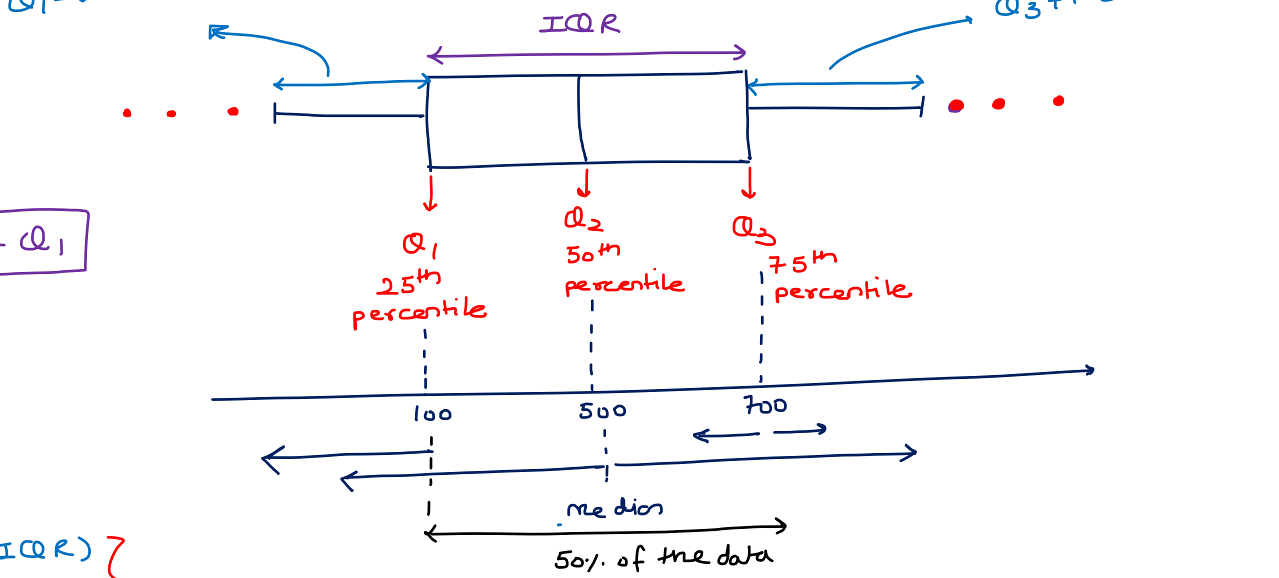
$\boxed{\sigma = 22.2}$

Standard deviation

Measures Of Variance

S-ID	English Marks	Error	Error	e^2
1	55	55-53	2	4
2	54	54-53	1	1
3	58	58-53	5	25
4	59	59-53	6	36
5	51	51-53	-2	4
6	52	52-53	-1	1
7	48	48-53	-5	25
8	58	58-53	5	25
9	49	49-53	-4	16
10	49	49-53	-4	16
Average	53.3		Average Of e^2	15.3
			Std-Dev	3.9

Interquartile Range and Percentile



$\boxed{IQR = Q_3 - Q_1}$

mean = median
= Balanced data

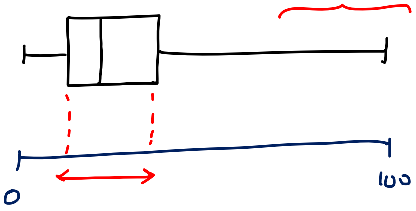
$$\left.\begin{array}{l} x > Q_3 + 1.5(IQR) \\ x < Q_1 - 1.5(IQR) \end{array}\right\} \rightarrow \text{outliers}$$

25th percentile value = 100 (Below 100 we have 25% of the observations)

50th percentile value = 500 (Below 500 we have 50% of the data as same above 500)

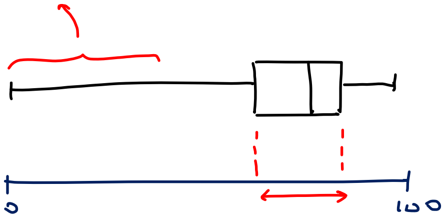
75th percentile value = 700 (Below 700 we have 75% of the observations and 25% above it)

Interquartile Range and Percentile



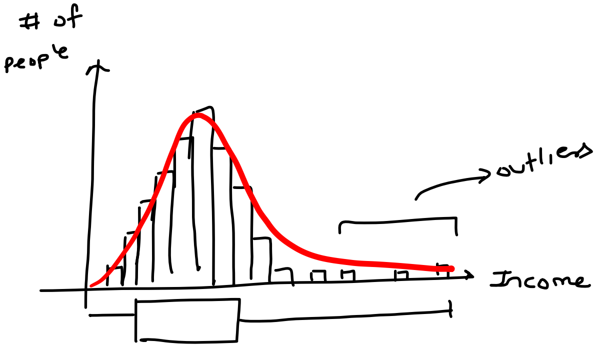
few observations on the higher side (potential outliers)

(50%) most of the data is concentrated on the lower range



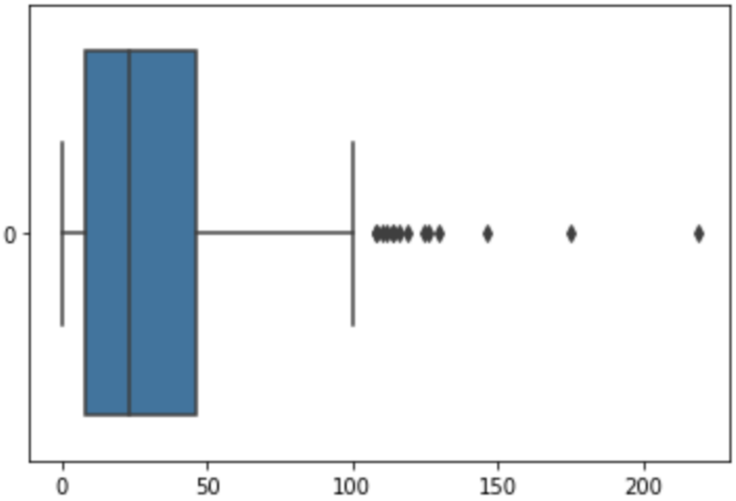
outliers

most of the data concentrated on the higher range

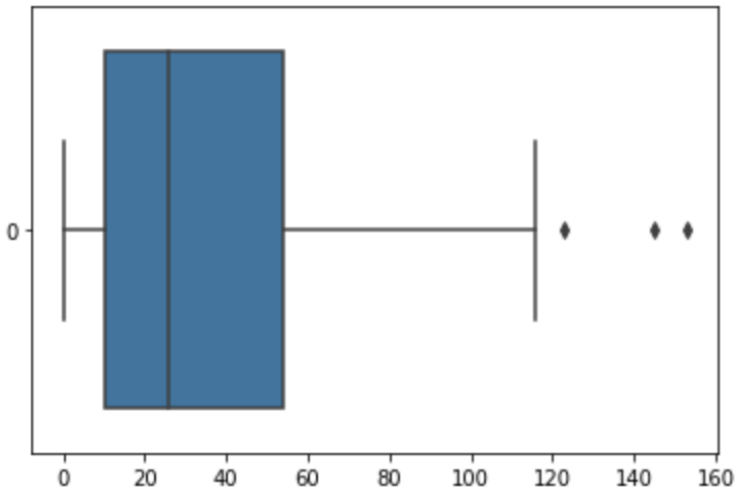


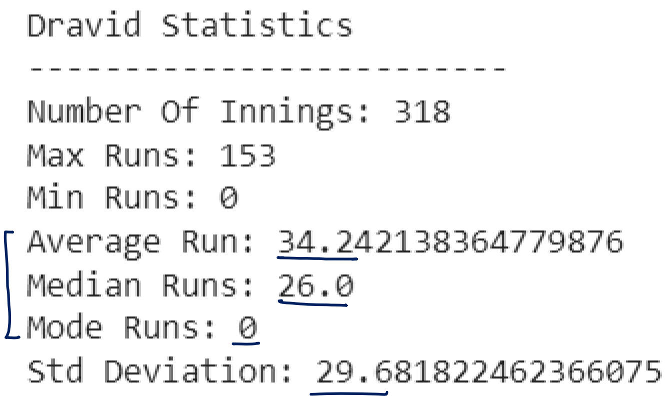
Sehwag VS Dravid

Sehwag



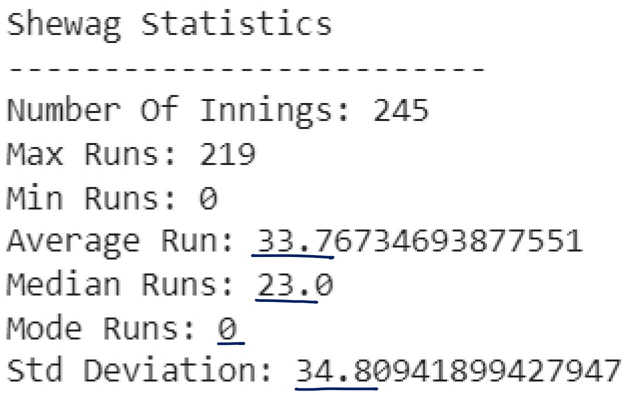
Dravid





```
Dravid Statistics
---------------------------
Number Of Innings: 318
Max Runs: 153
Min Runs: 0
Average Run: 34.242138364779876
Median Runs: 26.0
Mode Runs: 0
Std Deviation: 29.681822462366075
```

less spreadness



```
Shewag Statistics
---------------------------
Number Of Innings: 245
Max Runs: 219
Min Runs: 0
Average Run: 33.76734693877551
Median Runs: 23.0
Mode Runs: 0
Std Deviation: 34.80941899427947
```

more spreadness of the data around avg

Random Variable

Quiz-1: There are 4 people whose average age is 24. We know the age of three people: 20, 22, and 28. What is the median age of these 4 people?

$x_1 \quad x_2 \quad x_3 \quad x_4$

$$\frac{x_1 + x_2 + x_3 + x_4}{4} = 24$$

$$\frac{20 + 22 + 28 + x_4}{4} = 24$$

$$x_4 = (24 \times 4) - (20 + 22 + 28)$$

$$\boxed{x_4 = 26}$$

20, 22, 28, 26

$20, \boxed{22, 26}, 28$

$\longrightarrow \frac{22+26}{2} = \boxed{24}$

Quiz-2: A survey of number of pets in a town saw that - 30% people had 0 pets, 40% had 1 pet, 10% had 2 pets, 20% had 3 pets. What is the average number of pets?

$30 = 0$

$40 = 1$

$20 = 3$

$10 = 2$

$$\text{avg} = \frac{(0+0+0+\cdots+0)_{30} + (1+1+1+\cdots+1)_{40} + (3+3+3+\cdots+3)_{20} + (2+2+2\cdots+2)_{10}}{100}$$

$$= \frac{0\times 30 + 1\times 40 + 3\times 20 + 2\times 10}{100} = \frac{40+60+20}{100} = \frac{120}{100} = \boxed{1.2}$$

$$0.3\times 0 + 0.4\times 1 + 0.1\times 2 + 0.2\times 3 = \boxed{1.2}$$

weighted avg

Quiz-3: The mean weight of 2 children in a family is 40 Kgs. If the weight of the mother is included, the mean becomes 45. What is the weight of the mother?

x_1 = weight~~to~~ c1

x_2 = weight of c_2

x_3 = weight of m

$$\frac{x_1 + x_2}{2} = 40 \quad \text{———} \; \textcircled{1}$$

$$x_1 + x_2 = 80$$

$$\frac{x_1 + x_2 + x_3}{3} = 45$$

$$80 + x_3 = 45 \times 3$$

$$x_3 = (45 \times 3) - 80$$

$$\boxed{x_3 = 55}$$

=