Intermediate Stats

1. Measure of Central Tendency
2. Measure of Dispersion
3. Percentiles And Quartiles
4. 5 Number Summary (Box plots)

1 Measure of Central Tendency

Distribution

$\{23, 24, 28, 29, 31, 32, 33\}$

Central position

1. Mean ✓
2. Median ✓
3. Mode ✓

A measure of CT is a single value that attempts to describe a set of data by identifying the central position

1 Mean (Average)

$x = \{1, 2, 3, 4, 5\}$

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

$= \frac{15}{5} = \boxed{3}$

Population (N)

Sample (n)

$$\text{Population mean}(\mu) = \sum_{i=1}^{N} \frac{x_i}{N}$$

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

$n > N$ or $n < N$

Random choosing

Population data (N) $= \{24, 23, 2, 1, 28, 27\}$

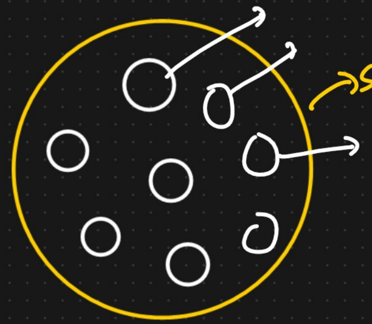
Sample data (n) $= \{23, 2, 28, 27\}$

$$\text{Population mean} = \mu = \frac{24+23+2+1+28+27}{6} = 17.5$$

$$\text{Sample mean} = \bar{x} = \frac{23+2+28+27}{4} = 20$$

Population Vs Sample

EXIT POLL



State (N) ⟹ population

Party is winning

$\boxed{n = 1000}$

Practical Application

DATASET

O NAN → Not a Number ⟹ Empty

Empty values

① Delete → Loss of Information

② Avg (column) → NAN

Age	Salary	Family Size
–	–	–
–	–	–
–	–	–
~~NAN~~	~~–~~	~~–~~
~~–~~	~~NAN~~	~~–~~
~~–~~	~~–~~	~~NAN~~
~~NAN~~	~~NAN~~	~~–~~

⟹ Analyze

$\{24, 26, \boxed{NAN}, 21, 20, 18\}$ ⇑ 21.8

Median

$$\text{Average} = \frac{24+26+21+20+18}{5} = 21.8$$

② Median

Range = (5−1) = 4

$\{\underbrace{1, 2, 3, 4, 5}\}$ ⟺ $\{1, 2, 3, 4, 5, \boxed{100}\}$

Range = (100−1) = 99

outlier

$$(\mu)\ \text{Average} = \frac{1+2+3+4+5}{5} = 3$$

$$\mu = \frac{1+2+3+4+5+100}{6} = \frac{115}{6} = 19.16$$

Defn outlier ?? :- Outlier is a number that is completely different than the entire distribution

$\{1, 2, \boxed{3, 4}, 5, 100\}$ ⟶ $\frac{3+4}{2} = 3.5$

$\{1, 2, \boxed{4}, 5, 100\}$ ⇓

Steps to find out Median

Median = 3.5

Median = 4

1. Sort the numbers
2. Find the central number

if no. of elements are even we find the average of central elements

if no. of elements are odd we find the central element

3. **Mode** : { Most frequent occuring element }

Dataset

Types of flowers

→ Lily

→ Sunflower

→ Rose

→ Lily

→ Rose

→ NAN ← Replace with something → Rose

→ Rose

→ Rose

→ Sunflower

→ Sunflower

Yes?

{categorical variable}

{ Mean, Median } X

→ [Rose, Sunflower] ←

OR

[Rose, Sunflower]

Randomly

Mean, (Avg) X outliers

Median → outliers

Mode → categorical replacement

① $X = \{24, 25, 26, 27, 28, 90, 100, 1000, 1200, 1400, 1400, 1400\}$

Average (μ) = 560

Median ( ) : 95

Mode ( ) : 1400 {Frequent occuring elements}.

	Age	Salary	
μ	23	40,000	μ median
median	24	41,000	Average
	28	72,000	Median
	27	NAN	
	NAN	18,000	
	31	24,000	
	$\boxed{58}$	$\boxed{10,00,00}$	
	NAN	50,000	

Age $\mu = 31.8$

Median: 27.5

② Measure of Dispersion

① Variance (σ^2)

② Standard Deviation (σ)

$X = \{1, 1, 2, 2, 4\}$ $\Leftrightarrow$ $Y = \{0, 2, 2, 3, 3\}$

No

$\mu = \frac{1+1+2+2+4}{5}$

$\boxed{\mu = 2}$

$\boxed{= \quad 2}$

$\{1, 1, 2, 2, 4\}$

① Variance

$\text{Population} \xrightarrow{} (N)$ variance $(\sigma^2) = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

μ → population mean

{Spread of the distribution}.

① Range

$A \rightarrow \{1, 2, 2, 3, 4, 5, 10\}$ ✓

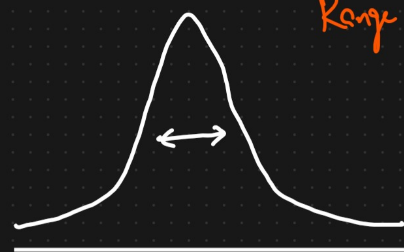
Range $= 10 - 1 = \boxed{9}$ }

$B \rightarrow \{2, 2, 4, 6\}$

Range $= 6 - 2 = 4$

Range will be higher

⇐⇒



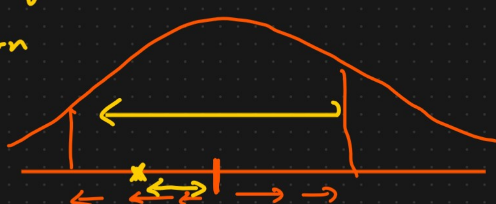
Range will be lesser

Variance will be higher

Variance will be lower

Measure of Dispersion

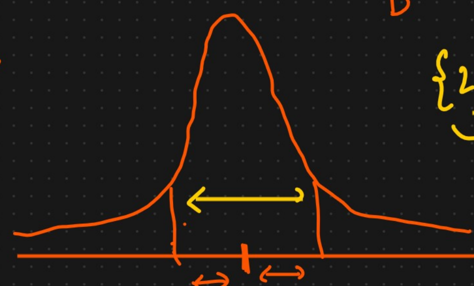
A



$$\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Variance is higher

B



$\{2, 2, 4, 6, 20\}$ 18

$\{2, 2, 4, 100\} \Rightarrow$ Range $= 100 - 2 = 98$

Population Variance

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Assignment

Sample Variance

Bessel's Correction

Degree of freedom

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

⇑

Assignment

Population Standard deviation

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

$$= \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

⇑ Sample

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

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

$$\begin{array}{r} 2.83 \\ -1.472 \\ \hline .358 \end{array}$$

$$\begin{array}{r} 2.83 \\ 1.472 \\ \hline 4.302 \\ 1.472 \\ \hline 5.774 \end{array}$$

$x = \{1, 2, 2, 3, 4, 5\}$

⇑ Sample

x	$\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$
1	2.83	−1.83	3.34
2	2.83	−0.83	0.6889
2	2.83	−0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
2.83			10.84

$$\sigma^2 = \frac{10.84}{5} = 2.168$$

$$\sigma = \sqrt{2.168} = \{1.472\}.$$

$\bar{x}$ $\quad (x-\bar{x})^2 \quad \sum(x-\bar{x}^2)/n-1$

$x = \{23, 21, 20, 19, 24, 27, 28\} \Rightarrow$ Sample

$[(23-23.14)^2 + (21-23.14)^2 + (20-23.14)^2 + (19-23.14)^2 + \quad]$

10 Min Break

$\bar{x} = 23.14$

Variance & Std ?

$\{ Var = 11.8 \quad \sigma = \sqrt{Var} = 3.44 \}$.

$\{\boxed{2, 3}, 4, 5, 6\}$

④ Percentiles And Quartiles

9th maths

Percentage = $\{1, 2, 3, 4, 5, 6, 7, 8\}$

Percentage of Even Numbers = $\dfrac{\text{No. of even Numbers}}{\text{Total no. of Numbers}} = \dfrac{4}{8} = \dfrac{1}{2} = 0.5$

⇓

50%

Percentage of odd numbers = $\dfrac{4}{8} = \dfrac{1}{2} = 0.5 \Rightarrow 50\%$

Percentiles: GATE, GRE, IELS, SAT, CAT ⇒ Percentile $\left\{ \begin{matrix} 99\% \\ 85\% \\ 84\% \end{matrix} \right\}$

Defn: A percentile is a value below which a certain percentage of observation lie

A ↓

99 percentiles → It means the person has got better marks than 99% of the entire students

Sorted Average DATASET

1 ⇒ 5%

Dataset: 2, 2, 3, 4, $\boxed{5}$, $\boxed{5}$, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, $\boxed{10}$, $\boxed{11}$, 11, $\boxed{12}$

1 2 3 4 5

⇓ 55%

$\boxed{25 \text{ percentile}}$ $\boxed{X}$

100 percentiles

What is the percentile ranking of 10 ??

$\boxed{100\%}$

$\boxed{99\%} \Leftrightarrow$ $\boxed{\dfrac{17}{20} \times 100} = 85$ percentile.

99% 100% CAT

99%.

Percentile Rank of x = $\dfrac{\#\text{ of values below } x}{n} \times 100$

20

55 percentile

$\Rightarrow \boxed{\dfrac{19}{20} \times 100 = 95 \text{ percentile}}$

$$\boxed{\frac{99}{100} \times (21)}$$

$\Downarrow$ n+1

$\frac{\cancel{16}^{4}}{\cancel{20}_{\cancel{5}\,1}} * \cancel{100}$ = 80 percentile

(2) What is the value that exists at 25 percentile

$\{ (n+1) \Downarrow \}$

$$Value = \frac{Percentile}{100} * n$$

$$= \frac{\cancel{25}^{\cancel{5}\,1}}{\cancel{100}\,\cancel{20}\,4} \times (20) = \frac{20}{4} = 5 \Rightarrow \text{Index}$$

$\boxed{5.5}$ Index

$$= \frac{\cancel{55}^{11}}{\cancel{100}_{\cancel{5}\,1}} \times \cancel{20}\,1 = 11^{th} \text{ Index}$$

40 percentile $\Rightarrow$

5.25

$\boxed{5^{th} \quad 6^{th}}$

$\Downarrow$

Avg

5 Numbers Summary

1. Minimum
2. First Quartile (25%) (Q1) ✓
3. Median
4. Third Quartile (75%) (Q3) ✓
5. Maximum.

$\Rightarrow$ Removing the outliers

$\boxed{n+1} \Rightarrow$ Reason

-5 -7 25 <13

-6

$\{\boxed{1}, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, \boxed{\cancel{27}}\}$ → outlier

25%

$7-3$

[Lower Fence ⟷ Higher Fence]

75 percentile

25 percentile

$IQR = Q3 - Q1$

$$\begin{cases} \text{Lower Fence} = Q1 - \boxed{1.5}(IQR) \\ \text{Higher Fence} = Q3 + \boxed{1.5}(IQR) \end{cases}$$

Inter Quartile Range (IQR)

$Q1 = (25 \text{ percentile}) = \frac{25}{100} \times \boxed{(n+1)} \Rightarrow$

$= \frac{\cancel{25}^{1}}{\cancel{100}_{4}} (\cancel{20}^{5}) = 5^{th}$ Index $= 3$

$Q3 = \frac{75}{100} \times (20) = 15^{th}$ Index $= 7$

$IQR = Q3 - Q1 = 7 - 3 = 4$

Lower Fence $= 3 - 1.5(4) = -3$

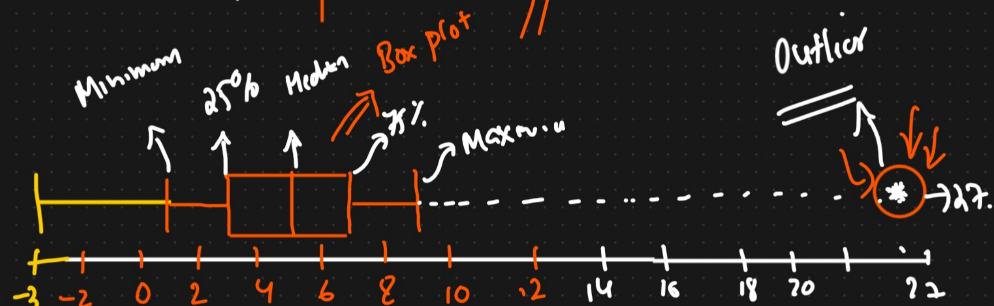
Higher Fence $= 7 + 1.5(4) = 13$

Set

$\{\boxed{1}, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27$

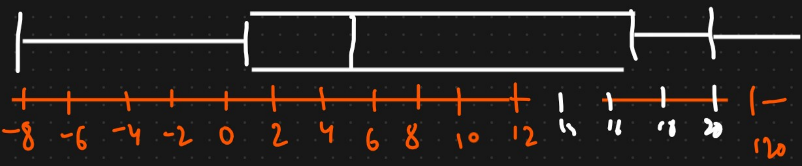
Box plot

Identifier outliers



1. Minimum = 1 ✓
2. Q1 = 3 ✓✓
3. Median = 5 ✓
4. Q3 = 7 ✓✓

⑤ Max = 9 ✓



5.25

0th 6th

$\{-8, 1, 2, 4, \boxed{5, 6}, 8, 15, 20, \boxed{120}\}$

$\frac{20+15}{2} = 17.5$

$\frac{5+6}{2} = 5.5$

- Minimum
- Q1
- Q3
- Median
- Maximum

Lower Fence = $Q1 - \boxed{1.5}(IQR)$

Higher Fence = $Q3 + \boxed{1.5}(IQR)$

$Q1 = 25 \text{ percentile} = \frac{25}{100} \times (11) = 2.75 \Rightarrow \text{Index} = 1.5$ (n+1)

$Q3 = \frac{75}{100} \times (11) = 8.25 = 17.5$

$IQR = 17.5 - 1.5$
$= 16.0$

Low Fence $= 1.5 - 1.5(16) = -22.5$

Higher Fence $= 17.5 + 1.5(16) = 41.5$

$\{-8, 1, 2, 4, \boxed{5}, \boxed{6}, 8, \boxed{15, 20}, \boxed{120}\}$

Avg $= \frac{1+2}{2} = 1.5$

python

$\frac{15+20}{2} = \frac{35}{2} = 17.5$ [Lower Fence ⟷ Higher Fence]

5 Number Summary

Lower Fence = $Q1 - (1.5)\, IQR$

Higher Fence = $Q3 + (1.5)\, IQR$

① Minimum = -8

② Q1 (First Quartile) 25 percentile ↳1.5

③ Median ⇒ 6.5

④ Q3 17.5

⑤ Maximum ⇒ 20

$Q1\ (25 \text{ percentile}) = \frac{25}{100} \times (11) = 2.75$ Index

$Q1 = 1.5$

$Q3\ (75 \text{ percentile}) = \frac{75}{100} \times (11) = 8.25$ Index

$\boxed{IQR} = Q3 - Q1$

$= 17.5 - 1.5$

$= 16$

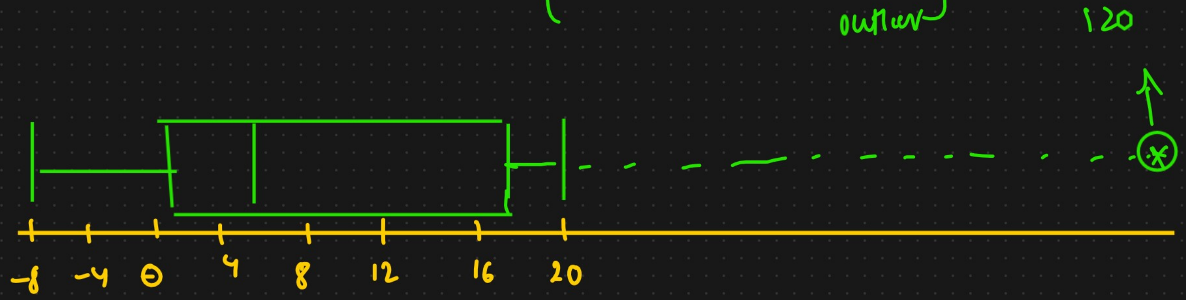
$Q3 = 17.5$

$\leftarrow [-22.5 \longleftrightarrow 41.5] \rightarrow$

Lower Fence $= 1.5 - 1.5(16) = -22.5$

Higher Fence $= 17.5 + 1.5(16) = 41.5$

{ Boxplot identify the outlier }



{ Shivan@ineuron.ai }