BASICS To Intermediate Stats

1. Measure of Central Tendency
2. Measure of dispersion
3. Gaussian Distr
4. Z score
5. Standard Normal Distr

1. Arithmetic Mean for Population & Sample

Mean (Average)

Population (N) → $X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 3.2$$

Sample (n)

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$= 3.2$$

2. Central Tendency

1. Mean ✓ 2. Median ✓ 3. Mode ✓

Refers to the measure used to determine the centre of the distribution of data.

{ 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100 }

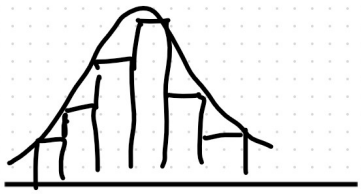
100 → Outliers

Mean = $\frac{32+100}{11}$

$= \frac{132}{11} = 12$

$\mu = 3.2$ + 100 → $\mu = 12$

Distribution



Median { 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112 }

1) Sort the numbers

12

$\mu = 3.2$

$\mu = 12$

Odd number = 11

$\mu = 12$

Median = 3

Median = 3.5

Avg = $\frac{3+4}{2} = 3.5$

Median = 3

Median 3.5

{ Median works well with outlier }

② Mode = { 1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 100, 100, 100 }

2 3

Mode = { Most frequent Element }

Mode : 6 → Measure of Central Tendency

Type of flower | petal length | petal width

Rose

Lilly

Sunflower

Mode

{ Missing Value → Most frequent occ element }

DATASET

10% Missing data

(Categorical Variable)

Ages of Students

Age
25
26
—
—
—
32
34
38

⇒ Mean? ✓
Median?
Mode ??

(F) Measure of Dispersion
① Variance
② Standard Deviation

{ Dispersion } ⇓ Spread

μ

{ 1, 1, 2, 2, 4 } → $\frac{10}{5} = 2$

{ 2, 2, 2, 2, 2 } → $\frac{10}{5} = 2$

① Variance ✓

Population Variance

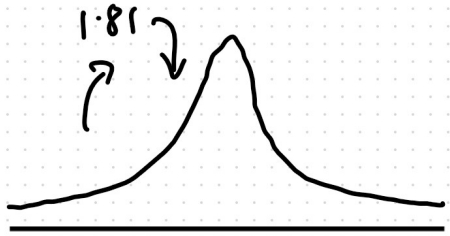
$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \frac{10.84}{6} = 1.81$$

Sample Variance

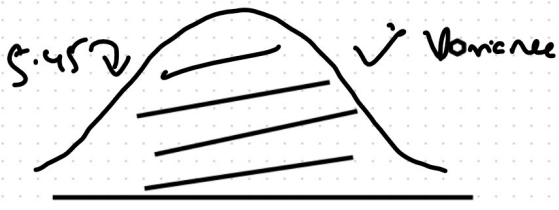
$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

x	μ	$x-\mu$	$(x-\mu)^2$
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
			10.84

$\mu = 2.83$



Variance is more ??

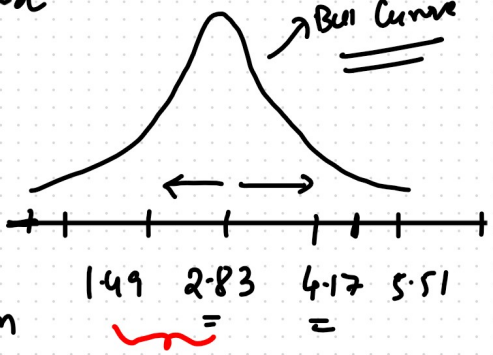


$\sigma = \sqrt{\text{Variance}} = \sqrt{1.81} = 1.345$

Variance = Spread

S.D = 5

↳ 1.5 σ from the μ



2.83 − 1.34 = 1.49

2.83 + 1.34 = 4.17

4.17 + 1.34 = 5.51

ⓨ Percentiles And Quartiles {Find outliers}

Percentage : 1, 2, 3, 4, 5

% of the numbers that are odd?

$\% = \frac{\#\text{ of numbers that are odd}}{\text{Total Numbers}}$

$= \frac{3}{5} = 0.6 = 60\%$

Percentiles (GATE, CAT, GMAT, SAT)

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

Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

5.25

11?

What is the percentile ranking of 10? n = 20

Percentile Rank of $x = \frac{\#\text{ of values below } x}{n} \times 100$ ($x = 10$)

$= \frac{16}{20} \times 100 = 80\%$

11?

$= \frac{17}{20} \times 100 = 85\%$

(2) What value exists at percentile ranking of 25%? 75%??

$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{25}{100} \times (21) = 5.25$ → Index position

5 → 25%

$= \frac{75}{100} \times (21) = 15.75 \Rightarrow 9$ Answer

↳ index

Five Number Summary

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

Removing the outliers

Outlier ?? [27 > 13]

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

25% 75%

$\frac{25}{100} \times (19+1)$

$\frac{25}{100} \times 20 = 5 \rightarrow$ index

$= 3 \Rightarrow$

so

[Lower fence ⟷ Higher fence]

Q1 = 3

Q3 = 7

Lower fence = Q1 − 1.5(IQR)

Upper fence = Q3 + 1.5(IQR)

Q3 = (75%)

Q1 = (25%)

Interquartile Range (IQR) = Q3 − Q1

= 7 − 3 = 4

Lower Fence = Q1 − 1.5 (IQR)

= 3 − 1.5 (4)

= 3 − 6 = −3

[Lower Fence ⟷ Higher Fence]

[−3 ⟷ 13]

Higher Fence = Q3 + 1.5 (IQR)

= 7 + 1.5 (4)

= 7 + 6 = 13

Remaining data

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

Minimum = 1

Q1 = 3

Median = 5

Q3 = 7

Max = 9

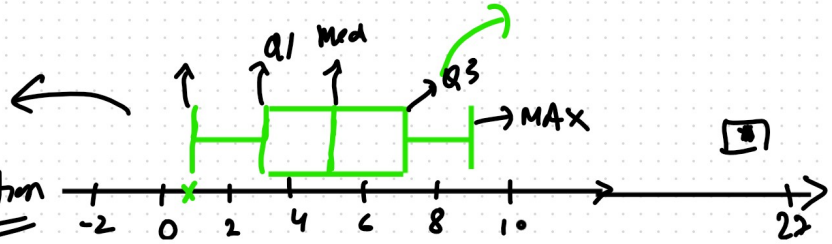
→ 5 Number Summary

⇓

Box plot

27

Box plot



DATA Visualization

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

$n-1$ → Bessel's correction
↳ Degree of freedom

STATS

Why sample variance is $n-1$?