Day 2 - Statistics

Agenda

1. Histograms ✓
2. Measure of Central Tendency ✓ } ⇐ 1.15 hrs
3. Measure of Dispersion ✓
4. Percentiles And Quartiles
5. 5 Number Summary (Box plot) }

① Histogram

Ages = { 0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }

1. Sort the Numbers
2. Bins → No. of groups
3. Bins size → Size of Bins

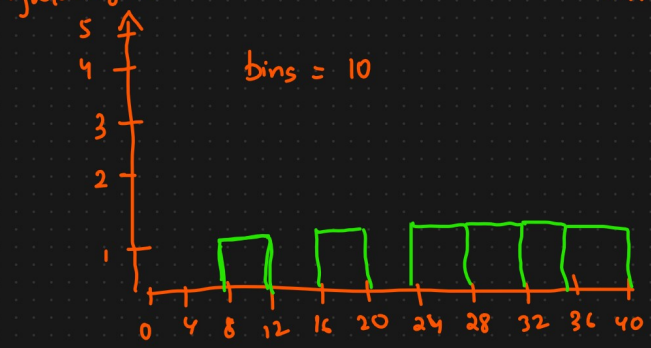
[10, 20, 25, 30, 35, 40]

min = 10

max = 40

bin size

$\frac{40}{10} = 4$ //



bins = 20

bins = 10

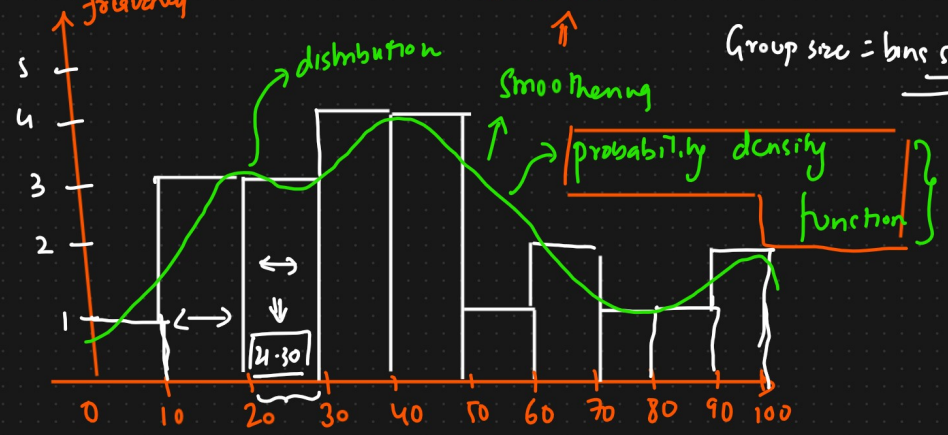
bin size = $\frac{100}{20} = 5$

bin size = $\frac{100}{10} = 10$

Group = bins

Group size = bins size

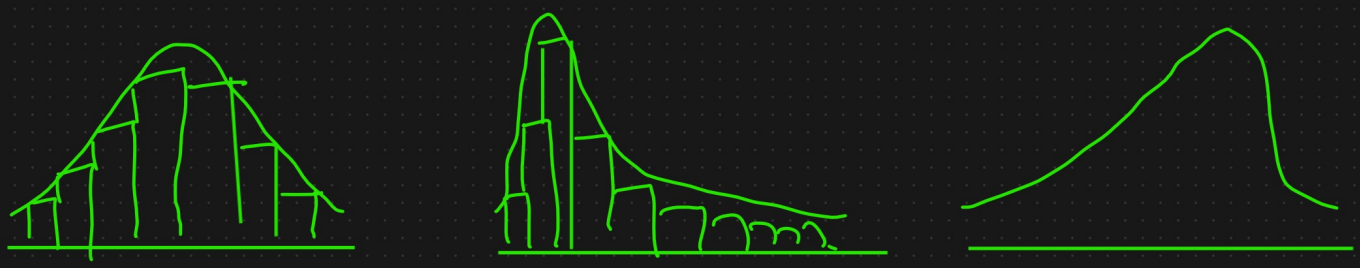
⇒ Kde



Bin size = $\frac{\text{Max} - 0}{\text{bins}}$

bins:

Ages = { 0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }



Assignment.

Weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

bins = 10

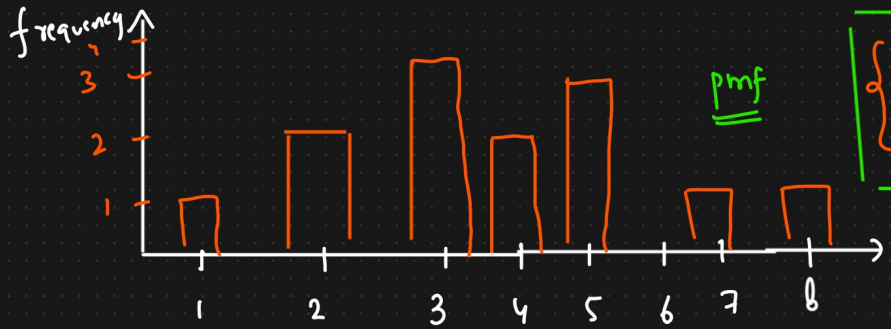
bin size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$

{Continuous value}

② Discrete :

pdf

No. of Banks accounts = [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]



pmf

{Probability mass function}

⇓ pmf

pdf : probability density function → Continuous

pmf : probability mass function → discrete.

Measure of Central Tendency

① Mean ✓

② Median

③ Mode.

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

Mean $X = \{1, 2, 3, 4, 5\}$ Average / Mean $= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

Population (N) $N \gg n$ Sample (n)

Population mean (μ) $= \sum_{i=1}^{N} \frac{x_i}{N}$ $n \not\approx N$

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

$N = 6$ $N > n$ $n = 4$

Population Age $= \{24, 23, 2, 1, 28, 27\}$ Sample Age $= \{24, 2, 1, 27\}$

$\{24, 23, 28, 27\}$ ← Average

$\bar{x} \neq \mu$

13
54
$\bar{x}_1$

Population mean (μ) $= \frac{24+23+2+1+28+27}{6}$ Sample mean ($\bar{x}$) $= \frac{24+2+1+27}{4}$

$\mu = 17.5$ $\bar{x} = 13.5$

$\mu \gtrless \bar{x}$

$\bar{x} \gtrless \mu$

np.nan ← Null values

Practical Application (Feature Engineering)

Age	Salary	Family Size
–	–	–
–	–	–
→ NAN	–	–
–	–	–
–	NAN	–
–	–	NAN
–	NAN	–
→ NAN	–	–

← Loss of Info

Age: 29.6 → 38

NAN ⇒ Not A Number

Null → Val = 10/4

Mean 10/4 = 1, 2, 3, 4 NAN NAN → mean

Age	Salary
24 ✓	45
28 ✓	50
29 ✓	NAN
NAN ×	60
31 ✓	75
36 ✓	80
NAN ×	NAN

Salary: 62 → 85

Outliers ∈ [80] [200] ∈ ∈

④ Median

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

↓outlier

$\bar{x} = 3 \longrightarrow \bar{x} = 19.16$

$\frac{1+2+3+4+5+100}{6}$

$= \frac{115}{6} = 19.16$

Steps to find out median

① Sort the Numbers

② Find the central number { ① if the no. of elements are even we find the average of central elements

② if the no. of elements are odd we find the central elements.

Sorted

{0,1,2,3,4,[5,6],7,8,100,120}

Mean = 25.6

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

median = 5

③ Mode : {Most frequent occuring elements}

{1,2,2,[3,3,3],4,5}

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

[2,3]

Dataset

Types of flower {Categorical variable}

Lily

Sunflower

Rose

NAN ← ROSE

Rose

Sunflower

Rose

NAN ← ROSE

[

FE

40% ⇒ Blood

Under 19

17, 18, 19, 16, 15, 32 → Outlier

(F) Measure of Dispersion

① Variance (σ^2) ← Spread of Data.

② Standard deviation (σ) ←

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

Variance

Population Variance (σ^2) { Degree of freedom

$$\sigma^2 = \sum_{i=0}^{N} \frac{(x_i - \mu)^2}{N}$$

{ Bessels Correction }

Sample Variance (s^2)

$$s^2 = \sum_{i=0}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

⇓

Assignment ⇐

{1-1000} = ⇓ First one.

{1-100} = Second one. ⇓

$\{1,2,3,4,5,6,7,8,9,10\}$ (1-10) $=$ Variance

$\{1,2,3,4,50,60,70,100\}$ (1-100) $=$ Variance

Variance {verify}.

$21 + 80 = 101$, $101/7$

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

$\mu = 3$

$\sigma^2 = \dfrac{[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]}{5}$

$= \dfrac{4+1+0+1+4}{5} = \dfrac{10}{5} = 2$

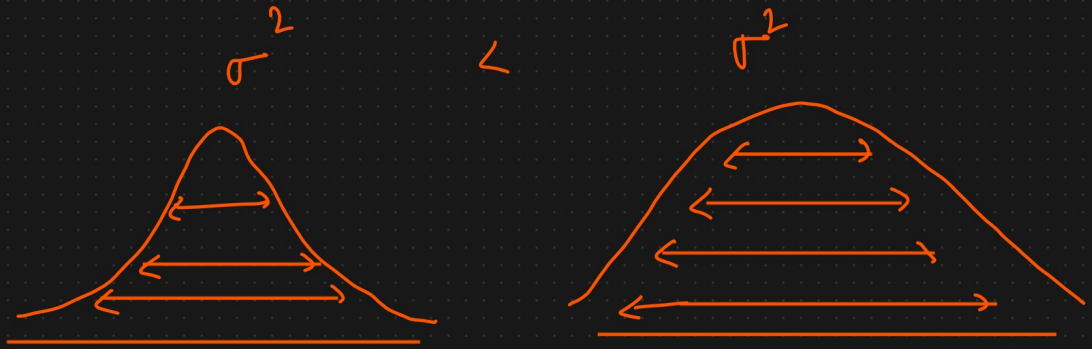
$\{1,2,3,4,5,6,80\}$

$\mu = 14.4$

$\sigma^2 = \dfrac{(1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (4-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + (80-14.4)^2}{7}$

Variance ⇑ Spread ⇑

$\sigma^2 = 719.10$



3.00
1.41
1.59

⑧ Standard deviation $\left(\sqrt{\sigma^2}\right) \Rightarrow \boxed{\sigma}$ $\boxed{4}$

$\{1, \boxed{2}, 3, \boxed{4}, 5\}$

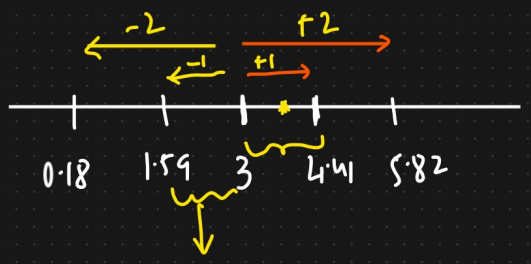
$\mu = 3$

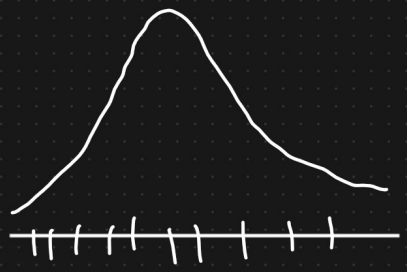
$\sigma^2 = 2$

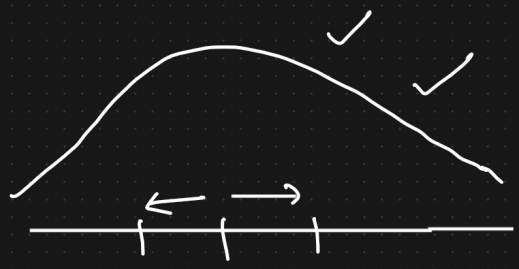
$\sigma = \sqrt{2} = 1.41$

3 ⇒ 0 Standard deviation

4.41
1.41
5.82







⑨ Percentiles And Quartiles

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

$$\text{Percentage of Even Number} = \frac{\text{No. of even Numbers}}{\text{Total no. of Number}} = \frac{4}{8} = 0.5 = 50\%$$

Percentiles : GaK, CAT, IELS, SAT, GRE, JEE, NEET ⇒ Percentiles

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

99 percentile = It means the person has got better marks than 99% of the entire students

Ascending order

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

Rank

What is the percentile rank of 10 ?

Yes

CAT Exam 99%

Next class

Percentile Rank of x = $\frac{\#\text{No. of value below } x}{n}$ = $\frac{16}{20}$ = 80 percentile.

=0.8

45 percentile

= $\frac{14}{20}$ = 70 percentile

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

95%

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

#n

$$= \frac{25}{100} \times 20 = 5^{\text{th}} \text{ Index}$$

Op = 5

$$= \frac{95}{100} \times 21$$

(5) 5 number Summary

1. Minimum
2. First Quartile (25 percentile) (Q1)
3. Median
4. Third Quartile (75 percentile) (Q3).

Box plot

=> Remove the outliers.

⑤ Maximum

{1, 2, 2, 2, 3, 3.3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 24}

24 ↓ outlier

15 16

5.25

[Lower Fence ⟷ Higher Fence]

[-3.65 ⟷ 14.25]

Lower Fence = Q1 − 1.5 (IQR)

Higher Fence = Q3 + 1.5 (IQR)

IQR = Q3 − Q1 (75 − 25)

Inter Quartile Range (IQR)

$Q1 = \frac{25}{100} \times 21 = 5.25$ Index = 3

$Q3 = \frac{75}{100} \times 21 = 15.75$ Index: $\frac{8+7}{2} = 7.5$

Lower Fence = $3 - (1.5)(4.5) = -3.65$

Higher Fence = $7.5 + (1.5)(4.5) = 14.25$

{1, 2, 2, 2, 3, 3.3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 24}

-5

5 Number Summary.

① Minimum = 1 ✓

② Q1 = 3 ✓

③ Median = 5 ✓

④ Q3 = 7.5 ✓

⑤ Maximum = 9 ✓

Box Plot

To Treat outliers