09 | TUESDAY OCTOBER 2018

SEPTEMBER 2018

M	T	W	T	F	S	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

STATISTICS

Descriptive Stats

↳ Measure of central tendency
↳ Measure of dispersion

"Summarising the data"

- Histogram
- Pdf
- Cdf
- Probability
- Permutation
- Mean, Median, Mode
- Standard deviation
- Variance

* Gaussian distribution
* Binomial "
* Bernauli "
* Parito "
* Standard Normal "
* Transformation & Standardization
* QQ plot
*

Inferential Stats

- Z - test
- t - test
- ANOVA test (F test)
- CHI SQUARE.
- Hypothesis testing [P value]
- Confidence interval

Notes

Statistics – Collect → Organise → Analyse.

09 { for better decision making }

10

Descriptive Stats –

11 organising & summarizing data

Inferential stats –

12 drawing a conclusion.

01 * Population (N) and Sample (n)

Sample → part of entire population.

02

Sampling Techniques –

03

(1) Simple Random Sampling

04 ↳ Every member has an equal chance of being selected

05

(2) Stratified Sampling.

06 ↳ Population is split into non-overlapping grps

Eg → Male & Female

Notes → Based on age group.

(3) Systematic Sampling

↳ Pick up every n^{th} person

④ Convenience Sampling

↳ only people having domain knowledge or interest.

Variable –

Quantitative
Age, ht.
- Discrete
  - no of bank acc
  - no of children
- Continuous
  - height (decimal v.)
  - amt of rainfall

Qualitative / Categorical
Gender, Blood Grp
[categories]

Variable Measurement Scale –
- Nominal → categorical data
- Ordinal → order of data matters not value
- Interval → order & value matter, No 0.
- Ratio

Notes

Frequency distribution –
Rose – 3
Lily – 4
:
:

↳ charts –

NOVEMBER 2018

M	T	W	T	F	S	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

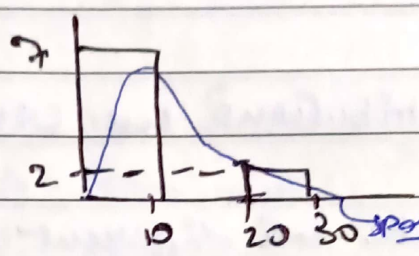
FRIDAY OCTOBER 2018 **12**

① Bar Graph → plot for discrete values.

② Histogram →

{ 1, 2, 3, 4, 5, 6, 10, 21, 23 }

Bin = 10



→ Plot for continuous values

Pdf. (probability density fn)

Pdf : Smoothening of histogram.

#Central Tendency - measure used to determine the center of distribution

Mean - Average

Population - N

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

Sample - n

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

Median - due to outliers mean changes a lot

Notes

∴ Median is considered

- Sort
- find middle no/ avg of

Mode - most frequent no.

(work with categorical variable)

Missed value can be replaced with most freq.

Measure of dispersion -

09

- Variance - σ^2
- Standard Dev. - σ

10

$\Rightarrow$ Spread of data.

11

* for different distributions we can have same mean. -

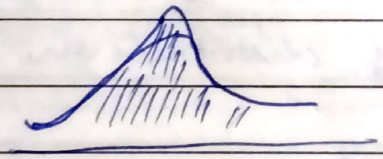
How are data sets different from each

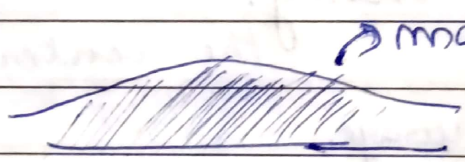
01 other - acc. of to their spreads.

02

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} \qquad s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

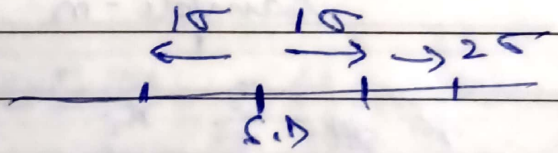
03





04

05



V = Spread of data

S.D = range of spread

Unit to tell the distance.

06

* Percentile & Quartile.

SUNDAY 14

Percentile $\Rightarrow$ $\frac{\text{no. of values below } x}{\text{total no of values}} \times 100$.

What value exist for a particular percentile -

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

↳ get index.

Five numbers Summary -

1. Minimum
2. Maximum
3. 1st Quartile
4. Median
5. 3rd Quartile

Removing Outliers -

[Lower fence ⟷ Higher fence]

Lower fence - $Q1 - 1.5(IQR)$

Higher fence - $Q3 + 1.5(IQR)$

IQR = Inter Quartile Range = $Q3 - Q1$

$Q3 \Rightarrow 75\%$. $Q1 \Rightarrow 25\%$.

Notes

Ex: [1, 2, 2, 2, (3), 3, 4, 5, 5, 5, 6, 6, 6, 6, (7), 8, 8, 9, 27]

$Q1 \Rightarrow \frac{25}{100} \times (19+1) \Rightarrow \frac{25 \times 20}{100} \Rightarrow (5)^{th} \Rightarrow 3$

$Q3 \Rightarrow \frac{75}{100} \times 20 \Rightarrow 15^{th} \Rightarrow 7$

$IQR = 7 - 3 \Rightarrow 4$

$L \Rightarrow 3 - 1.5 \times 4 \Rightarrow (-3)$

$R \Rightarrow 7 + 1.5 \times 4 \Rightarrow 13$

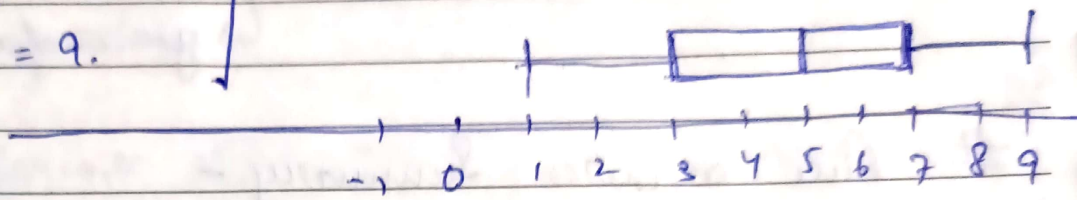
16 TUESDAY OCTOBER 2018

SEPTEMBER 2018

M	T	W	T	F	S	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Minimum = 1
$Q_1 = 3$
Median = 5
$Q_3 = 7$
Max = 9.

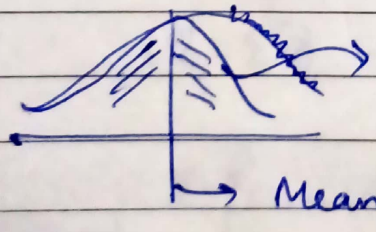
⇒ Box Plot.



Distributions -

- Normal / Gaussian
- Standard Normal dis.
- Z-score
- Log normal distribution
- Bernoulli "
- Binomial "

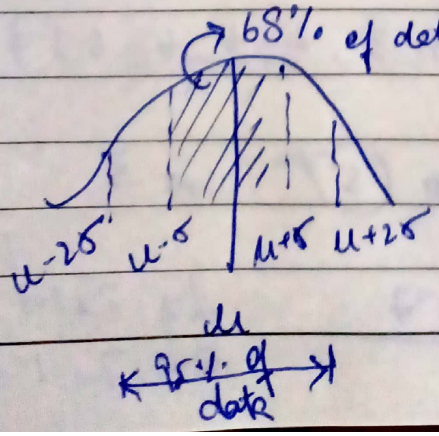
① Gaussian / Normal -



→ amt of data at both sides same.

→ Mean / Median / Mode

Notes



Empirical formula -

68 - 95 - 99.7

Example – Height - Normally distributed
cat → "
IRIS – "

Zscore – how much standard deviation is it away from the mean

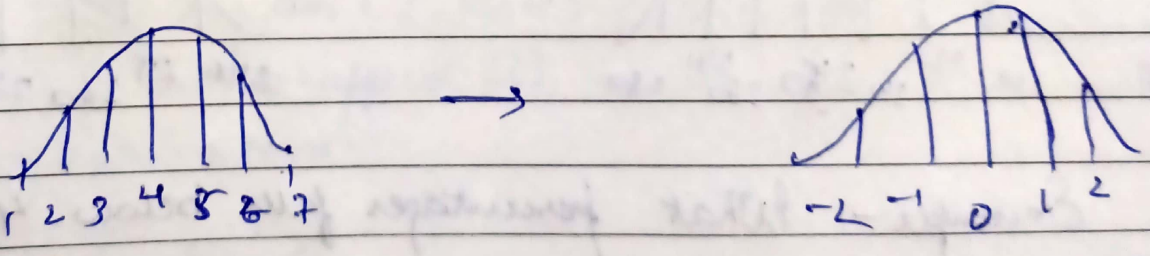
Zscore ⇒ $\frac{x_i - \mu}{\sigma}$

Say $\sigma = 1$ $\mu = 4$

for $x = 4.75$ $\frac{4.75 - 4}{1} \Rightarrow 0.75$

+ve → Right direct^n -ve – left direct^n.

Standard Normal distribution –
On applying Zscore to all values of sample the distribut^n obtained is called SND



Notes

$\{1, 2, 3, 4, 5, 6, 7\} \longrightarrow \{-3, -2, -1, 0, 1, 2, 3\}$

$\mu = 0$

$\sigma = 1$

Standardizat^n – Conversion into Standard Normal distrib.

18 THURSDAY OCTOBER 2018

SEPTEMBER 2018

M	T	W	T	F	S	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Normalizat^n → convert to (0 to 1) data set.

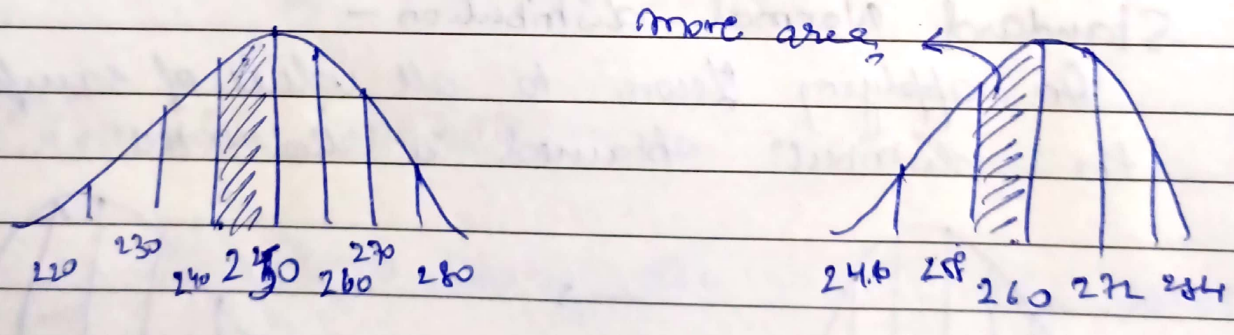
✓ Min Max Scalar
↳ (0,1)

Example –

	2021	2022
Series aug	250	260
S.D.	10	12
Pcam score	240	245

Which score is better

Z score → $\frac{240-250}{10} \Rightarrow -1$ $\quad$ $\frac{245-260}{12} \Rightarrow -1.25$



Example – What percentages fall below 4.25.

Notes

* find Z - score → $\frac{x-\mu}{\sigma}$

Look at Z table to find the area

= % of score fall above | below

Example - Avg IQ = 100
σ = 15
IQ lower than 85?

09

10

Z-score → $\frac{85-100}{15} \Rightarrow \frac{-15}{15} \Rightarrow -1$
↳ Look into table

11

12 Removing Outliers using python

01 outliers=[]

02 def detect_outliers(data):
threshold=3

03 mean = np.mean(data)
std = np.std(data)

04

for i in data:

05 Zscore = (i - mean)/std
if np.abs(z-score) > threshold:

06 outlier.append(i)
return outliers.

Notes IQR: $Q_3 - Q_1$

① Sort the data
② Calculate Q_1 & Q_3
③ IQR ($Q_3 - Q_1$)
④ find lower fence
⑤ find upper fence.

dataset = sorted(dataset)

09 Q_1, Q_3 = np.percentile(dataset, [25, 75])

10 IQR = $Q_3 - Q_1$

lowerfence = $q_1 - (1.5 * IQR)$

11 higherfence = $q_2 + (1.5 * IQR)$

12 **Probability** –

measure of likelihood

01 **Addition Rule** –

✓ Mutual Exclusive – cannot occur at same time

02 $P(A \text{ or } B) = P(A) + P(B) = 1$

03 ✓ Non – mutual Exclusive

04 $P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

05 **Multiplication Rule** –

✓ Independent Event –

06 one event not dependend on another event

$P(A \text{ AND } B) = P(A) * P(B)$

✓ Dependent Events –

SUNDAY 21 E_1 -> 3 Red marbel 2 White marbels

$P(1) \rightarrow 3/5, 2/4$

Name Bayes $P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$

09 ~~At~~ Permutation - ${}^{n}P_r = \frac{n!}{(n-r)!}$ ✓

10 Combination -
unique w.r.t. elements (no swaps allowed)

11

12 ${}^{n}C_r = \frac{n!}{(n-r)!\, r!}$

01 P value :
P = 0.8 out of 100 ; 80 times will be this.

02

03 * Influential Statistics -

04 Eg → Test whether a coin is fair or not by performing test for 100 times.

05 if 50 times H comes then → fair -

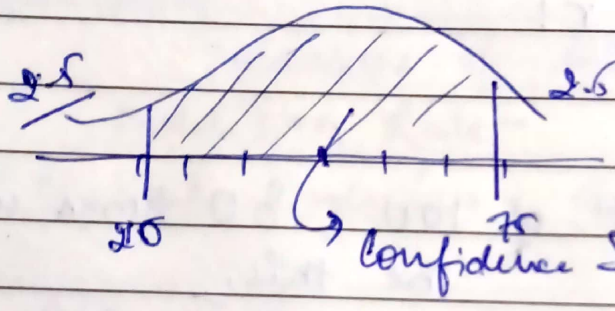
06 Hypothesis testing -
(1) Null Hypothesis → coin is fair
Notes (2) Alternate " → coin is unfair
(3) Experiment.
(4) Reject / Accept null hypothesis.

Say → 30 times heads,

"define how a far can the value be away from the mean"

↳ Significant Value - $\alpha = 0.05$ (5%)

$100 - 5 = 95\%$ ≡ Confidence Interval.



defined by domain expert.

Notes