Statistic

→ Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data.

Descriptive Stats

① Analyzing data, Summarizing data, Organizing the data. Provide whole info. of data, in the form of "Numeric and Graph"

② Bar plot, Histogram, pie chart PCDF, PDF.

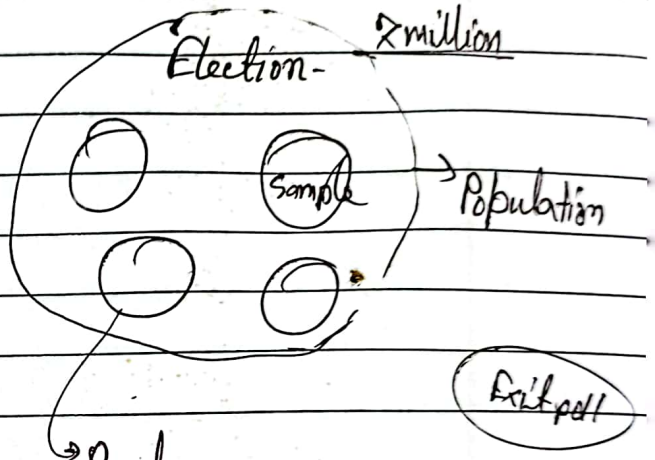
"Visualization"

③ Measure of central tendency "Mean, median, mode"

④ Measure of variance.

⑤ All Distribution function

Inferential stats



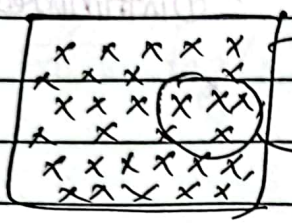
from entire population taking some sample, we are try inferencing some info- and come to the conclusion which party win (chance)

→ Z-test, T-test, chi-square test.

→ Confidential interval.

→ Population - the whole group of entire observation or data.

→ Sample → A Collection of data from part of the Population.



→ Population

→ Sample.

mean $\mu = \frac{\Sigma x_i}{N}$

mean $\bar{x} = \frac{\Sigma x_i}{n}$

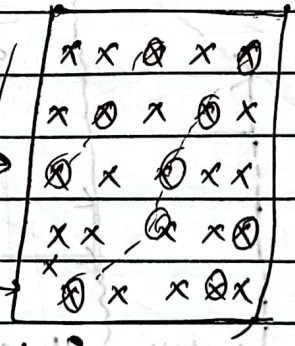
Random variable

① Increate ② Continious.

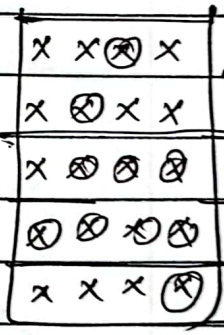
How we choose the sample from population?

Best

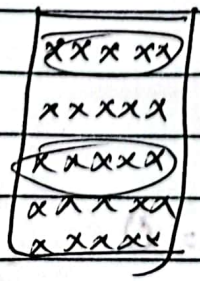
① Random Sample (Pick Randomly from the list)



② Systematic Sample (Such as every 3rd). follow a pattern



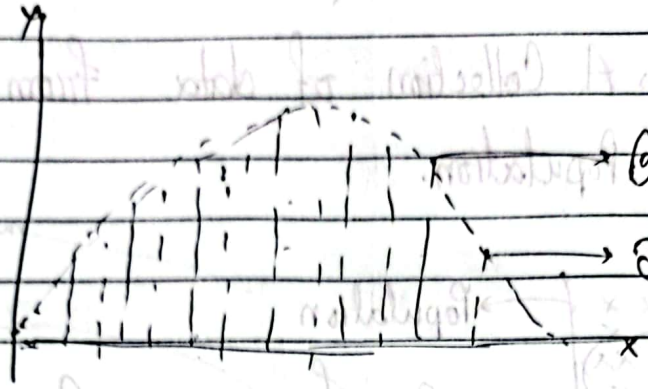
③ Stratified Sample (Randomly, but in ration to group size)

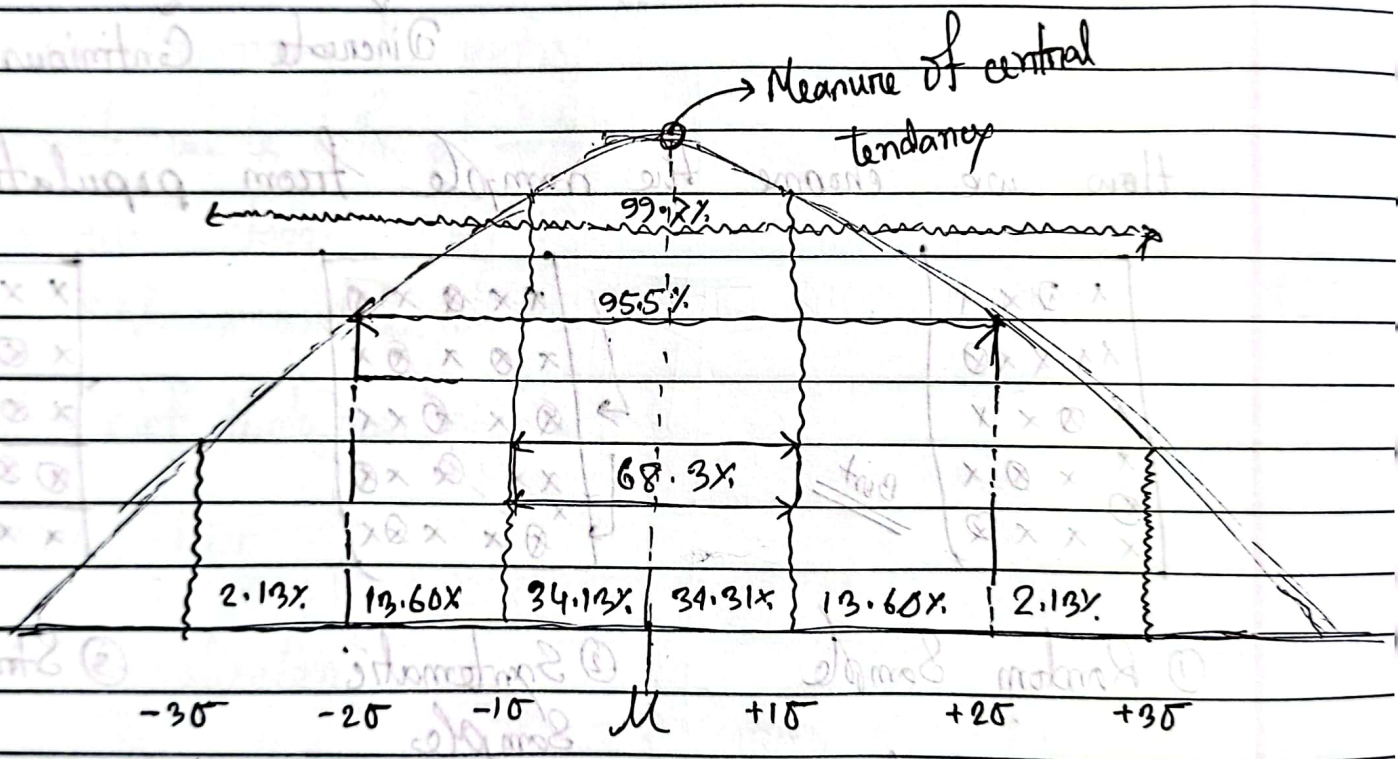


④ Cluster Sample

Choose whole groups randomly.

* Gaussian Distribution or Normal Distribution.





σ → Standard deviation

Outlier.

0 → Skew

Emperical Formula.

Standard Normal Distribution

$\mu = 0$

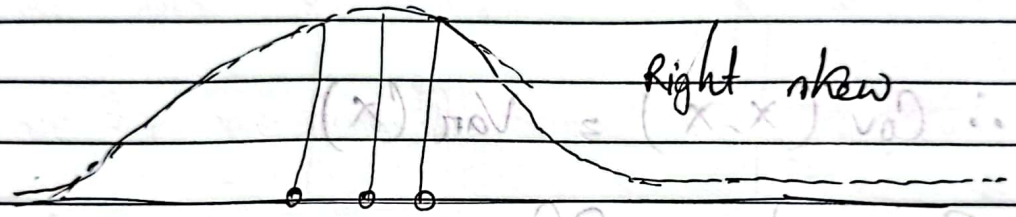
$\sigma = 1$

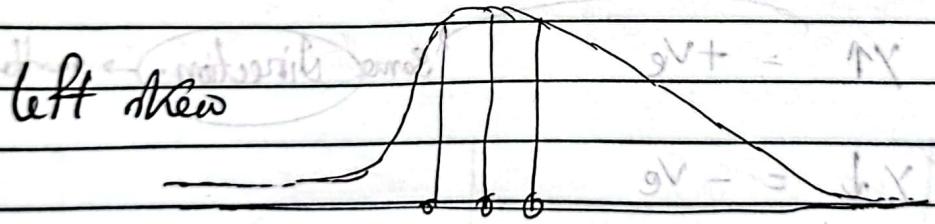
$$\frac{x_i - \mu}{\sigma}$$

Log Normal distribution.

we have Random variable x.

→ Normal Distribution belongs to log normal distribution if log(x) is Normally distributed.





log Normal Distribution.

Feature 1

follow log ND

↓

Standard Normal Distribution

↓

Scale the feature ← them.

Feature 2

follow GeND or ND

↓

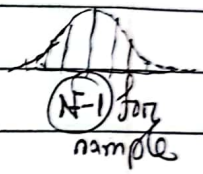
Standard Normal Distribution

→ Two Random variables x, y.

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$Var(x) = \frac{\Sigma (x_i - \mu_x)^2}{N} \qquad \sigma = \sqrt{v}$$

$$= \frac{\Sigma (x_i - \mu_x)(x_i - \mu_x)}{N}$$



$\therefore Cov(x,x) = Var(x)$

কীভাবে তাদের গুলো বিচ্ছিন্ন অবস্থায় থাকে বা রয়েছে → variance.

$x\uparrow \quad y\uparrow = +ve$ → Same Direction. → matter.

$x\uparrow \quad y\downarrow = -ve$

Pearson Correlation Co-efficient

① $$Cov(x,y) = \frac{\Sigma (x_i - \mu_x)(y_i - \mu_y)}{N}$$

② Pearson CC: $$\rho_{(x,y)} = \frac{Corr(x,y)}{\sigma_x, \sigma_y}.$$

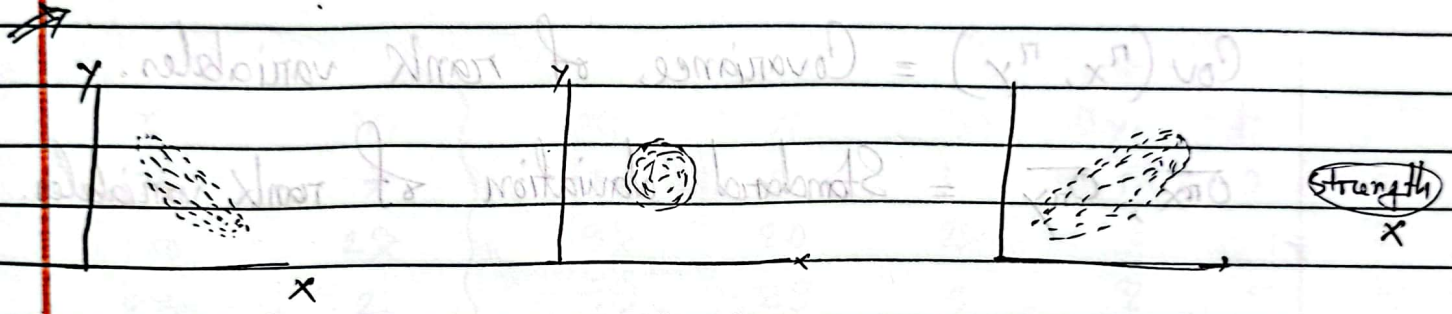
Covariance ⇒ Direction.

+ve and -ve value, but how much may be $-\infty$ ---$+\infty$ (actually we dont know)

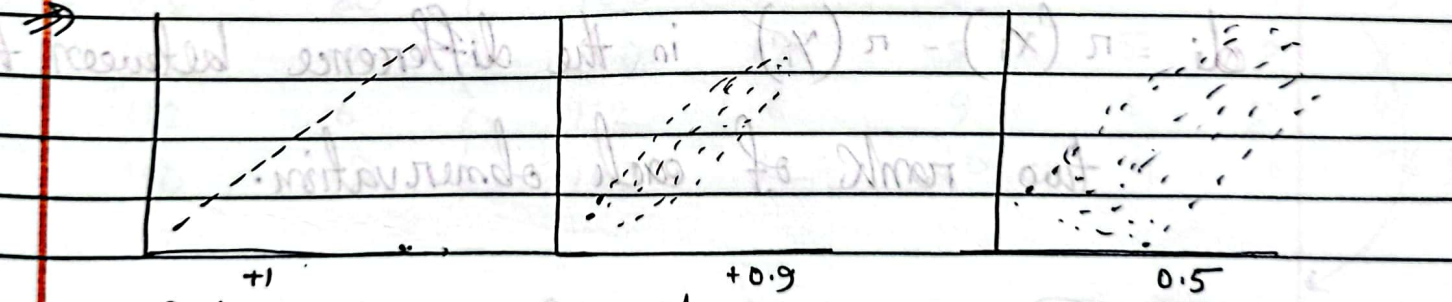
Correlation → "Direction" and "Strength"

+ve and -ve correlation

-1 to 1.

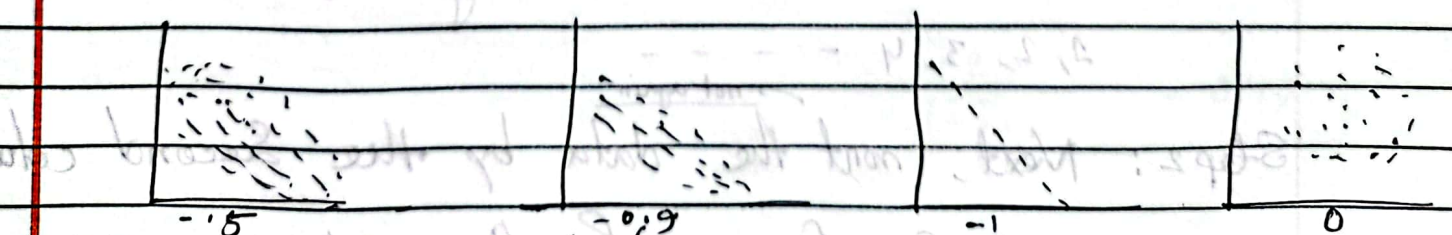


Cov (x, y) < 0 | Cov (x, y) ≈ 0 | Cov (x, y) > 0



+1 | +0.9 | 0.5

Perfect +ve correlation | High +ve Correlation | Low +ve Correlation



-.5 | -0.9 | -1 | 0

Low -ve corrn. | high -ve corrn. | Perfect -ve corre | No Corrn.

"Spearman Correlation."

(wikipedia.)

Random variable X, Y

$$r_s = \rho_{(r_X, r_Y)} = \frac{Cov(r_X, r_Y)}{\sigma_{r_X} \cdot \sigma_{r_Y}}$$

r = rank.

ρ = Pearson correlation co-efficient.

$Cov(r_X, r_Y)$ = Covariance of rank variables.

$\sigma_{r_X}, \sigma_{r_Y}$ = Standard deviation of rank variables.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$d_i = r(x_i) - r(y_i)$ in the difference between the two ranks of each observation.

Step 1: Sort the data by the first column (x_i). Create a new column r_{x_i} and assign it the rank $1, 2, 3, 4 - - - -$ → not again.

Step 2: Next, rank the data by the second column y_i. Create a fourth column r_{y_i}.

and Similarly assign it the rank values. 1, 2, 3 ---- n

③ Create a fifth column d_i to hold the difference between to rank Columns (r_{x_i}, r_{y_i}).

④ Create one final Column d_i^2 to hold the value of column d_i squared.

x_i	y_i	x_i	y_i	r_{x_i}	r_{y_i}	d_i	d_i^2
106	7	86	2	1	1	0	0
100	27	97	20	2	6	-4	16
86	2	99	28	3	8	-5	25
101	50	100	27	4	7	-3	9
99	28	101	50	5	10	-5	25
103	29	103	29	6	9	-3	9
97	20	106	7	7	3	+4	16
113	12	110	17	8	5	3	9
112	6	112	6	9	2	7	49
110	17	113	12	10	4	6	36

$$\rho(r_x, r_y) = 1 - \frac{6 \times 194}{10(10^2-1)}$$

$$= \frac{-29}{165} = 0.17$$

→ Mean, Median, Mode → Math is fun.

∴ outlier থাকলে remove or use Median.

কোথায় কি ব্যবহার হবে Krish NAIK.

→ Central Limit Theorem.

Population mean $\mu = \frac{\Sigma x_i}{N}$

Sample mean $\bar{x} = \frac{\Sigma x_i}{n}$

⇒ Actually আমরা যে ডাটাসেটগুলো নিয়ে কাজ করি সেগুলো Sample ডাটা।

⇒ Random variables →
- Numerical →
  - Discreate
  - Continuous.
- Categorical →
  - Nominal
  - Ordinal.

Central Limit Theorem.

$X \neq GD(\mu, \sigma^2)$

S_1 ---------- $x_{30} = \bar{x}_1$

S_2 ---------- $= \bar{x}_2$

S_3 ---------- $= \bar{x}_3$

S_4 ---------- $= \bar{x}_4$

S_5 ---------- $= \bar{x}_5$

$\bar{X} \approx GD\left(\mu, \frac{\sigma^2}{n}\right)$

Chebyshev Inequality.

if $X \approx GD(\mu, \sigma)$ — X = Random variable follow GD

$P_r(\mu - \sigma \le x \le \mu + \sigma) \approx 68\%$

$P_r(\mu - 2\sigma \le x \le \mu + 2\sigma) \approx 95\%$

$P(\mu - 3\sigma) \le x \le \mu + 3\sigma) \approx 99\%$

if $Y \not\approx GD$ — Y = Random variable not follow GD

$$P_r(\mu - k\sigma < x < \mu + k\sigma) \ge 1 - \frac{1}{k^2}$$

$k = 1, 2, 3 -$

Let, $k = 2$,

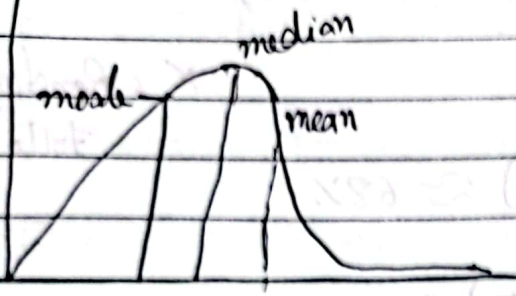
$$P_r((\mu - 2\sigma) < x < (\mu + 2\sigma)) > 1 - \frac{1}{2^2}$$

$= 75\%$

Skewness.

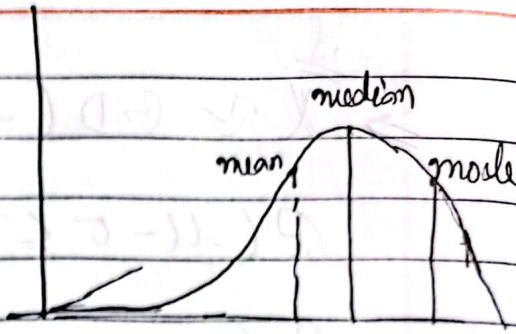
→ In Probability theory and Statistics, skewness is a measure of the asymmetry of the probability distribution of real-world valued random variable about its mean. The skewness value can be +ve, -ve, 0.

⇒ mean > median > mode

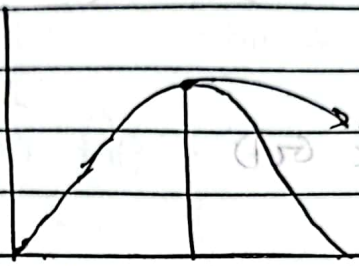


Right skew or Positive skew

⇒ mode > median > mean.



Left skew or Negative skew

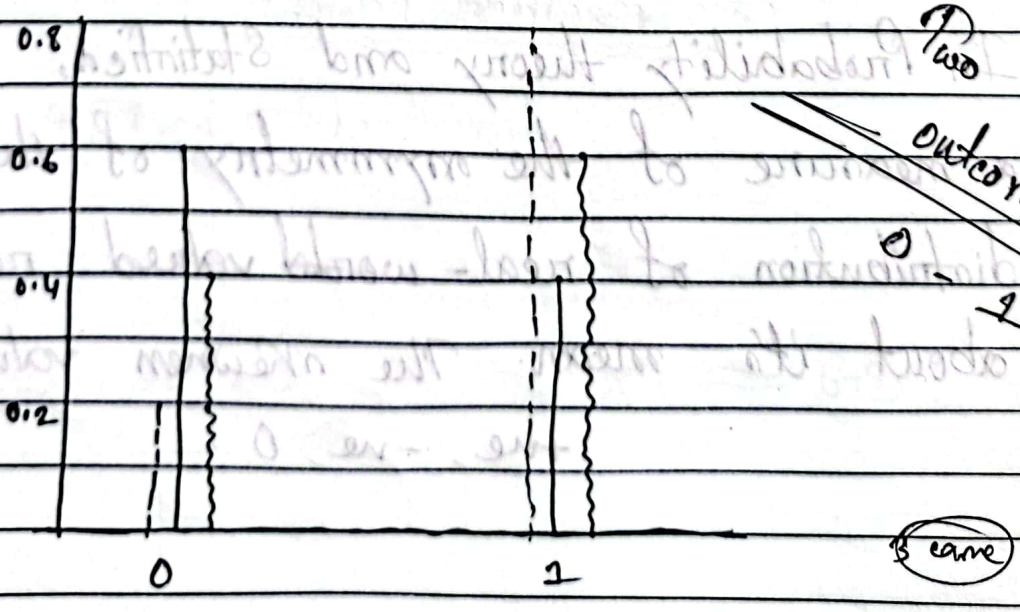


Symmetrical Distribution.

⇒ mean ≈ median ≈ mode.

Bernouli Distribution.

Probability mass function.



-----→ $P(x=0) = 0.2$ and $(P(x=1) = 0.8$

——→ $P(x=0) = 0.6$ " $(P(x=1) = 0.4$

~~~~→ $P(x=0) = 0.4$ " $P(x=1) = 0.6$

$$PMF = \begin{cases} p & \text{if } k=1 \\ q = (1-p) & \text{if } k=0 \end{cases} \qquad p^k (1-p)^{1-k} = (P(x=k)$$

→ Discrete value.

$\therefore P(X=x) = p^x (1-p)^{1-x}$

outcome continuous value → pdf → Probability Density function.

$x = 0$ or 1.

$P(\text{Success}) = p$

$P(\text{fail}) = 1-p = q$

$P(x=0) = p^0 (1-p)^{0|1-0} = 1-p = q$

$P(x=1) = p^1 (1-p)^0 = p$

"Interview Question"

Mean, variance, Std.

① Mean. $E(x) = \sum_{i=1}^{n} x \cdot p(x)$

$= (0 \times 0.4) + (1 \times 0.6)$

$= 0.6$

$= p$

$x = 0, 1$

$P(x=0) = 0.4$

$P(x=1) = 0.6$
~~~~

② Variance = $p(1-p) = pq$

③ Std = $\sqrt{pq}$

④ if. $q > p$ median = 0

$p = q$ median = 0.5

$p > q$ median = 1

⑤ Mode?

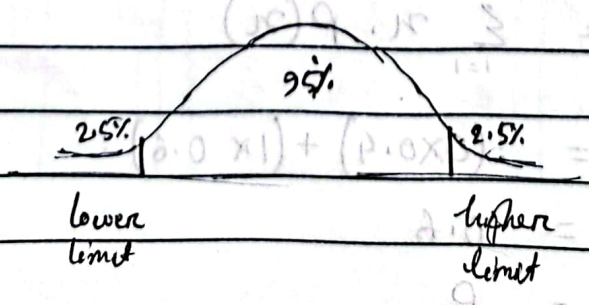
Confidence Intervals.

→ Point Estimate : $\bar{x}$ → μ → Population mean.

→ Range of values for an unknown parameter.

→ 95% Confidence level is used.

lower limit — higher limit.



→ Average size of shanks in the sea.

take sample.

Suppose

Population std $\sigma = 100$

sample $n = 30$

mean $\bar{x} = 500$

C.I = Point estimate ± margin error.

$$= \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$= 500 \pm z_{.05/2} \frac{100}{\sqrt{n}}$$

α = Significance value

$= 5\% = 0.05$

tail point.

Z-table

$1 - 0.025 \leftarrow z_{0.025}$

$= 0.925$

$$= 500 \pm 1.96 \times \frac{100}{\sqrt{30}}$$

lower limit $= 500 - 1.96 \times \frac{100}{\sqrt{30}} \approx 386$

higher limit $= 500 + 1.96 \times \frac{100}{\sqrt{30}} \approx 613$

$\mu \rightarrow 95\%$

$386 \longleftrightarrow 613$

when z in not given apply t-test

5 Number Summary and handle outliers using IQR. → github → codebasics.
z-test
t

① Minimum → 0%
② 25%
③ Median → 50%
④ 75%
⑤ Maximum → 100%

Percentile.

1, 2, (3) 4, 5, (6), 7, 8, 99, 100

Percentile = n% * (n+1)

25% = .25 * (12+1)
= $\frac{1}{4} \times \frac{13}{1}$
= 3.25 → 3

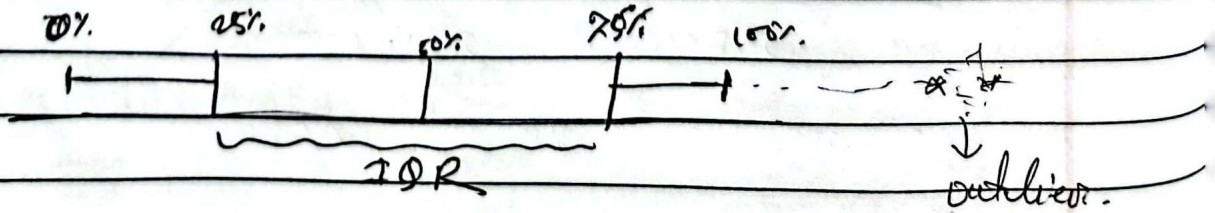
50% = $\frac{13}{2}$ = 6.5 → 6

handle Outliers → ① IQR → Percentile (50%)
② z-test
③ σ - standard deviation

IQR = $Q_1 - Q_3$

Low range = $Q_1 - 1.5$ (IQR)
High " = $Q_3 + 1.5$ (IQR)

} out of this range outliers.



P-Value

→ It is the probability for the "Null hypothesis" to be true.

→ A P value used in "Null hypothesis" to help you support or reject the null hypothesis. The P value is the evidence against null hypothesis. The "smaller" the P-value, the stronger the evidence that you should "reject" the null hypothesis.

(statisticshowto.com)

Null hypothesis : Treats everything same or equal.

H_0 = The coin is fair.

H_1 = The coin is not fair.

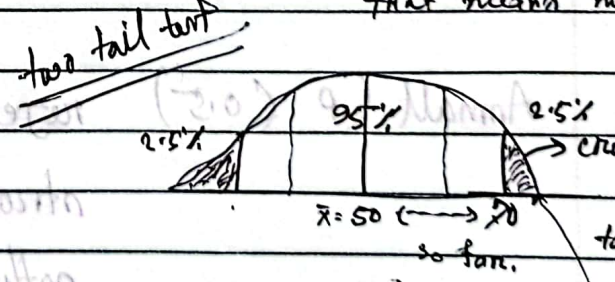
P = 0.05

Significance value

Toss experiment 100 time.

↳ 50 time (head).

that means mean = 50



5% = 0.05

(65) times

(30) times

Null hypothesis true

So we have fair coin.

Head (25) (20) (29)

away from the mean.

not fair

reject null hypothesis.

if $p > .10$ → no Significant

if $p \leq .10$ → marginally Significant.

if $p \leq 0.5$ → Significant.

if $p \leq 0.1$ → highly Significat.

P value vs Alpha level..

→ Alpha levels are controlled by the researcher and related to confidance level. You Get an alpha level by Subtracting your confidance level from 100%. For example, for if you wan to be 98 Percent confidant in your reresearch, the alph level would be 2% (100-98). when you ran the hypothesis test, the test will give you a value for p.

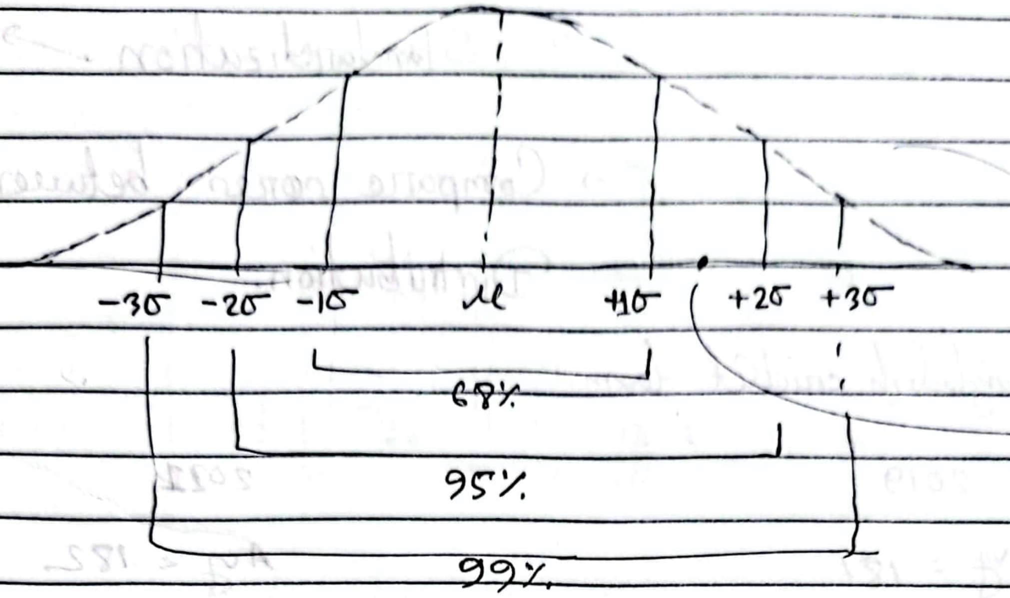
Asmall ($p \leq 0.5$) reject the null hypothesis. This is strong evidance that the null hypothesis is invalid.

Alarge (> 0.05) means the alternative hypothesis is weak, so we do not reject the H_0.

Variance $\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$

Standard $\sigma = \sqrt{\Sigma(x_i - \mu)^2 / N}$

Z - Score



Z-score table

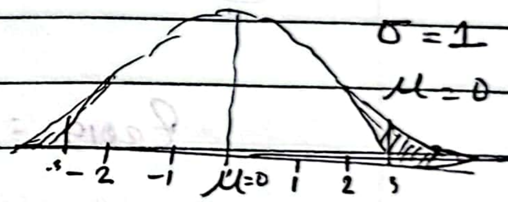
given, μ σ → ?

Convert it

SND

Standard Normal Distribution.

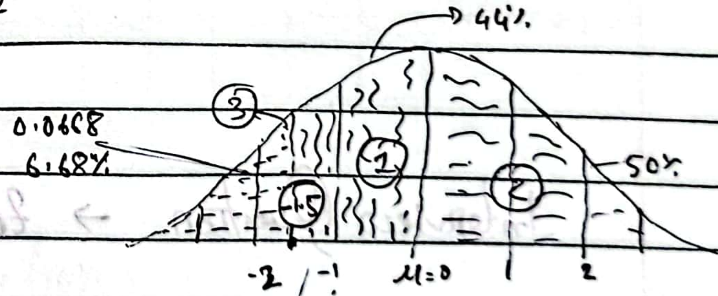
$$Z\text{-score} = \frac{x_i - \mu}{\sigma}$$



1, 2, 3, 4, 5

$\mu = 3$ z = -2, -1, 0, 1, 2

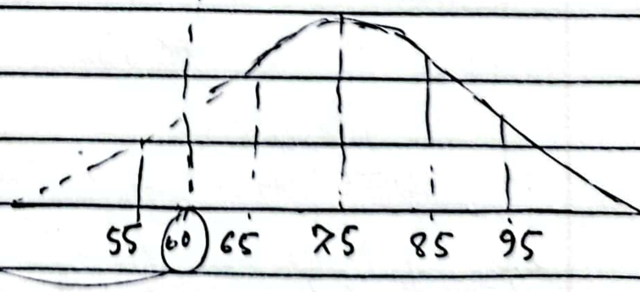
$\sigma = 1$



lets Problem $\mu = 25$

$\sigma = 10$

→ $P(x > 60)$



এই problemটি business dir. দিয়ে solve করা যাবে।

100 = x + 50% + 6.68

x = 100 − 56.68

≃ 44%

Z-score - Application -> [Outlier remove]

-> Standardization → same scale.

-> Compare scores between different Distribution.

Bangladesh cricket team.

2019	2021
Avg = 181	Avg = 182
σ = 12	σ = 5
Total score = 187 (final)	Total score = 185 (final)

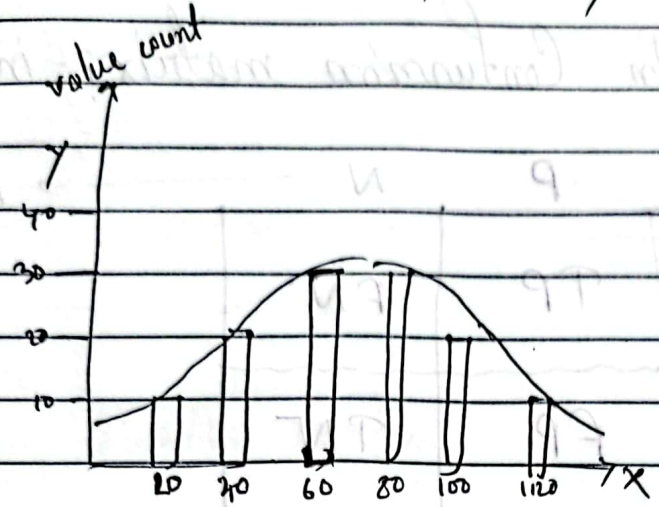
$$Z_{2019} = \frac{182-181}{12} = 0.5$$

$$Z_{2021} = \frac{185-182}{5} = 0.6$$

Better Performance

Interview Question → Z-score - Application.

⇒ Probability Density Function.

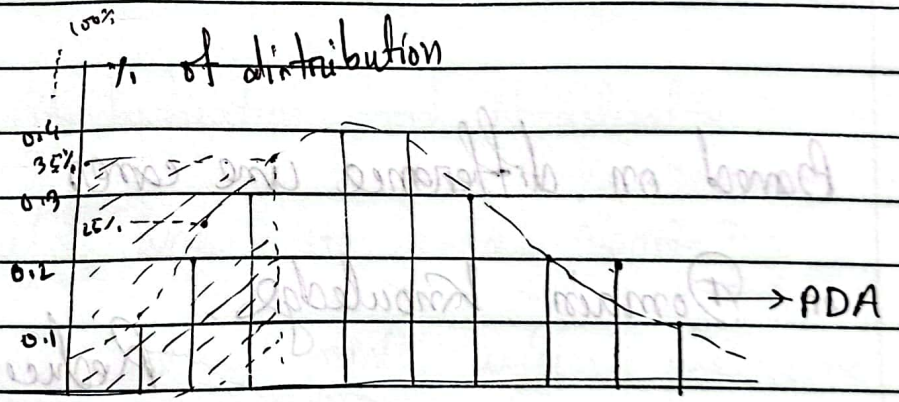


But

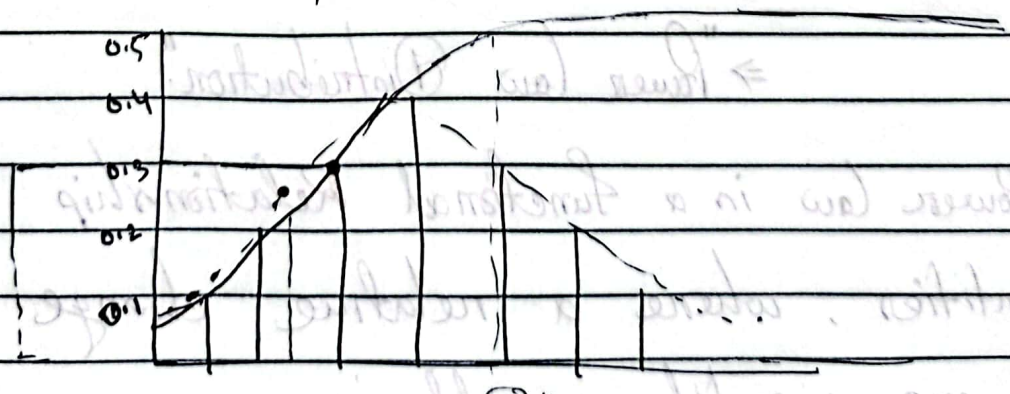
Its a Histogram!

But in a Probability Density Function. in Y curve axis not a value of count, their in percentage of Distribution.

Area



Cumanalative Density Function.

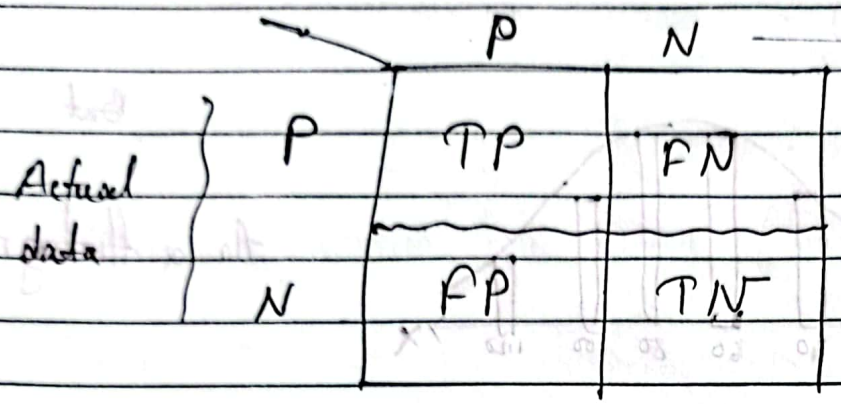


0.1+0.2+0.3 = 60%

50% data under 150 ng

অর্থাৎ, 50% data > 150 ng.

Weather we should Reduce false Positive or Negative In Confusion matrix? - interview Q?



(TP) ↑ (TN) ↑ (FN) ↓ (FP) ↓

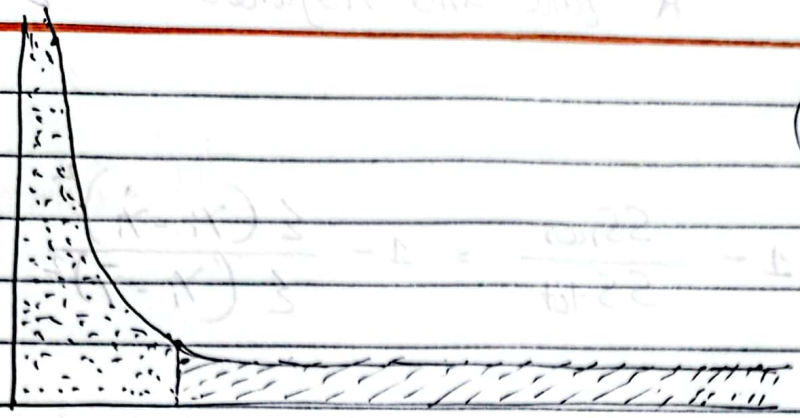
Both are important.

Based on difference use cases

on (1) Domain Knowledge, Reducing (FP) ↓ then (FN) ↓

⇒ "Power Law (1) Distribution"

A Power Law is a functional Relationship between two quantities, where a relative change in one in one quantity result in a proportional relative change in another quantity.



Pareto Principle
80/20 rule.

80% result get
20% effort.

Interview Question. Example.

Main focus → 20% Effort.
get 80% result.

Standardization and Normalization.

→ why we should use std ... and Norm.

→ when we use std --- and when we use Norm.

Brief - towardsai.net. most important topics

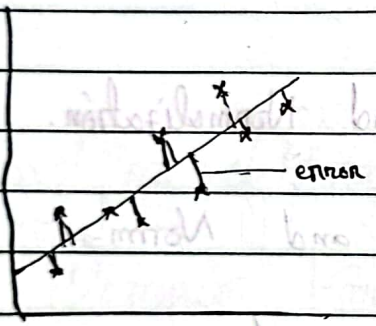
"R Squre" and "Adjusted R Square"

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

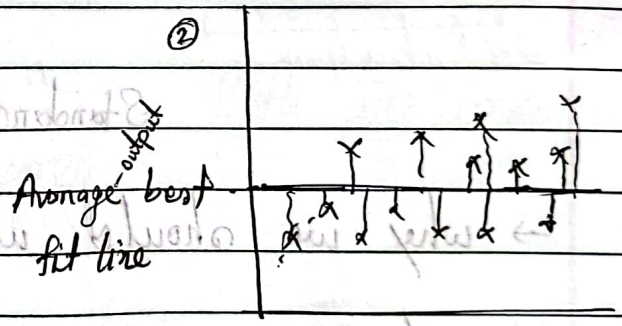
① SS_{res} = Residual or Error (Sum of square)

② SS_{tot} = Sum of Avarage total

①



②



$SS_{res} = \text{Sum} (y_i - \hat{y})^2$

$SS_{tot} = \text{Sum} (y_i - y_{i(avg)})^2$

$R^2 < 0$. when my best fit line worse than than the avarage best fit line.

$SS_{res} > SS_{tot} \Rightarrow > 1$ and get -ve.

R^2 → check goodness of best fit line.

⇒ Follow Krish Naik and Analytics vidhya article.

$$\text{Adjusted } R^2 = \left\{1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]\right\}$$

① added feature not correlated to the target value. $R^2 \downarrow$

when, $k \uparrow$

$(n-k-1) \downarrow$

$\left[\frac{\uparrow}{\uparrow}\right] \uparrow$

Adjusted R^2 decrease

② added feature correlated to the target variable.

when $R^2 \uparrow$ and $(n-k-1) \downarrow$

$(1-R^2)(n-1) \downarrow$

⇒ Adjusted R^2 Increase

Adj $R^2 =< R^2$

Analytics vidhya article ⇒

R^2 vs Adjusted R^2

① Difference ?

Hypothesis testing.

→ Evaluates 2 Mutual Exclusive statement,
↓ on
Population
↓
Using Simple data.

→ P values
→ T test
→ Anova test
→ Z - test
→ f - test

Steps

① Make Initial Assumption (H_0 - Null Hypothesis)
② Collecting data [Info. Evidence]
③ Gather Evidence to Re check.
Reject or Not Reject Null Hypothesis.

example
Criminal Case
H_0 Innocent
H_1 Guilty

or Confusion matrix

	H_0	H_1 → Truth
Do not Reject (accept) H_0	OK	Type II error
Reject	Type I Error	OK

Type I and Type II Error play major role.

Gender	Age	Weight	Height
M	Teen	20	1.4
F	Adult	65	1.2
M	Adult	65	1.4
M	child	20	1
F	Adult	25	1.3
M	Teen	80	1.3

$P \leq 0.05$ Significance value.

Two categoric (Gender, Age)

H_0 ---

H_1 ---

Test → Chi-Square test

One categorical

H_0 --- There is no difference

H_1 There is difference.

Proportion in Gender

Test → One Sample Proportion Test

$P \leq 0.05$ Reject H_0

Accept H_1

height mean test → 1.3

based on previous sample this mean is difference or not.

One Continous feature

H_0 ---

H_1 ---

Test → T-Test

Population mean vs Sample mean.

Two Continous feature

Test → Correlation.

Test → ~~T-test~~

Let's One Numerical or Continous feature :

One or two categorical feature

↓

many category.

more than two

Then we apply

Anova test

→ if only two category.

Suppose one numerical
one categorical
and in categorical variable
we have two category
we apply T-test

→ Chi-Square test

① Chi-Square Test for Independence

② Chi-Square test for Goodness of fit.

[Statlectures.]

↓

youtube.