Stats

Basic

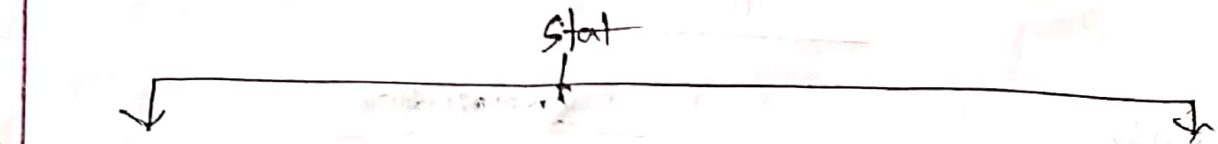
1. Variable
2. Rand "
3. Population, Sample
   " mean, " mean
4. Population distribution
   Sample "
   Sampling "
5. Mean, Median, Mode
6. Range
7. Measure of Dispersion
8. Variance
9. STD
10. Gaussian/Normal distribution,

Intermediate

12. Standard Normal distribution
13. Z score
14. Prob density function
15. Cumulative distribution function
16. Hyp testing
17. Graphs
18. Kernel-Density estimation
19. Central limit theorem
20. Skewness of data
21. Covariance
22. Pearson corr coeff
23. Spearman Rank corr
24. Hypo test

Advanced

25. Q-Q Plot
26. Pow law distribution
27. Box Cox transform
28. Poisson Distribution
29. Discrete and cont distribution
30. App of non Gauss Distrn
31. Bernoulli and Binomial "
32. Log Normal Distribution



* Inf-Stats: is a statistical method that deduce from a small but representative sample the characteristic of a biggen population.

ex. Asking people of diff election area and conclude which panity might win. By statistical testing

Ex of inferential stats that we do →

① Z-test → hypothesis testing
② T-test
③ Chi square test

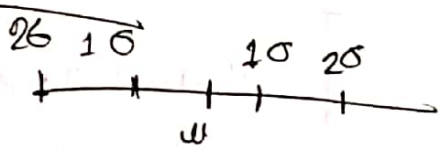
② Mean $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

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

$\sigma = \sqrt{\text{Var}}$

Variance and std helps to understand, from the mean, how far the values are distributed.

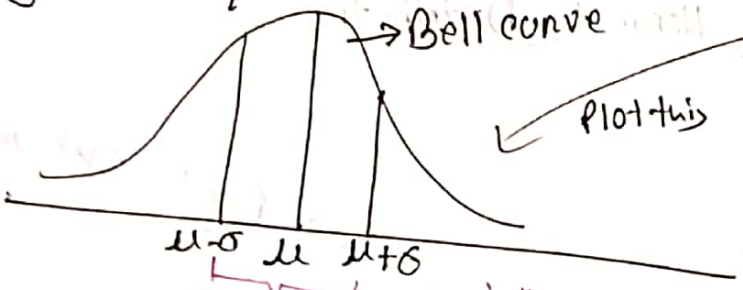
Gaussian / Normal distribution



Consider x as random variable

x belongs to Gaussian distribution with some mean and sigma value / standard deviation.

$x \approx GD(\mu, \sigma)$



Empenical formulas: → Probability that 68% of data points belonging to the rand var x fall within the range of first STD.

As it is Normal distribution

① $Pr[\mu - \sigma \leq x \leq \mu + \sigma] \approx 68\%$

② $Pr[\mu - 2\sigma \leq x \leq \mu + 2\sigma] \approx 96\%$

③ $Pr[\mu - 3\sigma \leq x \leq \mu + 3\sigma] \approx 99.7\%$

Use of them to remove outliers

Using std:

std = df.column.std

= 5

df[[(df.column ≤ 3*std) | (df.colum ≥ 3*-std)]]

Using Z score:

$Z\text{score} = \frac{x - \mu}{\sigma}$ (mean, std)

* Remove if Z score is greater than 3.

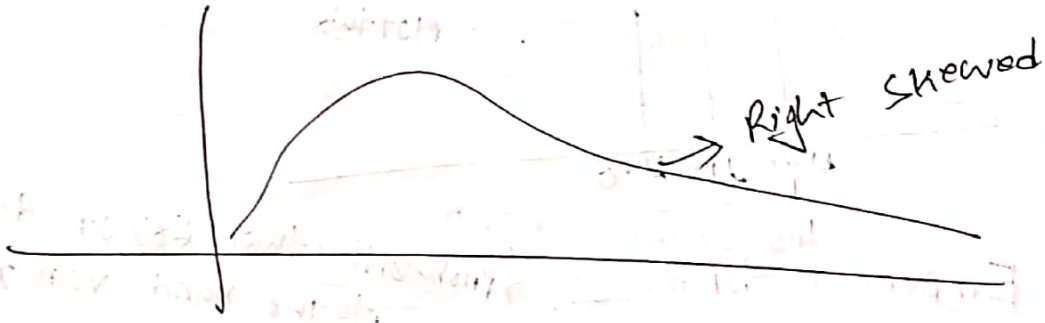
It is similar to std.

To remove:

df_no_outlier = df [(df.zscore(df) < 3) & (df.zscore(df) > -3)]

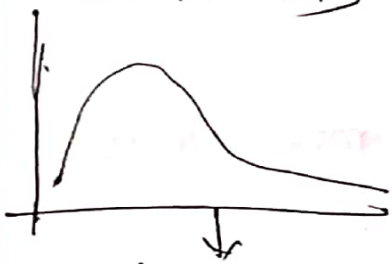
* We can also use percentile.

Log Normal Distribution

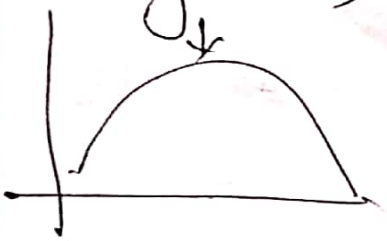


This is used to normalize the data.

for the above graph, if we apply log(x) to the x axis, it becomes normal distribution.



log(income)



If you get a normal distribution by applying a log function to a dataset the dataset is log normally distributed.

Gaussian → Standard Normal distribution

$Z = \frac{x - \mu}{\sigma}$ → Standard scaler

Normalization $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$

Std, $\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N}}$ ③

* Normalization or Standardization.

⇒ Normalization, when data doesn't follow gaussian distribution $\frac{x - x_{min}}{x_{max} - x_{min}}$

⇒ Standardization, when data follow gaussian distribution. $\frac{x - \mu}{\sigma}$ → It converts the dataset as mean → 0 and std = 1

Covariance → Direction of the relationship

* It maps relationship between two features.

$$Cov(size, price) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y - \mu_y)$$

(size → x, price → y)

It is similar to variance

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

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (x_i - \mu_x)$$

So, $Cov(x, x) = Var(x)$

It just says that there is positive or negative relationship but doesn't say about how much positive / Negetive.

important for feature selection

Pearson Connelation Coefficient

① Strength
② Direction of Relationship.

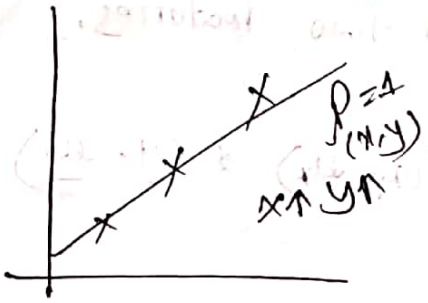
Range = **-1 to +1**

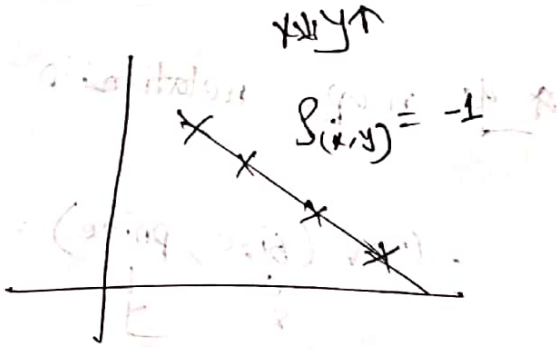
$-1 \leq \rho \leq 1$

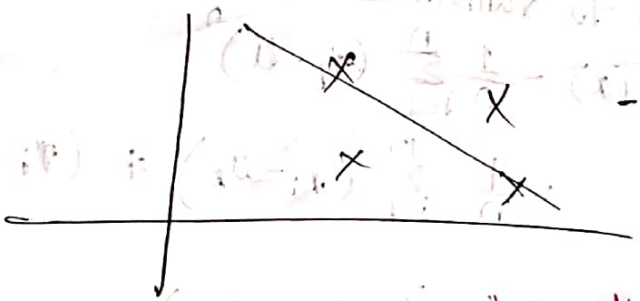
formula $\rho_{(x,y)} = \frac{cov(x,y)}{\sigma_x * \sigma_y}$ → covariance

$\frac{1}{n}\sum_{i=1}^{n}(x-\mu_x)*(y-\mu_y)$

Standard deviation. → $\sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$







$-1 \leq \rho_{(x,y)} \leq 0$

① Findout which variables are impacting most for ~~target~~ variable

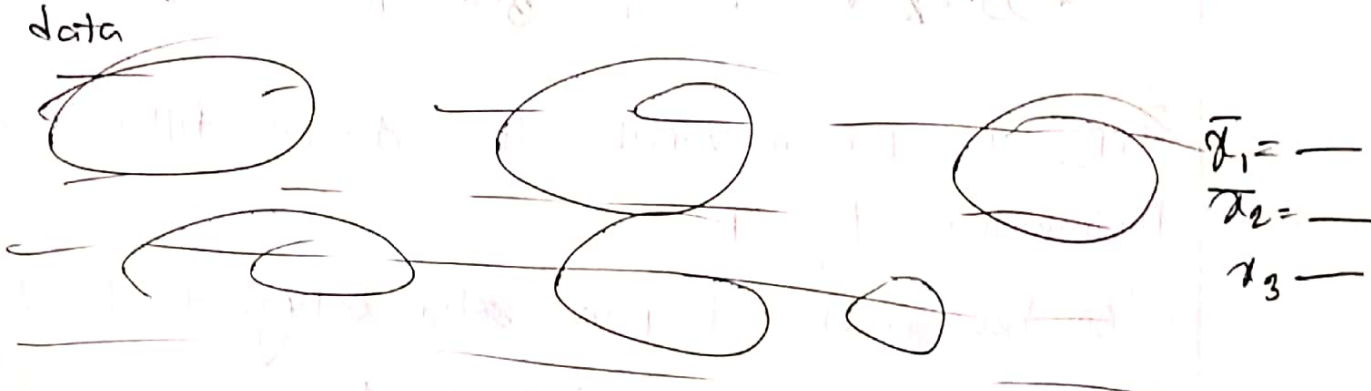
② find out multicolinianity and remove them.

without target variable → ~~two~~ different ~~value~~ features are highly ~~connelated~~

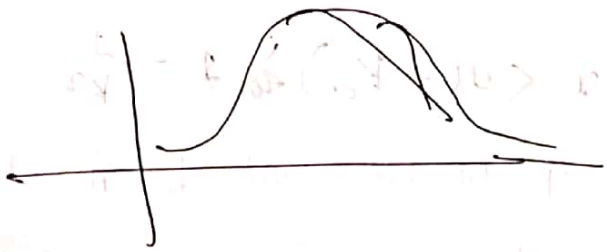
The presenve similar info, we can remove them

Central limit theorm

* No matter population distribution, if we take enough sample, calculate the mean of them, the distribution of the mean will be normal/gaussian.



If we plot $\bar{x}_1, \bar{x}_2, \bar{x}_3$, distribution will look like this:



And their mean will be closer to population mean. And standard deviation will be n time less

$$N \sim \left(\mu, \frac{\sigma^2}{n}\right)$$

CHEBYSHEV'S INEQUALITY

In Gaussion distribution:

* 68% data belongs to 1st Standard Deviation
* 95% " " " 2nd " "
* 99.7% " " " 3rd " "

Suppose y is random variable that doesn't follow Gaussian Distribution y ∉ GD.

So, how much data points belongs to 1st STD

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

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

k = which range of std, we want to find the value.

Suppose, k = 2 → for 2 STD

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

$$= 1-\frac{1}{4}$$

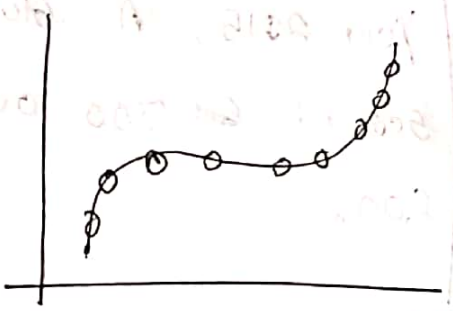
$$= \frac{3}{4} \approx 75\%$$

for k = 3 → for 3 STD

$$P_r > 1-\frac{1}{3^2} = 1-\frac{1}{9} = \frac{8}{9} \approx > 85\%.$$

→ better than pearson correlation
↓
As accurate on NON LINEAR DATA

Spearman's rank correlation Coefficient



Pearson correlation
$= \frac{cov(x,y)}{\sigma_x * \sigma_y}$

Spearman correlation = 1
Pearson correlation = 0.88

formula, $r_s = \rho_{rg_x, rg_y} = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x} * \sigma_{rg_y}}$

x	y
5	7
1	2
4	3
2	1

Sort →

x_{sorted}	y	$Rank_x$	$Rank_y$	d_i	d_i^2
1	2	1	2	−1	1
2	1	2	1	1	1
4	3	3	3	0	0
5	7	4	4	0	0

d_i = Rank x − Rank y

If all n ranks are distinct integers, then,

Spearman's, $r_s = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$

$= 1 - \frac{6 \times 2}{4(16-1)} = 1 - \frac{12}{60} = \frac{60-12}{60}$

$= \frac{48}{60}$

$= 0.8$ → highly correlated.

Quartile & Percentile

Year 2014, A student scored 450 out of 600	Year 2015, A student scored ~~60~~ 500 out of 600.

Who is better? Can't decide as year is different. The question might be higher.

Better comparatively — We need more info

82% students got less than 450.	73% students got less than 500.

Ex. The 63rd percentile of the mark in mathmatics in a class is 82.

↓

This means 63% students of that class got 82 on less

⇒ Higher the percentile, rarer your data.

* There are total 99 percentile.

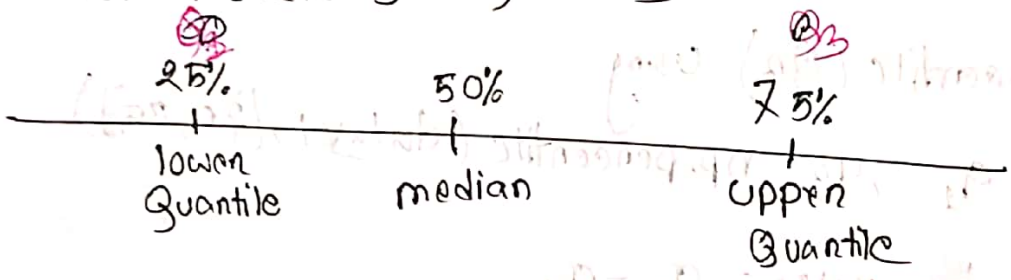
90th percentile is also called upper 10th percentile.

Quantile

Percentile is a version of quantile

Percentile considers 100 equal section to locate our data	Quantile can consider any number of equal sections

If 4 sections = Quantiles



Interquantile Range

Q_1 , Q_3

$IQR = Q_3 - Q_1$

Data 3, 5, 7, 1, 1, 8, 4, 6

↕ sorted

1, 1, 3, 4 | 5, 6, 7, 8

$Q_1 = 2$ $Q_3 = 6.5$

Interquantile Range = $Q_3 - Q_1 = 6.5 - 2$

$= 4.5$

Outlier Removal Detection

Two methods:

① Z-score

② Interquartile Range (75% − 25%) $Q_3 - Q_1$

Interquartile Range:

We can find lower percentile (Q_1) and upper percentile (Q_3) using

q_1, q_3 = np.percentile(dataset, [25, 75])

iqr_value = $q_3 - q_1$

lower_bound_value = quantile − (1.5 * iqr)

upper_bound_value = quantile + (1.5 * iqr)

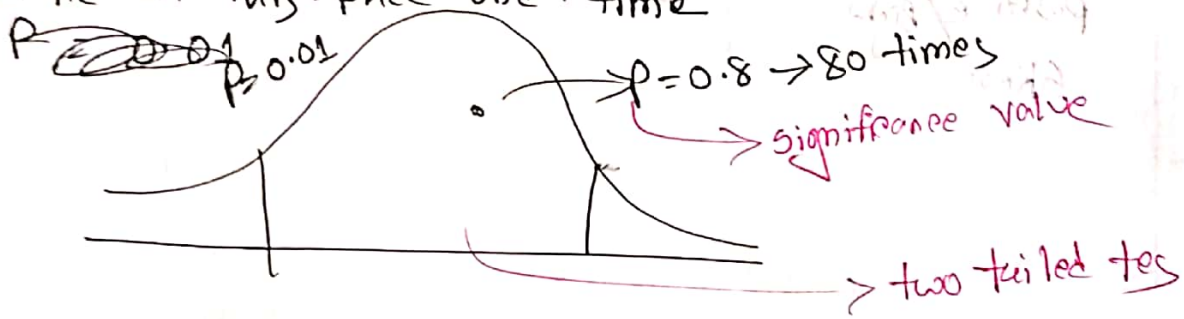
We can filter the values between lower and upperbound value and remove rest to remove outliers.

Z-score: Provides insight about how much standard deviation the data is away from the mean. We can remove data where z score is ~~−3 and +3~~ less than −3 and more than +3

P value → used to accept / reject null hypothesis

experiment on type on keyboard.

If we repeat this experiment 100 times and find people stroke on this place one time

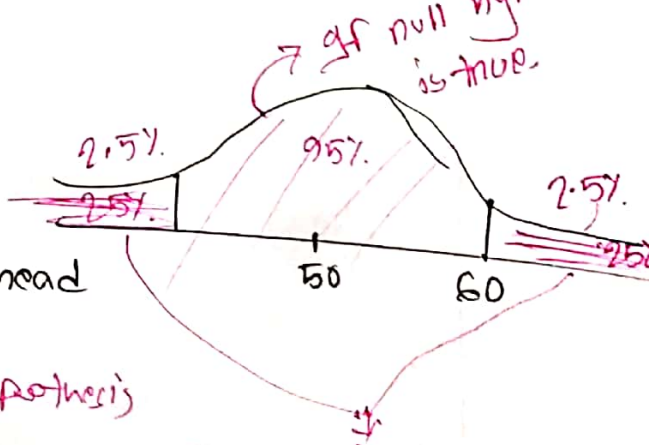


P-value: Probability that the null hypothesis is true

ex: Toss a coin 100 times → 50 times it is head.

H_0 = The coin is fair.

H_1 = The coin is not fair.



=> Suppose we get 60 times head

If $p \leq 0.05$, reject null hypothesis

↓

as it is away from the mean value.

If they fall in this region we'll reject hypothesis and p value will be less than 0.05

left-skewed, Right skewed, and their mean, median, mode

positive / Right Skew

gaussian distribution, standardized normal distribution, log normal distribution

Bernoulli distribution

Two outcome: ex: Tossing a coin. → H, T

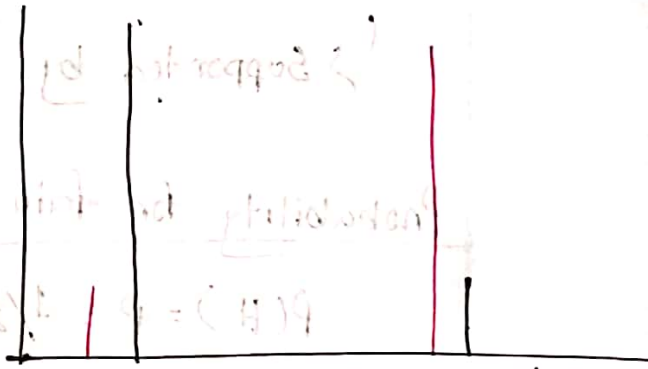
$X = 1$ → Head → Success

$X = 0$ → Tail → Fail

x is following bernoulli distribution

Prob of (X) will've some value x

$P(X=x) = p^x (1-p)^{1-x}$ → {Probability mass function}



Two examples of bernoulli distribution

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

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

Possible value for x is

$x = 0$ or $x = 1$

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

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

We know $P(X=x) = p^x (1-p)^{1-x}$

failure:

$\Rightarrow P(X=0) = p^0 (1-p)^{1-0}$

$= 1 - p$

$= q$

success:

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

$= p$

$$PMF = \begin{cases} q = 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

↳ Supported by $p^x (1-p)^{x-1} \Rightarrow PMF$

Probability for fair coin:

$$P(H) = p = 1/2$$

$$P(T) = 1-p = 1 - 1/2 = 1/2$$

unfair

$$P(H) = 0.4 = p$$

$$P(T) = 1 - 0.4 = 0.6$$

Mean, Variance and STD:

(expected value of x)

$$E(x) = \sum_{i=1}^{x} x \cdot P(x)$$

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

$$= 0.6$$

$x = 0$ / $x = 1$

$P(x=0) = 0.4 = p = 1-q$

$P(x=1) = 0.6$

formula for variance $= p(1-p) = pq$

" " STD σ, SD $= \sqrt{pq}$

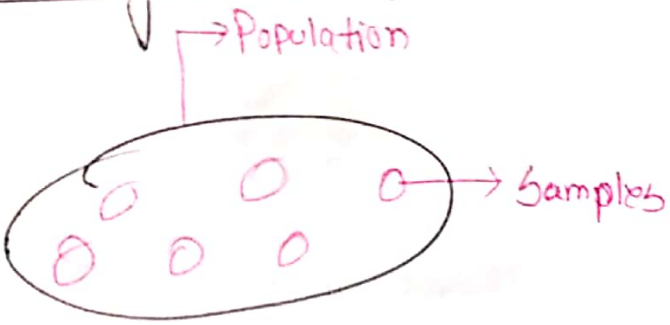
Mean	Variance	SD
p	pq	$\sqrt{pq}$

*If ratio is proportional, we can go toward random sampling better.

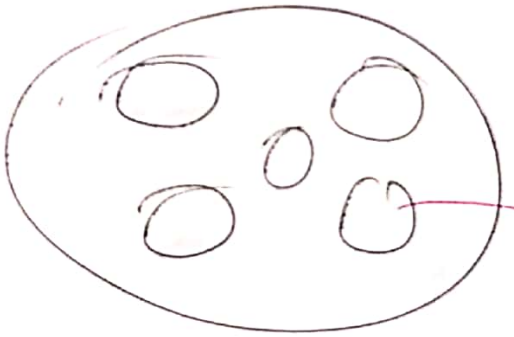
* For unbalanced class stratified sampling technique works

Sampling Techniques

① Random Sampling Technique:



② Stratified Sampling technique:



7:3 ratio caused biased.

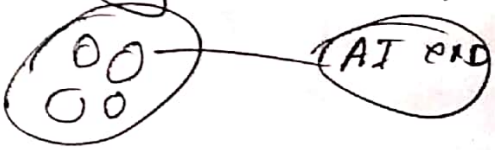
We've to provide 1:1 for good sampling.

③ Systematic sampling:

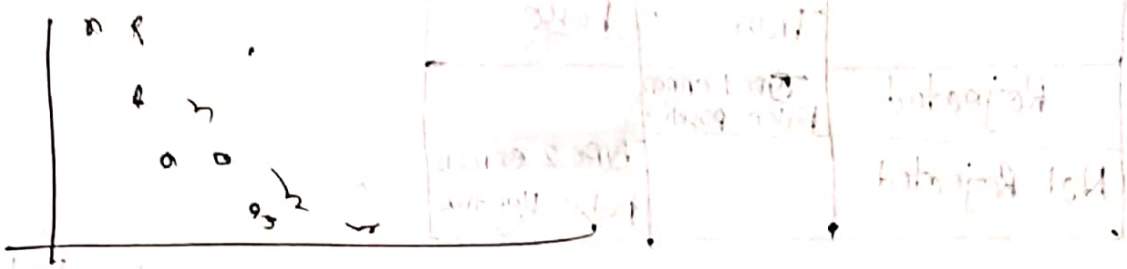
④ Cluster Sampling:

Ex: If AI will take our job.

We will consider only domain expert



L1 and L2 Regularization



math won = $\theta_0 + \theta_1 * age$

math won = $\theta_0 + \theta_1 * age + \theta_2 * age^2 + \theta_3 * age^3 + \theta_4 age^4$

These regularization reduce overfitting by penalize higher degree of polynomial.

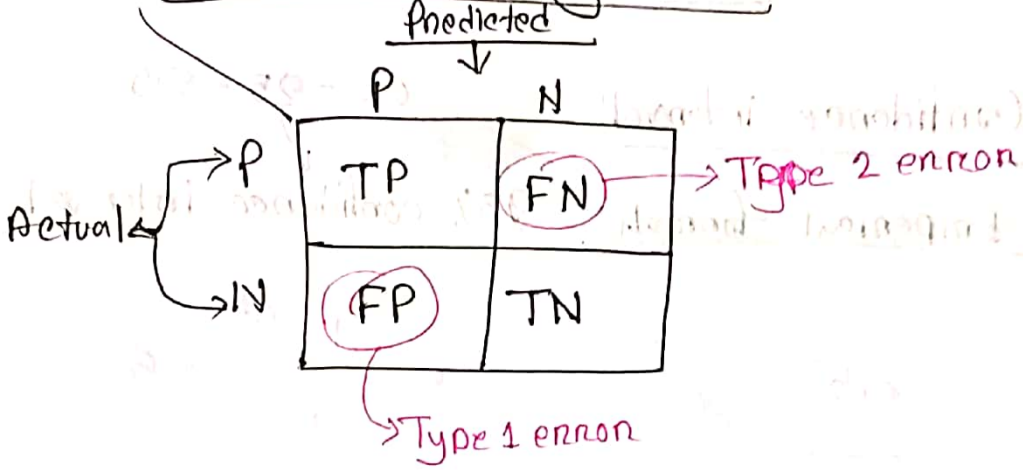
formula for L2 Regularization $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2 + \lambda\sum_{i=1}^{n}\theta_i^2$

Act like a knob. The bigger, theta value will be smaller.

n λ L1 n $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2 + \lambda\sum_{i=1}^{n}|\theta_i|$

Type 1 and Type 2 error

false positive / Negetive



3 cases to consider false positive / Negetive:

① Whether a person has disease or not

② Market will crash

③ Vaccination side effect.

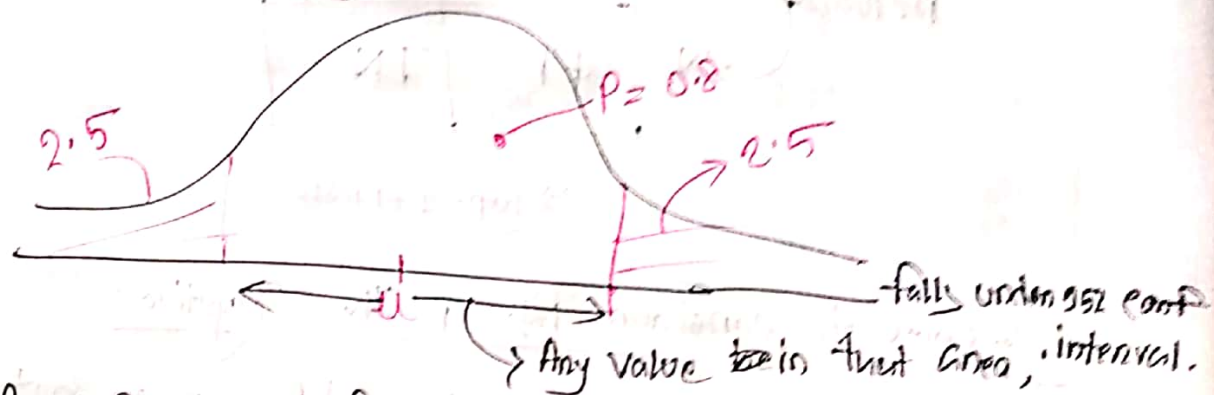
Pvalue also known as significance value. denoted by $\alpha = 0.05$

Hypothesis testing - Confidence interval

Z-test Statistics

Confidence interval: 68 - 95 - 99

Empenical formula: 95% confidence interval

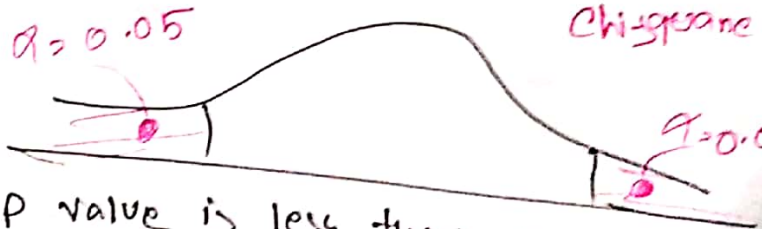


p = 0.8 means for 100 experiment, that specific thing can happen 80 times.

Hypothesis testing

① Null hypothesis → Alternative hypothesis

② Perform experiments → $\alpha = 0.05$ { t-test, z-test, Chi-square, Annova }



If p value is less than α, then possibility to happen that will be rare.

So, we reject the null hypothesis.

for $\alpha = 0.05$ → confidence interval is 95 percent
for $\alpha = 0.01$ → " " 99 percent

(3) Test value falls in extreme end → Reject the null hypothesis.

Accept the alternative hypothesis.

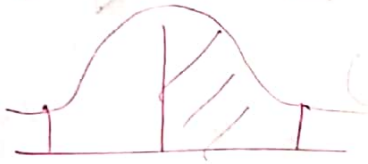
Test value falls in confidence interval → Accept null hyp
↳ Reject Attennative

Significance value ↑ confidence interval ↑
If $\alpha = 0.05$ then CI is 95%.

Experiment

1 tail test

$\bar{x} > \mu$



We just have to consider right part of the mean (μ)

2 tail test

$\bar{x} \leq \mu$

$\bar{x} \geq \mu$

We have to consider left and right part of μ

As it can be greater or less than equal to μ (mean value)

Case study

Task: Avg height of people in India.

hypothesis testing
- population → with some info
- Sample → perform some experiments. (→ statistical testing)
  - → conclusion ~~about~~ about the population

Statistical Test:

① T-test (Comparison of mean) → Sample (s) σ (Standard deviation) given

✓ ② Z-test (Comparison of mean) → σ of population given

③ Annova test (Analysis of variance)

④ F-test (Comparison of variance)

⑤ Chi-squine (~~Cate~~ " " Categorical variable)

not good

One sample Z test

Next page

IQ test for students:

Average, $\mu = 100$

Standard deviation, $\sigma = 15$

Doctor tested new medication to findout if it increases or decreased the IQ.

Medication
- > IQ
- < IQ

After 1 month ($n = 30$ 30 people) were taken.

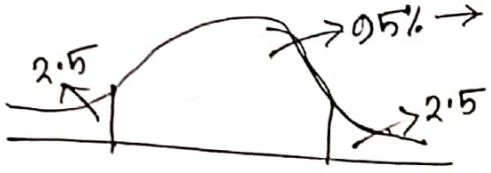
Whose mean, $\bar{x} = 140$

Did medication affect intelligence. [Significance value; $\alpha = 0.05$]

① Null hypothesis, $H_0 \Rightarrow \mu = 100$ → didn't affect.

Alternative " , $H_1 \Rightarrow \mu \neq 100$ → affect

So, this is a two tail test

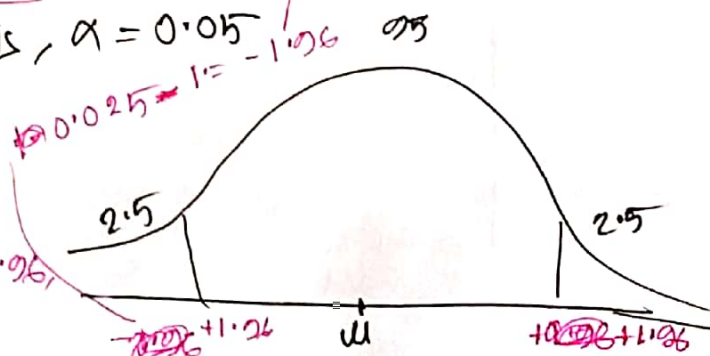


$1 - \alpha$
$1 - 0.025$
$= 0.9750$

We need to find area under the curve of 0.9750

② State the Alpha: Alpha is, $\alpha = 0.05$

③ State the decision rule → find specific region



If they range between -1.96 to +1.96, we accept the null hypothesis.

Otherwise if ≤ -1.96 or ≥ +1.96, we reject the null hypothesis and accept the alternative hypothesis.

Z-test: What is the specific range in confidence interval.

different from z-score → $\frac{x-\mu}{\sigma}$

④ z-test:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}}$$

(sample mean, population mean)

$$= 14.60$$

⑤ State Result:

Range to accept null hypothesis is -1.96 to +1.96

But z ~~score~~ stat is 14.60

As it doesn't fall into the range

So, ⇒ Reject the null hypothesis and accept the alternative hypothesis.

Scanned with CamScanner

AOC
ROC

(M)

One sample Z test ~~with proportion~~

In a population, the average IQ, μ = 100 with σ = 15. (mean, std)

Doctor tested a new medication to find out whether it increase the iq or decreases.

Medication
> IQ
< IQ

After 1 month

Sample of 30 (n=30) participants were taken. Sample mean for 30 peoples IQ is 140 ($\bar{x}$=140).

⇒ Did medication affect intelligence? Significance value is 0.05 (α=0.05)

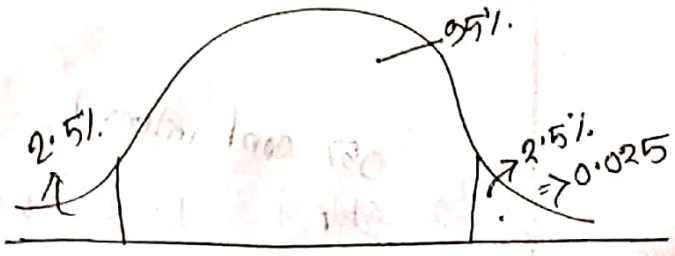
① ⇒ Hypothesis:

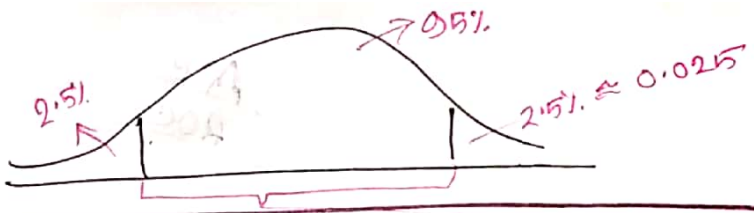
H_0 ⇒ Did not affect the IQ (μ=100)

H_1 ⇒ Affect the IQ ($\mu \neq 100$)

IQ can be increased or decreased. That's why it is a two tailed test.

② ⇒ State the alpha value: (α= 0.05)





③ ⇒ State the decision rule: ← this range

Entire region for the above distribution is = 1

finding the range

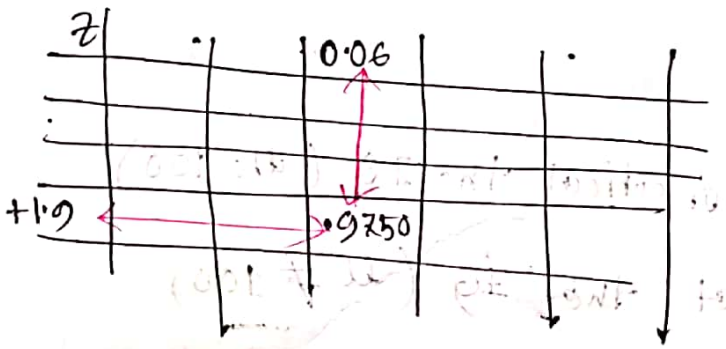
①⇒

$1 - 2.5\% \Rightarrow 1 - 0.025 = 0.9750$

Z-score: how many std away from the mean.

Z-test: find the range for 95% confidence interval.

②⇒ Now we've to find out the area under the curve for 0.9750. We can find it on "Z-table"

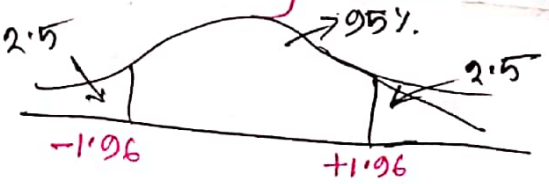


So, the area under the curve is: $1.9 + 0.06$

$= 1.96$

95% conf interval

⇒ So, ~~Std~~ is 1.96 away from the mean.



different from z-score

④ => **z-test:**

formula $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$

⑤ => **State the result:** We'll accept the null hypothesis if z-test value fall into the area under the curve range. (-1.96 to +1.96)

14.60 doesn't fall into -1.96 to +1.96

=> Reject the null hypothesis.

=> Accept " Alternative "

} Changes in IQ

(*) One sample z test with proportion:

A survey claim that 9 out of 10 doctors recommed aspirin for their patient with headache. To test this claim, a random sample of 100 doctors is taken. Out of this 100 doctors, 82 indicate that they recommend Aspirin. $\alpha = 0.05$

(1) ⇒ Hypothesis:

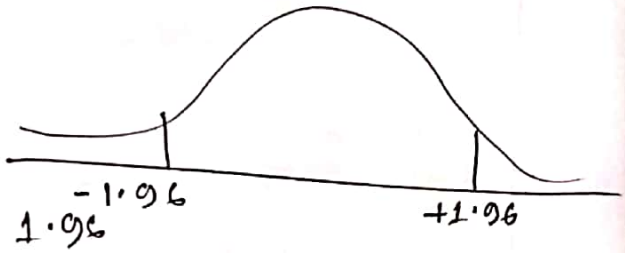
Null Hypothesis, $H_0 = P = 0.90$

Alternative " " $H_1 = P \neq 0.90$

(2) ⇒ Significance: $\alpha = 0.05$

Range: -1.96 to +1.96

As $1 - 0.05 = 0.0750 \rightarrow$ z table $\rightarrow 1.96$



(3) ⇒ Decision Rule:

(4) ⇒ Z-statistics:

formula, $Z_0 = \dfrac{\hat{P} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$

Here, $\hat{P} = 0.82$ ⇒ Probability respect to sample ⇒ $\frac{82}{100}$

$P_0 = 0.90$ ⇒ Probability respect to population ⇒ $\frac{9}{10}$

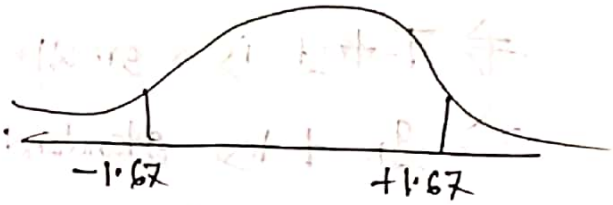
$n = 100$ ⇒ num of samples

$$= \frac{0.82 - 0.90}{\sqrt{\dfrac{0.90(0.10)}{100}}} = -2.66$$

(h) State the result:

Z value is = -2.667

Our range is :- -1.67 to +1.67
↓
from z table



As z value is out of the range, we reject the null hypothesis.

T-test

⇒ T-test is a small sample test.

⇒ In t-test standard deviation (σ) is given for sample, not population.

⇒ Formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{\sigma \rightarrow \text{sample}}{\sqrt{n}}}$$

Application of t-test

~~⇒ Ap~~

(*) Size of sample is small ($n < 30$)

(*) Degree of freedom is ($v = n - 1$) → sample size

⇒ T-test is used for test of significance of regression coefficient in regression model.

Statistical testing

Panametric test

Non-panametric test

① Panametric test: Suitable for normally distributed data.

② Non-Panametric test: Suitable for any continious data, based on ranks of the data values.

=> Tree types of t-test:

① One sample t-test
② Unpaired two sample t-test
③ Paired sample t-test.

=> Conditions for t-test:

① The data has to follow a continious on ordinal scale.
② The data has to be randomly selected.
③ The data should be normally distributed.

=> t-score:

The t-score formula enable us to transform a distribution into a standandize form, which we use to compare the score

formula, $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$

s_1^2, s_2^2 = ~~Variance~~ Std

N_1, N_2 = Num of samples

$\bar{x}_1, \bar{x}_2$ = mean of two groups

Group	method	n	$\bar{x}$	s
1	Intensive	12	46.31	6.4
2	Paced	10	42.79	7.52

$$t = \frac{46.31 - 42.79}{\sqrt{\frac{6.44^2}{12} + \frac{7.52^2}{10}}}$$

$$= \frac{3.52}{\sqrt{3.46 + 5.66}} = 1.166$$

t value is = 1.166. Now to find valid range, we need to look into t-table with significance value and degree of freedom.

degree of freedom is smallest among 12-1, 10-1

∴ Degree of freedom = 9

$\alpha = 0.05$

⇒ With df = 9 and $\alpha = 0.05$, from t-table, we find that critical value is = 1.833

As t-value 1.66 doesn't exceed 1.833, we accept the null hypothesis.

→ No std for population
→ Std for sample.

One Sample t-test

Z-tests are commonly done when population std is known

t-tests " " " " " " is unknown

* We use t-distribution as population std is unknown. that we have to estimate.

⇒ Example Avg IQ 100. Scientist experimend that medication has positive/negetive effect.

To test affect of the medication, 30 participants were taken with mean of 140 and std of 20. & Alpha = 0.05

① ⇒ Define Hypothesis:

$H_0 \Rightarrow \mu = 100$

$H_1 \Rightarrow \mu \neq 100$

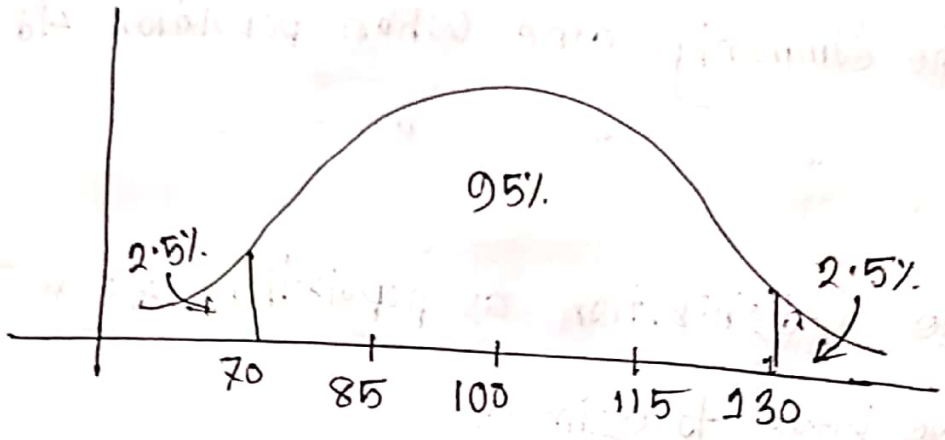
② ⇒ State alpha:

$\alpha = 0.05$

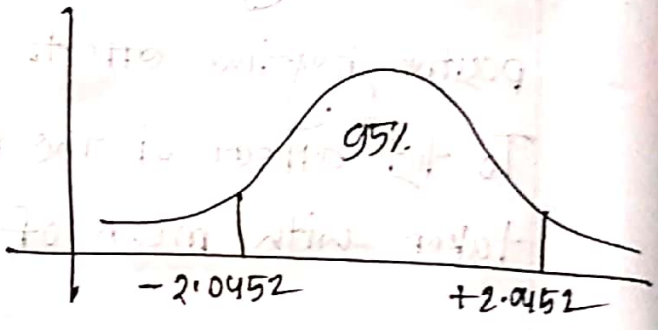
③ ⇒ Calculate Degree of freedom:

$n-1 = 30-1 = 29$; n = num of sample.

④ => State Decision Rule:



df		Critical value
2tailed test	0.1	0.5
29		2.0452



=> Critical value is => -2.0452 to $+2.0452$

if t is greater/ledger than critical value, reject the null hypothesis.

⑤ => Calculate test statistics:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\bar{x} = 140$
$\mu = 100$
$s = 20$
$n = 30$

$$= \frac{140 - 100}{\frac{20}{\sqrt{30}}}$$

$$= 10.96$$

6⇒ State result:

Critical value is: -2.0452 — 2.0452

$t = 10.96$

As t value is far away from critical value, reject the null hypothesis.

∴ It changed the IQ.

Independent sample t-test

⇒ Compare two independent sample

⇒ Ex: comp two class. Class A: 25 student, $\mu_1 = 70$, $s_1 = 15$

Class B: 20 student, $\mu_2 = 74$, $s_2 = 25$. $\alpha = 0.05$. Are they different

$H_0 \Rightarrow \mu_1 = \mu_2$

$H_1 \Rightarrow \mu_1 \neq \mu_2$

⇒ $\alpha = 0.05$

⇒ Degree of freed: $df = (n_1 - 1) + (n_2 - 1) = (25-1) + (20-1) = 43$

2⇒ decision rule: critical page from t-table

→ -2.0167 ⟷ 2.0167

(b) => test stat:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_p^2}{n_2}}} \qquad s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

$$= \frac{70 - 74}{\sqrt{\frac{401.74}{25} + \frac{401.74}{20}}} \qquad = \frac{5400 + 11875}{24 + 19}$$

$$= 401.74$$

$$= \frac{-4}{\sqrt{36.16}}$$

$$= -0.67$$

$df_1 = n - 1 = 25 - 1 = 24$

$df_2 = n - 1 = 20 - 1 = 19$

$SS_1 = s_1^2 (df_1) = 15^2 (24) = 5400$

$SS_2 = (25^2)(10) = 11875$

$s^2 (df)$

(c) => State result:

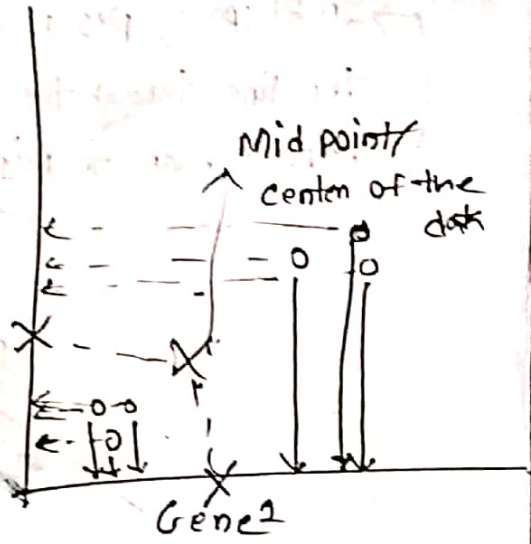
Decision Rule: Critical range $-2.0167 \longleftrightarrow 2.0167$

$t = -0.67$

Accept the null hypothesis

PCA

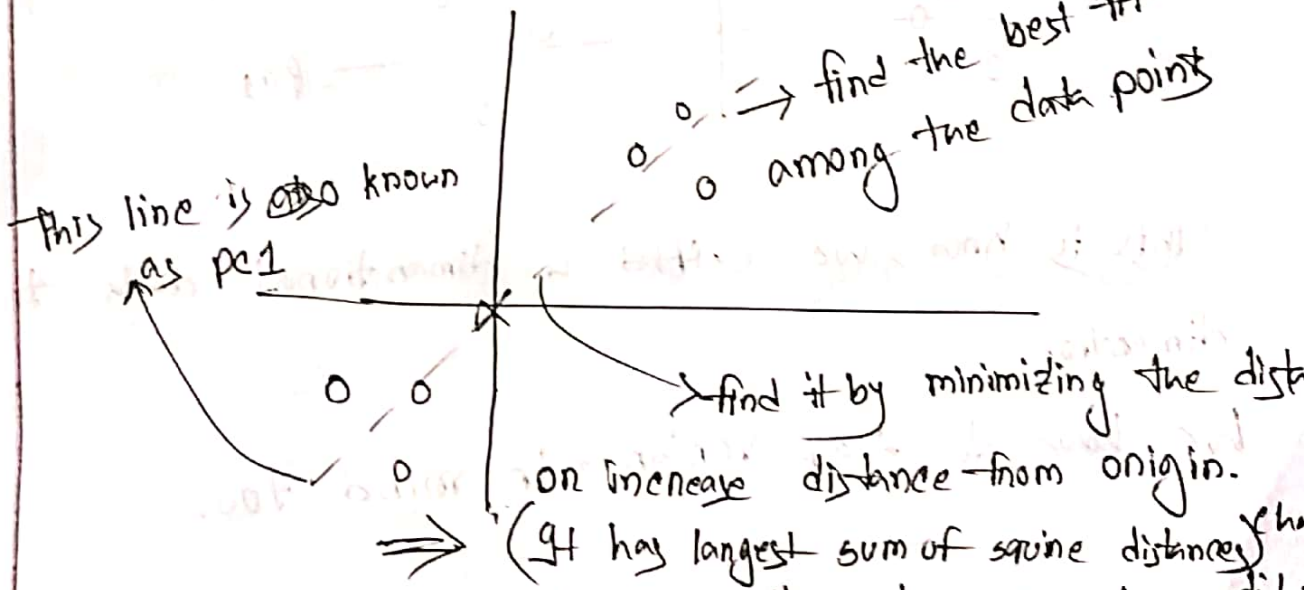
	Mouse1	M2	M3	M4	M5	M6
Gene1	10	11	8	3	2	1
Gene2	6	4	5	3	2.8	1



① Calculate average for Gene1

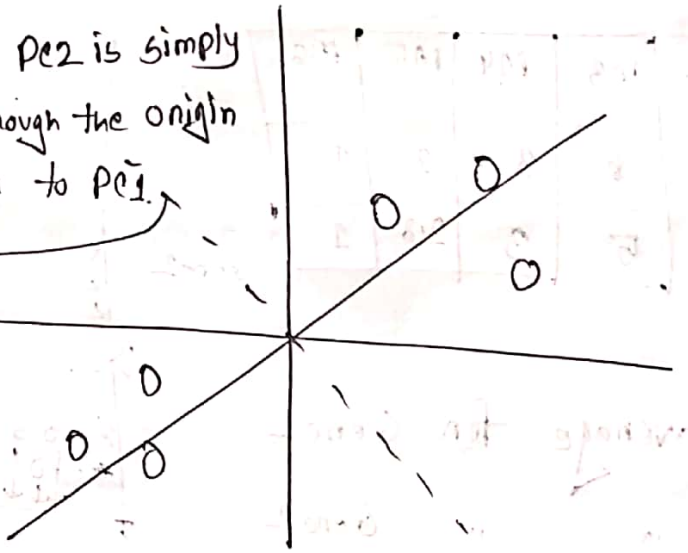
② " " " Gene2

* Shift the data, so that the center is on top of the origin (0,0) in the graph.

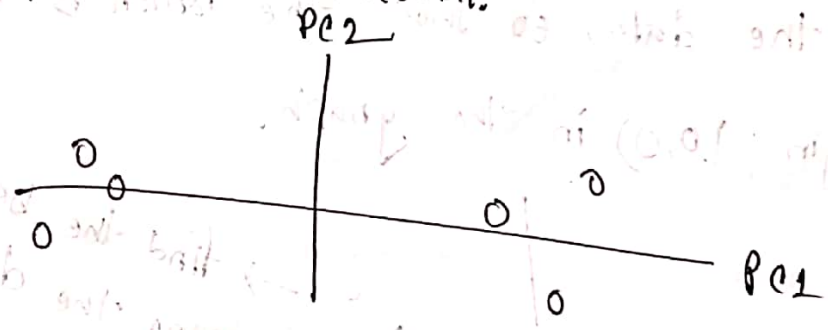


Note: Data points are positioned relative to each other. did not change

In 2D space, pc2 is simply the line through the origin perpendicular to pc1.



finally shift them to horizontal.

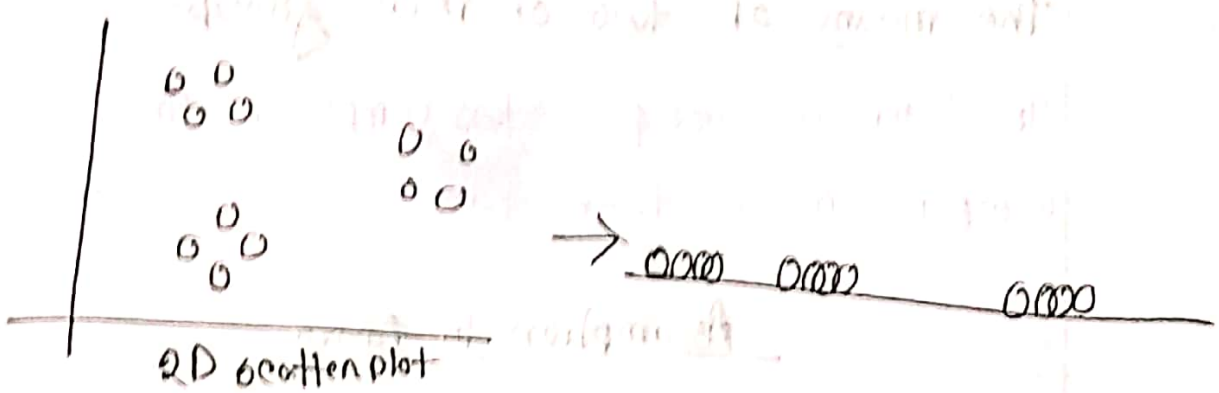


This is how, we shifted 6 dimentional data to two dimention.

We have to keep vanience in mind too.

t-SNE

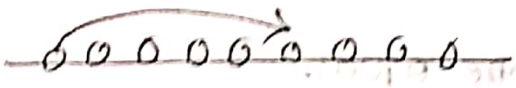
=> It takes high dimentional data. Reduce it to low dimentional graph that retain a lot of orginal information.



Let's see how t-SNE transform this graph into a flat, 1-D plot on a number line.

=> t-SNE finds a way to project data into low dimentional space, so that the clustering in the high dimentional space is preserved.

Step1: put the points on the number plot in a random order.



t-SNE move them a bit by bit until it cluster them

=> It attract by similar points and pushed by not similar once.

It moves according to the attraction / ditraction

Anova test

Anova ⇒ Analysis of vanience

Anova is a statistical method used to compare the means of two on more groups.

In t-test, we compare two sample, in Anova, we can compare more than two.

Assumption in Anova

1. Normality of sampling distribution of mean.
2. Independence of error
3. Absence of outliers.
4. Homogenity of vanience

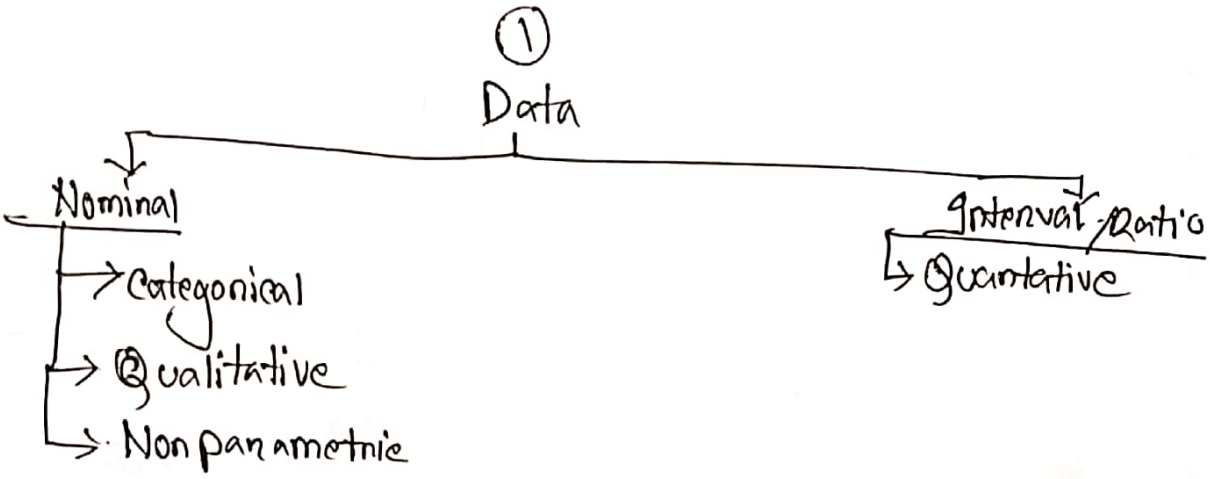
Hypothesis

$H_0 \Rightarrow \mu_{A_1} = \mu_{A_2} = \mu_{A_3}$

$H_1 \Rightarrow$ Not all means are equal.

Choosing the test

1. Data
2. Sample
3. Purpose.



Test statistics for nominal data:

1. Test for proportion
2. Difference of two proportion
3. Chi-squine test for independence

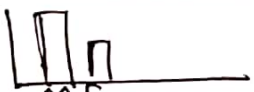
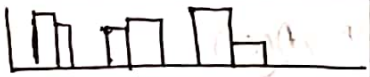
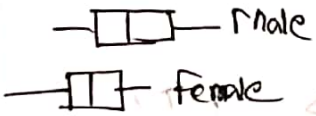
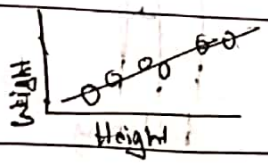
Test statistics for interval data:

1. Test for a mean
2. Difference of two means (paired)
3. Difference of two means (independent samples)
4. Regression Analysis.

What to use, where

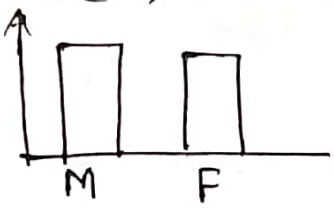
t-test, Chi-squine test, p vale

Gendan	Age	Hoight	Weight
Female	Adult	1.4	60
Male	Child	1.2	15
Male	Adult	1.5	85
Female	"	1.3	74
Male	"	1.6	77
Female	Eldenly	1.5	65

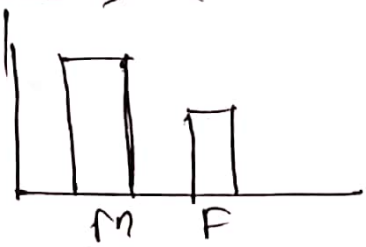
	What we observe in our sample data	Is it real
One categonical		1 sample Proportion test
Two "		Chi squined
One numenical		t-test
One numenical & One categonical		t-test on Annova
two numenica		connelation test

(1) => One Category: (Genden)

H_0 => There is no difference in genden propontion



H_1 => There is a difference:

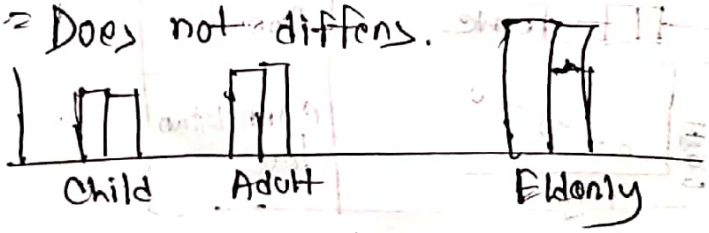


Test: We need to know how significance this difference is by One sample propontion test.

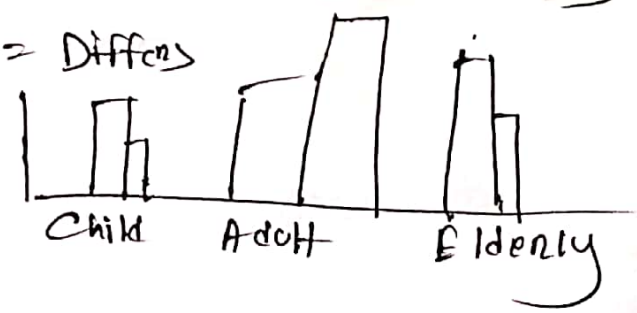
(2) => Two Categories: (Genden, Age)

Does the propontion of male and female, differ accross age groups?

H_0 = Does not differs.



H_1 = Differs



=> test: chi-square → calculate p value

if p is less than α, reject null hypothesis

(2) One numeric (height)

Is height different from previous year?

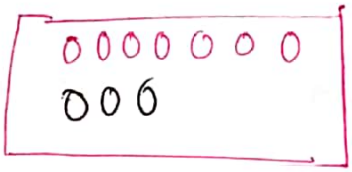
Bayes theorm

① Conditional probability

② Independent event → coin toss → doesn't depend on each other

③ Dependent event → one event affecting the probability of another event.

$P(\text{Black}) = \frac{3}{10}$



$P(\text{Pink}) = \frac{7}{10}$

dependent event

* If we randomly select two marbel from this box, what is the probability of drawing a ~~green~~ pink marbel and then a ~~blue~~ black marbel, without replacement.

First event

$P(\text{pink marbel}) = \frac{7}{10}$

$= 0.7$

Second event:

$P(\text{Black marbel}) = \frac{3}{9}$ → as we drawn a marbel

$= 0.33$

$P(A \cap B) = P(A) \times P(B)$

$= \frac{7}{10} \times \frac{3}{9_3}$

$= \frac{7}{30}$

$= 0.233$

Conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{n(A \cap B)}{n(B)}$$

* ~~Probability of~~ Given that probability of picking a pink manbel is $\frac{7}{10}$. What is the probability of picking black manvel ~~given~~ for given probability of pink manbol.

	পুরুষ	নারী	মোট
বিবাহিত	50	40	90
অবিবাহিত	40	60	100
মোট	90	100	

① পুরুষ যদি বিবাহিত হওয়ার সম্ভাবনা কত?

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{50}{90} = \frac{5}{9}$$

→ মোট পুরুষ

Bayes theorm

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$\Rightarrow P(A \cap B) = P(B \cap A)$$

$$P(A \cap B) = P(A|B) * P(B)$$

$$\updownarrow \text{ equal.}$$

$$P(B \cap A) = P(B|A) * P(A)$$

As $P(A \cap B) = P(B \cap A)$

$$= P(A|B) * P(B) = P(B|A) * P(A)$$

$$\therefore P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

- $P(B|A)$ → likelihood probability
- $P(A)$ → Prior
- $P(A|B)$ → Posterior Probability
- $P(B)$ → marginal

Naive Bayes classifier

Bayes theorm, $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

Dataset

$\{x_1, x_2, x_3, x_4, \cdots x_n\}$ $\{y\}$.

$$P(y|x_1, x_2 \cdots x_n) = \frac{P(x_1, y) P(x_2|y) P(x_3|y) \cdots P(x_n|y) * P(y)}{P(x_1) P(x_2) \cdots P(x_n)}$$

$$= \frac{P_y * \prod_{i=1}^{n} P(x_i|y)}{P(x_1) P(x_2) \cdots P(x_n)}$$

$$\therefore P(y|x_1, x_2 \cdots x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

$$y = \text{argmax}_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

→ Weill accept the highest value for that classifier

Suppose for y happy → 07

Yes = 0.7

NO = 0.3

=> Weill accept 0.7 ⇒ Yes

Outlook

	Yes	No	P(Y)	P(No)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(Y)	P(N)
HOT	2	2	2/9	2/5
MILD	4	2	4/9	2/5
COOL	3	1	3/9	1/5
Total	9	5	100%	100%

⇒ Suppose today is Sunny and Hot. What is the probability of playing?

Today (Sunny, Hot)

	Play	P(yes)
Yes	9	9/14
No	5	5/14
Total	14	

$$P(Y \mid Sunny, Hot) = \frac{P(Sunny \mid yes) * P(Hot \mid yes) * P(yes)}{P(Today) P(Sunny)} \rightarrow \text{ignored}$$

$$= \frac{2}{9} * \frac{2}{9} * \frac{9}{14}$$

$$= 0.031$$

$$P(N \mid Sunny, Hot) = P(Sunny \mid No) * P(Hot \mid No) * P(No)$$

$$= \frac{3}{5} * \frac{2}{5} * \frac{5}{14}$$

$$= 0.08571$$

Now, we have to normalize them to 1

$$P(yes) = \frac{0.031}{0.031 + 0.08571} \qquad P(No) = 1 - 0.27$$

$$= 0.27 \qquad = 0.73$$

So, output is No.

2 sample z-test for the difference between Means

A guidance counsilor claims that the student at college prep program has higher scon than general program.

45 student randomly selected from college prep with mean score of 24.1 and population std of 4.6.

42 stdent randomly selected from general program with mean score of ~~29.01~~ 20.2 population and std of 4.6.

With 5% significance, can you suppon the claim

(1) ⇒ **Hypothesis:**

$H_0 \Rightarrow \mu_1 = \mu_2$

$H_1 \Rightarrow \mu_1 > \mu_2$

μ_1 = std in college prep

μ_2 = " " general program

(2) **Z-test and pvalue:**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}}$$

according to null hypothesis diff is 0

$$= \frac{(24.1 - 20.2) - (0)}{\sqrt{\frac{4.6^2}{45} + \frac{5^2}{42}}}$$

$$= 3.695$$

pvalue is = $1E^{-4} = 0.0001$

③ ⇒ Decision Rule:

Compare p value to α.

pvalue is :- $0.0001 < 0.05$

Reject the null hypothesis.

Scanned with CamScanner

Z-test for proportion, two samples

64 out of 200 people taking medication report symptoms on anxiety.

Of the people receiving placebo, 92 out of 200 report symptoms of anxiety. Is medication working differently than the placebo. $\alpha = 0.05$

① Hypothesis:

$H_0 \Rightarrow P_1 = P_2$

$H_1 \Rightarrow P_1 \neq P_2$

② State Alpha:

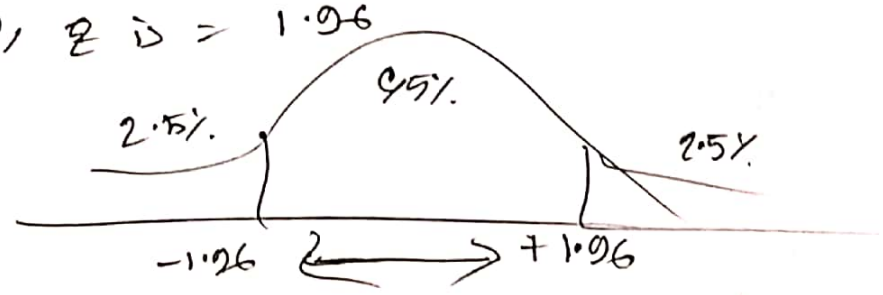
$\alpha = 0.05$

③ State the decision:

~~$1 - \alpha = 1 - 0.05 = 0.9750$~~

$1 - \frac{\alpha}{2} = 1 - 0.025 = .975$

So, Z is = 1.96



④ Calculate the test stat

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$n_1 = 206$

$n_2 = 206$

$P_1 = \frac{64}{200} = 0.32$

$P_2 = \frac{92}{200} = 0.46$

$$= \frac{0.32 - 0.46}{\sqrt{0.39(1-0.39)}\sqrt{\frac{1}{200}+\frac{1}{200}}}$$

$$= \frac{-0.14}{(.488)(0.1)}$$

$$= 2.869$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + x_2}$$

$$= \frac{64 + 92}{200 + 200}$$

$$= 0.39$$

(b) State the result:

if z is less than -1.96 or greater than 1.96, reject the null hypothesis

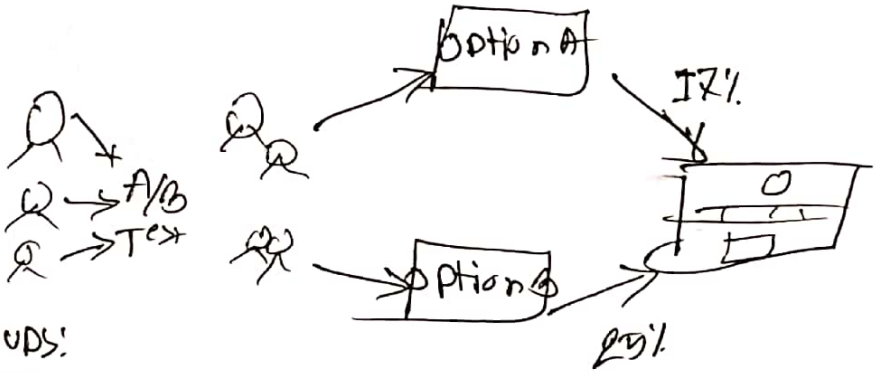
$$z = 2.869$$

=> Reject the null hypothesis.

A/B Testing

* We want to increase the sell of the product. Divide the products to A and B. A is unchanged an B has significant changes in packaging.

On the basis of the response from customer groups who used A and B respectively, which one performs better?



Two groups:

① Control
↓
Convention rate
16%

② Experiment/Test
↓
Convention rate
19%

⇒ Can we conclude that experimet group is doing better?

No, → for that we have to find that if the difference is statistically significant:

through two sample z test, proportion test

Type 1 error: Reject null hypothesis while it is true. Accept B while it is not performing better than A.

Type 2 error: Accept the null hyp while it is false. Concluding B is not doing good while it is doing good.