P-hacking:

> Exploit the statistical analysis.

> If it is false they reject null hypothesis.

P values:

evaluate the data the null hypothesis is true

> Manipulating data or analyses to artificially get significant p values.

In null hypothesis testing it is either Reject or fail

The binary decision process leads us to 4 possible scenarios.

	H_0 is True	H_0 is false
Fail to Reject the H_0	✓	X
Reject the H_0	X	✓

p hacking is analyses are being chosen based on the p values significant, not the best analysis Plan.

Eg:-

some people affect with virus we need to develop a drug in this A, B, C, D ...

check with people by this drug & measure

at last one drug does great job so

we calculate the means of the two groups.

then we get p value = 0.02

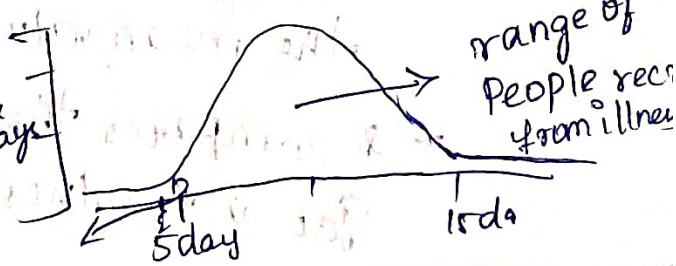
$0.02 < 0.05$ we reject the null hypothesis there is no difference b/w not taking a drug & taking drug Z.

Use p. hacking

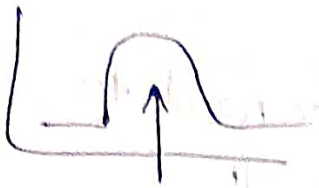
↓

refer to misuse & abuse of analysis techniques & results in being false positives.

we measured: recovery times for a whole lot of people, did not take any drugs to fight virus.

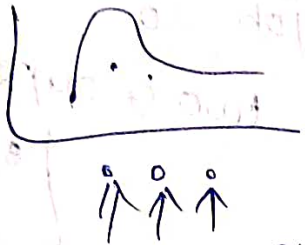


Indicating that 95% of people recovered in less than 5 days.

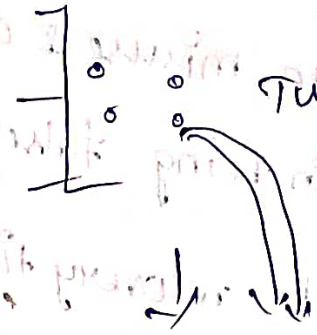


95% of area under the curve is blw 5 to 15 days.

Indicating that 95% of people recovered blw 5 to 15 days



3 People recover from illness there's good chance but say something blw 5 to 15d



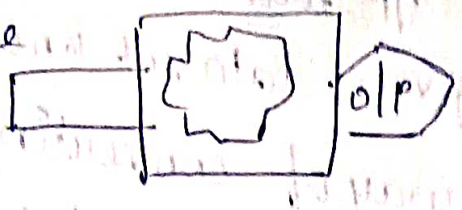
Two group of people.

→ Then we calculate the mean values for a group

→ & compares these two means & get a P-values = 0.86.

0.86 > 0.05 would fail to see difference b/w the 2 groups of observation

False Discovery Rate



P = 0.65
= 0.05
= 0.02

Prblm to adjust all the Pvalue by using mechanism called false discovery Rate (FDR)

Mathematical adjustment of Pvalues increases them by some values & in the end the Pvalues are incorrectly came lower adjusted to get higher values 0.05.

Multiple Testing Problem:

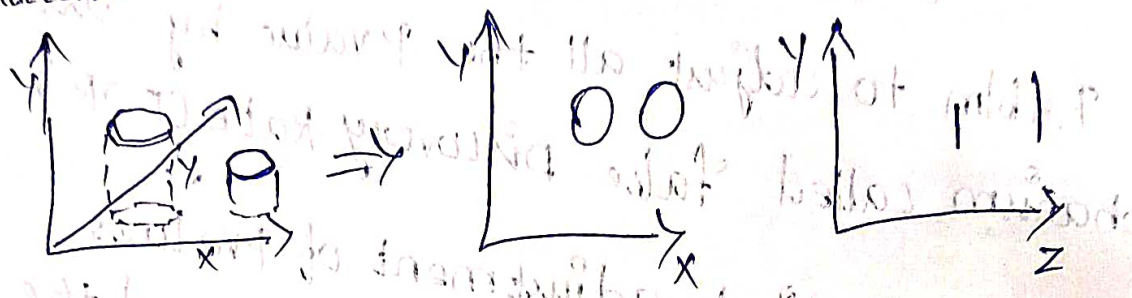
↑ no. of tests.
maximum no. of tests are resulting in rejection of null.
More test will mean that more false Positives
(5% of total tests in our case)
5 out of 100, 50 out of 1000, 500 out of 10,000
Called multiple Testing problem.

Dimensionality Reduction:

Process of reducing the no. of random variables or feature called DR

No. of I/P features, variables or columns Present in given data set known as dimensions

Technique → way of converting higher dimension dataset into lesser dimension data set.



Components DR..

Feature Selection:- find subset of original set of variable or features to get a smaller subset.

Used to model the prblm.

3 ways
- Filter
- Wrapper
- Embedded.

Feature extraction: reduces the data in a high dimensional space to lower dimension space with lesser no. of dimensions.

Methods

PCA Principal Component Analysis

LDA Linear Discriminant "

GDA Generalized Di " "

DR is may both linear ~~or~~ non linear de

PCA:

finds a linear projection of data into orthogonal basis sys.

that has minimum redundancy & preserves the variance in data.

Application PCA/SVD.

Dimensionality reduction

Latent Semantic Indexing

Kleinberg / Hits algorithm

Google / PageRank algorithm

Image - compression

Data visualization [Project data in 2D].

Eigen vector.

A is square matrix

nonzero vector v is eigenvector of A

if there is a scalar λ (eigenvalue)

$$Av = \lambda v$$

Eg: $\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \begin{pmatrix} 3 \\ 2 \end{pmatrix}$

Covariance of matrix of X

$$C_X = \frac{1}{n-1} X^T X$$

symmetric

$$\frac{1}{n-1} (X^T X) = C_X$$

Goal & assumptions of PCA.

$Y = XA$ → covariance matrices capture information

Y ↙ minimal noise & redundancy.

A ↓ best transformation (orthogonal basis vector)

Derivation PCA

C_y : covari of Y expressed in terms of A.

$$C_y = \frac{1}{n-1} Y^T Y$$

$$= \frac{1}{n-1} (XA)^T (XA)$$

$$= \frac{1}{n-1} A^T X^T X A$$

$$= \frac{1}{n-1} A^T (X^T X) A.$$

Assume $A = V$ .ie each column is an eigen vector of $X^T X$.

$$X = VD.$$

$$C_y = \frac{1}{n-1} V^T (X^T X) V$$

$$= \frac{1}{n-1} V^T (V D V^T) V$$

$$= \frac{1}{n-1} V^T V D V^T V$$

$$= \frac{1}{n-1} V^{-1} V D V^{-1} V$$

$$= \frac{1}{n-1} D.$$

Loss function:

Measure estimated value & true value.

$\hat{Y}_i$ → Predicted value given in data set

Loss used in linear loss.

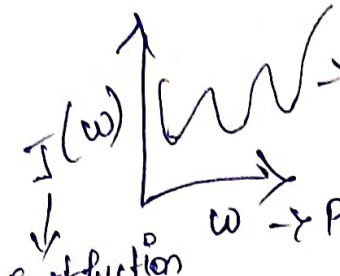
$$= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

↑ = function

Salary of person based on the experie about 20,000 but set. model 30,000

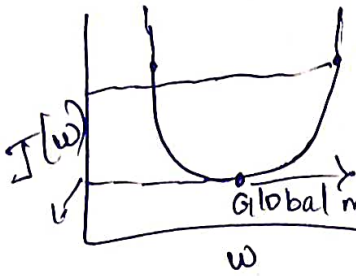
$Y = 1$ or 0

↳ cannot use this



Cost function → in which Parameter value the loss func is very less.

then model is work accurate.



$\hat{Y}$ lies blw 0 - 1

Binary cross Entropy Loss func or Log Loss

$$L(Y, \hat{Y}) = -(Y \log \hat{Y} + (1-Y) \log (1-\hat{Y})) \quad \text{*}$$

where $Y = 1$ $\quad L(1, \hat{Y}) = -(1 \log \hat{Y} + (1-1) \log (1-\hat{Y}))$

$$L(1, \hat{Y}) = -\log \hat{Y}$$

$Y = 0 = L(0, \hat{Y}) = -(0 \log \hat{Y} + (1-0) \log (1-\hat{Y}))$

$$L(0, \hat{Y}) = -\log (1-\hat{Y})$$

we need minimum loss func (L)

↓

Slight difference estimated value & true value

$y=1$

$\hat{y}$ value is large so $-\log \hat{y}^i$ is large negative num

$y=0$

$\hat{y}$ value is small so $-\log \hat{y}^i$ is large - value.

Cost function (J)

↳ $J(w,b) = \frac{1}{m} \sum L(y^i, \hat{y}^i) = -\frac{1}{m} \sum (y^i \log \hat{y}^i + (1-y^i) \log (1-\hat{y}^i))$

m denotes no. of data point in set