Computing for Data Analysis

Spring 2023

Set

Axioms.

Set: Collection of union subjects.

Peano's Axioms.

bucket: [5 oranges, 3 Apples, 4 Mango]

{orange, Apple, mango}

AND (INTERSECTION) (X)

OR (UNION) (+)

→ [5 oranges, 1 Banana] → [4 Banana, 5 Mango]

$A = \{O, B\}$ $|A| = 2$

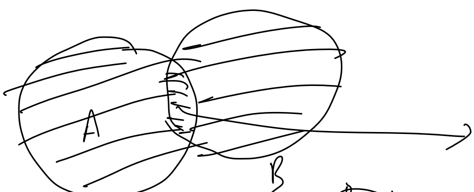
$\{B, M\} = 2$

Common elements: (intersection)

$\{O, B\} \cap \{B, M\} = \{B\}$

All: (Union)

$\{O, B\} \cup \{B, M\} = \{O, B, M\}$



Person who are male

Person who scored >80% in class Xth

$A \cap B =$ all males who scored >80%

$A \cup B =$

S1: I like mango.

S2: I like cricket.

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

$$= \frac{2}{4}$$

$= 50\%$

Greatest Integer function / ~~ceet~~ ceil } methods to round your number

least " " / floor

decimal numbers

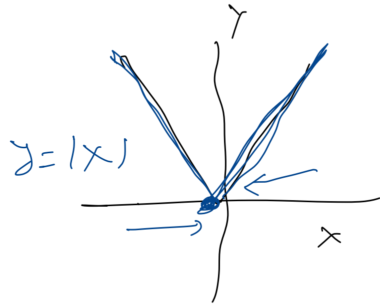
round(5) = 5

round(4.3) = 4

round(4.8) = 5

Ceil → next higher

floor → prev. number



Boolean Operator:

& $\wedge$, $\vee$, $\sim$, !
and, or, NOT

AND

$\wedge \rightarrow$ and

$1 \wedge 1 \rightarrow 1$

$1 \wedge 0 \rightarrow 0$

$0 \wedge 1 \rightarrow 0$

$0 \wedge 0 \rightarrow 0$

OR

$+ \rightarrow +$

$1 + 1 \rightarrow 1$

$1 + 0 \rightarrow 1$

$0 + 1 \rightarrow 1$

$0 + 0 \rightarrow 0$

1, 0

$\wedge$ $\vee$

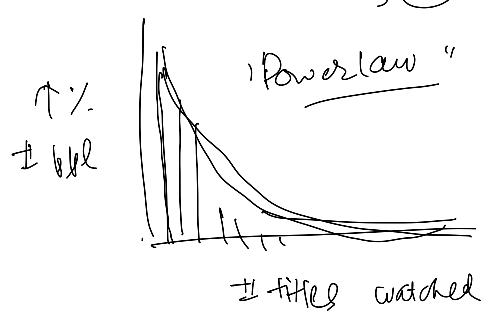
$> 5V \rightarrow 1$

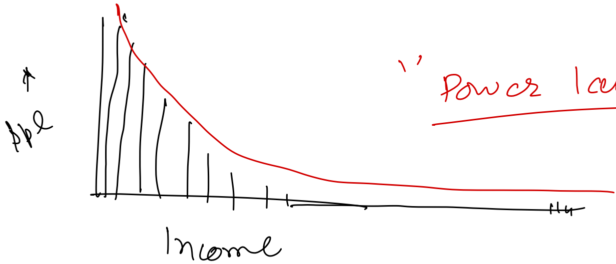
$< 5V \rightarrow 0$

NOT

$\sim 1 \rightarrow 0$

$\sim 0 \rightarrow 1$





"Power law"

← 220M →

users = [0, 1, 2, ... 220M]

17000 {

movie$_1$ = [0 0 0 1 0 0 0] $\underline{220\text{ MB}}$ x17000

movie$_2$ = [1 0 0 . . 0 1 0 0 0] = 3652 GB

⋮

index : 1 ; 2 5

$\underline{\text{sparse}}$

0	3	0	4	0	6

~~45~~

8

Association Rule Mining

grocery =['milk','butter','yogurt','rice’]

How many different pairs (2) of items you can build?

triplets (a, b, c) 4-lets

Milk, Butter
Butter, Yogurt
Yogurt, Rice
Rice, Milk
Milk, Yogurt
Rice, Butter

6

M, B, Y
B, Y, R
Y, R, M
R, M, B

4

(1) (M, B, Y, R)

Counting: (Combinatorics) Combinations

Total C pairs $= {}^{4}C_{2} = \frac{4 \times 3}{2 \times 1} = \frac{4 \times 3}{2!} = 6$

\# pairs $= {}^{4}C_{2} =$

Factorials:

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

$4! = 4 \times 3 \times 2 \times 1 = 24$

$6! = 720$

\# Triplets: $= {}^{4}C_{3} = \frac{4 \times 3 \times 2}{3!} = \frac{4 \times 3 \times 2}{3!} = \frac{4 \times \cancel{3 \times 2}}{\cancel{3 \times 2 \times 1}} = 4$

10 items, (# triplets) $= {}^{10}C_{3}$

$= \frac{10 \times 9 \times 8}{3!} = \frac{10 \times \overset{3}{\cancel{9}} \times \overset{4}{\cancel{8}}}{\cancel{3} \times \cancel{2}} = 120$

Support(X) = (Number of transactions containing X) / (Total number of transactions)
Confidence(X → Y) = (Number of transactions containing X and Y) / (Number of txn. containing X)

TID	Items				
100	A		C	D	
200		B	C		E
300	A	B	C		E
400		B			E

consider 100, 200, 300, and 400 are the unique identifiers of the four transactions: A = sugar, B = bread, C = coffee, D = milk, and E = cake.

The first step is to count the frequencies of k-itemsets

min supp > 50%
min con > 80%

Itemsets	Frequency
$\{A\}$	2
$\{B\}$	3
$\{C\}$	3
$\{D\}$	1
$\{E\}$	3

Itemsets	Frequency
$\{A, B\}$	1
$\{A, C\}$	2
$\{A, D\}$	1
$\{A, E\}$	1
$\{B, C\}$	2
$\{B, E\}$	3
$\{C, D\}$	1
$\{C, E\}$	2

Itemsets	Frequency
$\{A, B, C\}$	1
$\{A, B, E\}$	1
$\{A, C, D\}$	1
$\{A, C, E\}$	1
$\{B, C, E\}$	2

Itemsets	Frequency
$\{A, B, C, E\}$	1

The second step is to generate all the association rules from the frequent itemsets.

Association rules with 1-item consequences from 3-itemsets

RuleNo	Rule	Confidence	support
Rule1	$B \cup C \to E$	100%	50%
Rule2	$B \cup E \to C$	66.7%	50%
Rule3	$C \cup E \to B$	100%	50%

Association rules with 2-item consequences from 3-itemsets

RuleNo	Rule	Confidence	support
Rule4	$B \to C \cup E$	66.7%	50%
Rule5	$C \to B \cup E$	66.7%	50%
Rule6	$E \to B \cup C$	66.7%	50%

Association rules frequent 2-itemsets

RuleNo	Rule	Confidence	support
Rule7	$A \to C$	100%	50%
Rule8	$C \to A$	66.7%	50%

RuleNo	Rule	Confidence	support
Rule9	$B \to C$	66.7%	50%
Rule10	$C \to B$	66.7%	50%

RuleNo	Rule	Confidence	support
Rule11	$B \to E$	100%	75%
Rule12	$E \to B$	100%	75%

RuleNo	Rule	Confidence	support
Rule13	$C \to E$	66.7%	50%
Rule14	$E \to C$	66.7%	50%

	coffee	not coffee	
tea	20	5	25
not tea	70	5	75
	90	10	100

We can apply the support-confidence model to the potential association rule
tea → coffee
The support for this rule is 20%, which is fairly high.
The confidence is the conditional probability that a customer buys coffee, given that he/she buys tea, i.e., P[tea AND coffee]/P[tea]=20/25=0.8, or 80%, which is also fairly high. Hence, the rule tea→ coffee is a valid rule.

Numerical Precision and Stability Analysis

$f(x) =$ function we want to compute

$alg(x) =$ algorithm/ program to compute $f(x)$

$$\underbrace{|alg(x) - f(x)|}$$

How large this can be.

Case # 1

$$|alg(x) - f(x)| \le \epsilon$$

Forward Error

Case # 2

$$alg(x) = f(x + \Delta x)$$

$alg(x)$ is solution to ~~slightly~~ different problem.

Backward Error

If $\frac{\text{fwd}}{\text{bwd}}$ error is small; we say implementation is stable.

Stability is property of implementation of the system

$|alg(x) - f(x)| \le \epsilon$; fwd stable if ϵ is small.

$alg(x) = f(x + \Delta x)$; if Δx is small $\longrightarrow$ backward stable

$$f(x) = \frac{1}{x^4} \qquad \boxed{x = 0.00100} \text{ vs } \boxed{0.00101}$$

If slightly changing your input q drastically changes the result, the problem (math) is ill conditioned.

want: square of a number

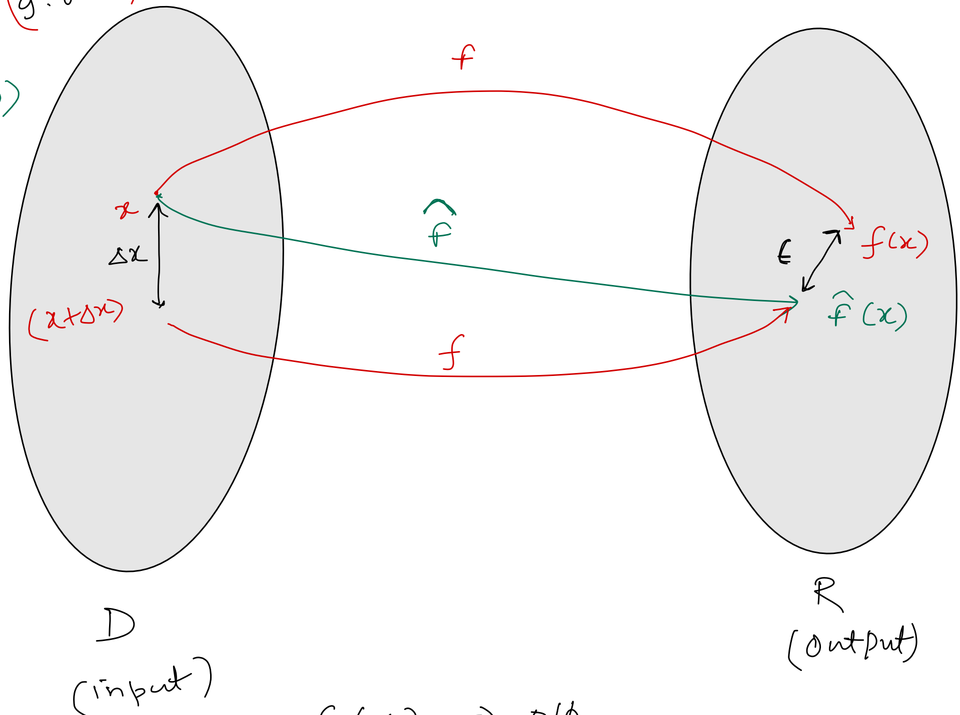
$(3.01)^2 = (9.0001)$

$\downarrow$

(9.00099)

(3.0001)

$f : D \rightarrow R$



$\underbrace{f(x)}_{\text{i/p}} \Rightarrow$ o/p

---

x_0 : $x_0(1+\delta_0)$

$$x_0 = x$$

$$fl(a+b) = (a+b)(1+\delta)$$

$$= a+b+a\delta+b\delta$$

$$= a+b+\underbrace{\delta(a+b)}_{\text{error}}$$

$$x_0+x_1 = (x_0+x_1)(1+\delta_0)$$

$$= x_0+x_0\delta_0+x_1+x_1\delta_0$$

adding x_2

$$= x_0+x_1+\delta_0(x_0+x_1)$$

$$x_0+x_1+x_2 = (x_0+x_1+x_2)(1+\delta_1)+\delta_0(x_0+x_1)$$

$$= x_0+x_1+x_2+x_0\delta_1+x_1\delta_1+x_2\delta_1+\delta_0(\quad)$$

$$= (x_0+x_1+x_2)+\delta_0(x_0+x_1)+\delta_1(x_0+x_1+x_2)$$

$$x_0+x_1+x_2+x_3 = \underline{\qquad}+\delta_0(x_0+x_1+x_2+x_3)+$$

$$\delta_1(x_0+x_2+x_3)$$

$$\delta_2(x_0+x_3)$$

$$\delta_3(x_3)$$

x_1, x_2, x_3, x_4

$0.1, 0.2, 0.3, 0.4$

$$x_1+x_2+x_3+x_4+x_1(4E)+x_2\times 3E+x_3\times 2E$$

$$+x_4\times E$$

$0.1\times 4\times 0.1$

$0.2\times 3\times 0.1$

$0.3\times 2\times 0.1$

0.4×0.1

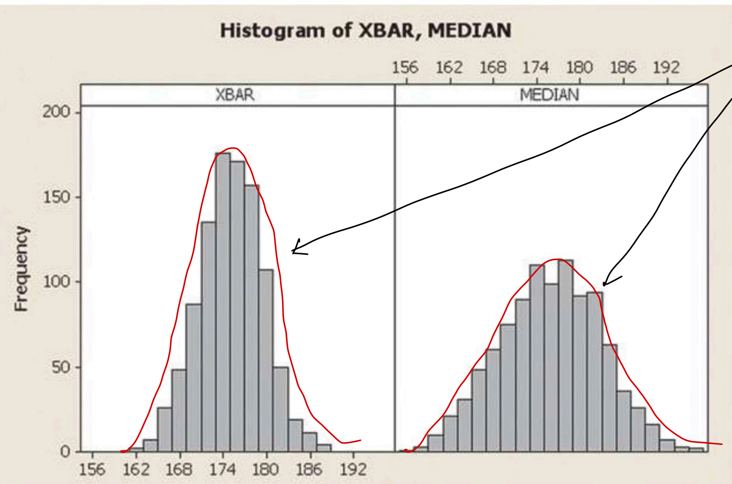
Inferential Statistics

A **parameter** is a numerical descriptive measure of a population. Because it is based on the observations in the population, its value is almost always unknown.

A **sample statistic** is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

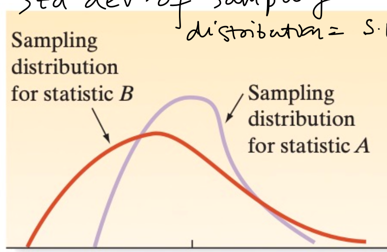
10 Samples of n=11 Height Measurements from XYZ University.

Sample	$\boxed{\mu = 182cm}$ Height ~~Thickness~~ Measurements (in cms.)											Mean	Median
1	173	171	187	151	188	181	182	157	162	169	193	174.00	173
2	181	190	182	171	187	177	162	172	188	200	193	182.09	182
3	192	195	187	187	172	164	164	189	179	182	173	180.36	182
4	173	157	150	154	168	174	171	182	200	181	187	172.45	173
5	169	160	167	170	197	159	174	174	161	173	160	169.46	169
6	179	170	167	174	173	178	173	170	173	198	187	176.55	173
7	166	177	162	171	154	177	154	179	175	185	193	172.09	175
8	164	199	152	153	163	156	184	151	198	167	180	169.73	164
9	181	193	151	166	180	199	180	184	182	181	175	179.27	181
10	155	199	199	171	172	157	173	187	190	185	150	176.18	173



<u>Sampling Distribution</u> is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population

Std dev. of sampling distribution = S.E



	Population Parameter	Sample Statistic
Mean:	μ	$\bar{x}$, $\hat{x}$
Variance:	σ^2	s^2 , $\hat{s}$
Standard deviation:	σ	s
proportion:	p	$\hat{p}$

$\hat{a}$

Properties of the Sampling Distribution of $\bar{x}$

1. The mean of the sampling distribution of $\bar{x}$ equals the mean of the sampled population. That is, $\mu_{\bar{x}} = E(\bar{x}) = \mu$.
2. The standard deviation of the sampling distribution of $\bar{x}$ equals

$$\frac{\text{Standard deviation of sampled population}}{\text{Square root of sample size}}$$

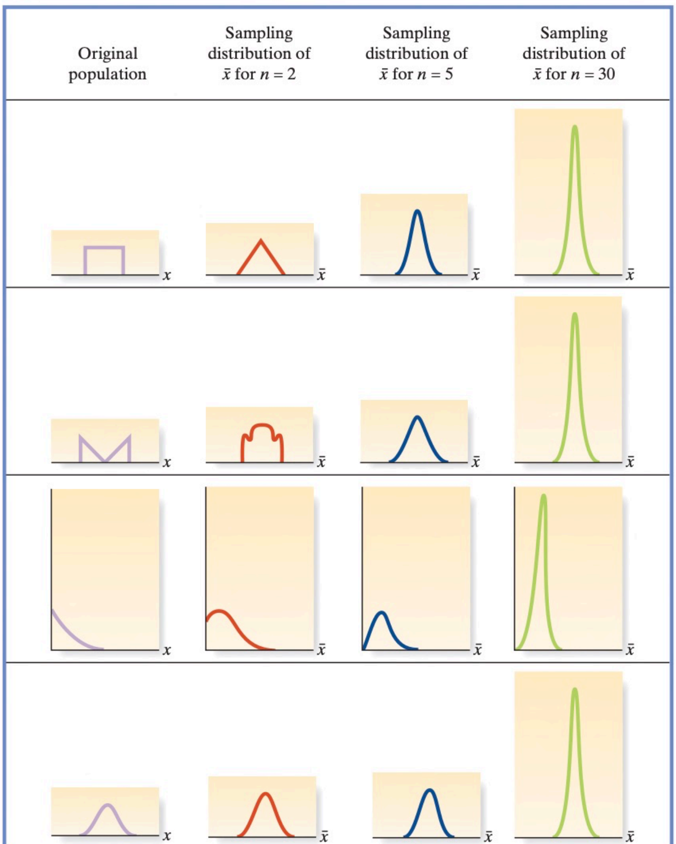
That is, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$*

The standard deviation $\sigma_{\bar{x}}$ is often referred to as the **standard error of the mean**.

"$n > 30$"

Central Limit Theorem

Consider a random sample of n observations selected from a population (*any* population) with mean μ and standard deviation σ. Then, when n is sufficiently large, the sampling distribution of $\bar{x}$ will be approximately a normal distribution with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The larger the sample size, the better will be the normal approximation to the sampling distribution of $\bar{x}$.*



$\sum_{i=1}^{5} x$

$\sum_{i=1}^{5} x$

$x + x + x + x + x$
$= 5x$

A **point estimator** of a population parameter is a rule or formula that tells us how to use the sample data to calculate a *single* number that can be used as an *estimate* of the target parameter.

https://www.youtube.com/watch?v=TqOeMYtOc1w

An **interval estimator (or confidence interval)** is a formula that tells us how to use the sample data to calculate an *interval* that *estimates* the target parameter.

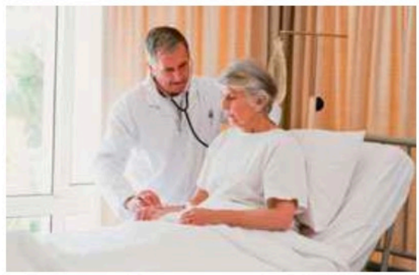
$$(\overline{X} - \text{something}, \ \overline{X} + \text{something})$$

Desired Confidence Interval	Z-Score
90%	1.645
95%	1.96
99%	2.576

$$\boxed{\bar{x} \pm \frac{1.96\sigma}{\sqrt{n}}}$$

For CI w/ 95% Confidence level

Problem Consider the large hospital that wants to estimate the average length of stay of its patients, μ. The hospital randomly samples $n = 100$ of its patients and finds that the sample mean length of stay is $\bar{x} = 4.5$ days. Also, suppose it is known that the standard deviation of the length of stay for all hospital patients is $\sigma = 4$ days. Use the interval estimator $\bar{x} \pm 1.96\sigma_{\bar{x}}$ to calculate a confidence interval for the target parameter, μ.



Solution Substituting $\bar{x} = 4.5$ and $\sigma = 4$ into the interval estimator formula, we obtain:

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm (1.96)\sigma/\sqrt{n} = 4.5 \pm (1.96)(4/\sqrt{100}) = 4.5 \pm .78$$

Or, (3.72, 5.28).

$$\boxed{N = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

$$\Pr\left(a \le \frac{\bar{x}-\mu}{\sigma} \le b\right) = \Phi(b) - \Phi(a)$$

CLT

It is known (as long as n is large enough) that $\bar{X}$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. so $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

$$Z = \left(\frac{\bar{X} - \mu}{\sigma}\right)$$

Find c such that $P\left(-c \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq +c\right) = 0.95$

$$\boxed{\phi(-a) = 1 - \phi(a)}$$

$$\phi(c) - \phi(-c) = 0.95$$

$$\phi(c) - (1 - \phi(c)) = 0.95$$

$$2\phi(c) = 1.95$$

$$\phi(c) = 0.975$$

$$c = \phi^{-1}(0.975)$$

$$c = \boxed{1.96}$$

$1.9 + 0.06 = \boxed{1.96}$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Transaction Management

A → +500 → B (₹)

Net Banking, P2P

$S_i \xleftrightarrow{T} S_f$ ✗ (Inconsistent)

A's Account (−500)

```
Open_Account(A)
Old_Balance = A.balance
New_Balance = Old_Balance - 500
A.balance = New_Balance
Close_Account(A)
```

Reading balance
Modifying balance
Updating balance
(R/W/U)

Transaction is a set of instructions that are executed on a database.

B's Account (+500)

```
Open_Account(B)
Old_Balance = B.balance
New_Balance = Old_Balance + 500
B.balance = New_Balance
Close_Account(B)
```

SQL Name, age, gender, sal, cntry
~~Sel~~
Delete * from table
where
country-code = 'IN'
and age ≥ 30
and gender = 'Male'

1000 → 58

(2nd march, Kolkata, NDL-KOL-RJD)

Date
Time
Boarding
Train

At begin DB is consistent

DB ← Transaction

Consistent

ACID

Atomicity : Entire txn should execute or none. Guaranteed.

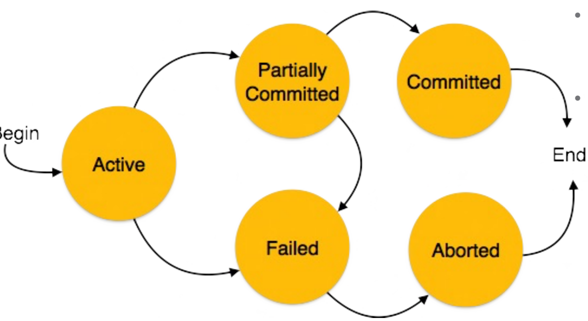
Consistency :

Isolation : logically Isolation

Durability : Changes should be permanent.

DB should also be consistent

roll | marks scored
5 | ~~95~~ ~~96~~ 94



- **Active** – In this state, the transaction is being executed. This is the initial state of every transaction.
- **Partially Committed** – When a transaction executes its final operation, it is said to be in a partially committed state.
- **Failed** – A transaction is said to be in a failed state if any of the checks made by the database recovery system fails. A failed transaction can no longer proceed further.
- **Aborted** – If any of the checks fails and the transaction has reached a failed state, then the recovery manager rolls back all its write operations on the database to bring the database back to its original state where it was prior to the execution of the transaction. Transactions in this state are called aborted. The database recovery module can select one of the two operations after a transaction aborts –
  - Re-start the transaction
  - Kill the transaction
- **Committed** – If a transaction executes all its operations successfully, it is said to be committed. All its effects are now permanently established on the database system.

Stochastic Process / Markov Matrix

Let A be an $n \times n$ matrix. A scalar λ is said to be an **eigenvalue** or a **characteristic value** of A if there exists a nonzero vector $\mathbf{x}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. The vector $\mathbf{x}$ is said to be an **eigenvector** or a **characteristic vector** belonging to λ.

$$A = \begin{pmatrix} 4 & -2 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\underbrace{\begin{bmatrix} 4 & -2 \\ 1 & 1 \end{bmatrix}}_{A \ 2\times 2} \underbrace{\begin{bmatrix} 2 \\ 1 \end{bmatrix}}_{x \ 2\times 1} = \underbrace{\begin{bmatrix} 6 \\ 3 \end{bmatrix}}_{2\times 1} = 3 \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

λ_i : eigen values

(1) $\sum_{i=0}^{n} \lambda_i = tr(A)$ $(a_{11} + a_{22})$ $(\lambda_1 + \lambda_2)$

$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$

(2) $\prod_i \lambda_i = \det(A)$ $\{a_{11} \times a_{22} - a_{21} \times a_{12})$ $\lambda_1 \cdot \lambda_2$

$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$

Find the eigenvalues and the corresponding eigenvectors of the matrix

$$A = \begin{pmatrix} 3 & 2 \\ 3 & -2 \end{pmatrix}$$

$\lambda_1 \cdot \lambda_2 \cdots \lambda_n = \det(A)$
sum $\lambda_i = tr(A)$

$4, -3$ $= 1$

Solution

The characteristic equation is

$$\begin{vmatrix} 3-\lambda & 2 \\ 3 & -2-\lambda \end{vmatrix} = 0 \quad \text{or} \quad \lambda^2 - \lambda - 12 = 0$$

$\begin{vmatrix} 3-\lambda & 2 \\ 3 & -2-\lambda \end{vmatrix}$
$= (3-\lambda)(-2-\lambda) - (3\times 2)$
$= -6 - 3\lambda + 2\lambda + \lambda^2 - 6$
$= \lambda^2 - \lambda - 12 = 0$

Thus, the eigenvalues of A are $\lambda_1 = 4$ and $\lambda_2 = -3$. To find the eigenvectors belonging to $\lambda_1 = 4$, we must determine the null space of $A - 4I$.

$$A - 4I = \begin{pmatrix} -1 & 2 \\ 3 & -6 \end{pmatrix}$$

Solving $(A - 4I)\mathbf{x} = \mathbf{0}$, we get

$$\mathbf{x} = (2x_2, x_2)^T$$

$\lambda^2 - \lambda - 12 = 0$
$\lambda^2 - (4-3)\lambda - 12 = 0$
$\lambda^2 - 4\lambda + 3\lambda - 12 = 0$
$\lambda(\lambda - 4) + 3(\lambda - 4) = 0$
$(\lambda - 4)(\lambda + 3) = 0$
$\lambda = 4, -3$

Hence, any nonzero multiple of $(2, 1)^T$ is an eigenvector belonging to λ_1, and $\{(2, 1)^T\}$ is a basis for the eigenspace corresponding to λ_1. Similarly, to find the eigenvectors for λ_2, we must solve

$$(A + 3I)\mathbf{x} = \mathbf{0}$$

In this case, $\{(-1, 3)^T\}$ is a basis for $N(A + 3I)$ and any nonzero multiple of $(-1, 3)^T$ is an eigenvector belonging to λ_2. ■

Q. B is a 5×5 matrix how many eigen values B will have? = 5

$\begin{bmatrix} \ddots & & \\ & -\lambda & \\ & & \ddots \end{bmatrix} = \lambda^5 + \dots$

A **stochastic process** is any sequence of experiments for which the outcome at any stage depends on chance. A **Markov process** is a stochastic process with the following properties:

I. The set of possible outcomes or states is finite.

II. The probability of the next outcome depends only on the previous outcome.

III. The probabilities are constant over time.

Page Rank is a 1^{st} order Markov chain/Process

also, it is a stochastic process.

A Markov chain is a mathematical process that transitions from one state to another within a finite number of possible states

If a Markov chain with an $n \times n$ transition matrix A converges to a steady-state vector $\mathbf{x}$, *then*

(i) $\mathbf{x}$ *is a probability vector.*

(ii) $\lambda_1 = 1$ *is an eigenvalue of A and* $\mathbf{x}$ *is an eigenvector belonging to λ_1.*

Let us denote the kth state vector in the chain by $\mathbf{x}_k = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})^T$. The entries of each $\mathbf{x}_k$ are nonnegative and sum to 1. For each j, the jth entry of the limit vector $\mathbf{x}$ satisfies

$$x_j = \lim_{k \to \infty} x_j^{(k)} \geq 0$$

and

$$x_1 + x_2 + \cdots + x_n = \lim_{k \to \infty} (x_1^{(k)} + x_2^{(k)} + \cdots + x_n^{(k)}) = 1$$

Therefore the steady-state vector $\mathbf{x}$ is a probability vector. ■

Cf: Steven J. Leon - Linear Algebra with Applications-Pearson (2014) ch6 pp.315-325

Simple Linear Regression

$(\hat{y}, y)$

$\hat{y} \downarrow$ estimate, $y \to$ actual.

$$\boxed{\hat{y} = \beta_0 + \beta_1 x_i}$$

$$\underbrace{(y - \hat{y})^2}_{\text{Sum of Squared Residual}} \qquad SSR = (y_i - \hat{y}_i)^2$$

minimize SSR

$$SSR = \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

$$SSR = \sum_i \left(y_i^2 - 2 y_i (\beta_0 + \beta_1 x_i) + (\beta_0 + \beta_1 x_i)^2\right)$$

$$SSR = \sum_i \left(y_i^2 + \beta_0^2 + \beta_1^2 x_i^2 + 2\beta_0\beta_1 x - 2 y_i \beta_0 - 2\beta_1 x_i y_i\right)$$

$$\frac{\partial SSR}{\partial \beta_0} = \sum_i \left(0 + 2\beta_0 + 0 + 2\beta_1 x_i - 2 y_i - 0\right)$$

$$= \sum_i \left[2\beta_0 + 2\beta_1 x_i - 2 y_i\right] \quad \text{———} \quad (1)$$

$$\frac{\partial SSR}{\partial \beta_0} = 0$$

$$\sum_i (2\beta_0 + 2\beta_1 x_i - 2 y_i) = 0$$

$$\Rightarrow 2\sum_{i=1}^{n} \beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i = 0$$

$$\Rightarrow n\beta_0 = \sum y_i - \beta_1 \sum x_i$$

$$\Rightarrow \beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\beta_0 = \frac{y_1 + y_2 + \cdots y_n}{n} - \beta_1 \times \frac{(x_1 + x_2 + \cdots x_n)}{n}$$

$$\boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$$

$$SSR = \sum_i \left(y_i^2 + \beta_0^2 + \beta_1^2 x_i^2 + 2\beta_0\beta_1 x - 2y_i\beta_0 - 2\beta_1 x_i y_i \right)$$

$$\frac{\partial SSR}{\partial \beta_1} = \sum_i \left(2\beta_1 x_i^2 + 2\beta_0 x_i - 0 - 2x_i y_i \right) = 0$$

$$= \sum_i -2x_i \left(y_i - (\beta_0 + \beta_1 x) \right) = 0$$

$$= \sum x_i \left[y_i - (\underline{\beta_0} + \beta_1 x) \right] = 0$$

$$= \sum x_i \left[y_i - (\bar{y} - \beta_1 \bar{x} + \beta_1 x_i) \right] = 0$$

$$= \sum x_i \left[y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i \right] = 0$$

$$= \sum x_i \left[(y_i - \bar{y}) - \beta_1 (x_i - \bar{x}) \right] = 0$$

$$= \sum x_i (y_i - \bar{y}) - \beta_1 x_i (x_i - \bar{x}) = 0$$

$$\Rightarrow \sum x_i (y_i - \bar{y}) = \beta_1 \sum x_i (x_i - \bar{x})$$

$$\boxed{\beta_1 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}}$$

$$\boxed{\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}}$$

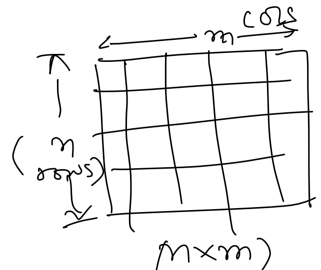
OLS & Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \underbrace{\epsilon}_{\text{residual}}$$

Consider our dataset ~~tabular~~ (n rows, m cols)

$$\underset{(n\times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}} = \underset{n\times(m+1)}{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \ddots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}} \underset{(m+1)\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}} + \underset{(n\times 1)}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

$$\underbrace{Y}_{\text{response}} = \underbrace{X}_{\text{design matrix}} \underbrace{\beta}_{\text{coeff matrix}} + \underbrace{\epsilon}_{\text{noise matrix}}$$



$$\epsilon = (Y - X\beta)$$

To minimize, inner product of ϵ

$$\epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta)$$

$$= \left(Y^T - (X\beta)^T\right)(Y - X\beta)$$

$$= \left(Y^T - \beta^T X^T\right)(Y - X\beta)$$

$$= Y^T Y - \underline{Y^T X \beta} - \underline{\beta^T X^T Y} + \beta^T X^T X \beta$$

$1\times n \times n \times (m+1)(m+1)\times 1$

$1 \times 1 = 1$

it is scalar

$(1\times(m+1)(m+1)\times n \times n\times 1$

(1×1)

Matrix formula

$$(A-B)^T = A^T - B^T$$

$$(AB)^T = B^T A^T$$

$$\beta^T X^T Y = Y^T X \beta = \text{scalar}$$

$$\epsilon^T \epsilon = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta$$

$$\frac{\partial \epsilon^T \epsilon}{\partial \beta} = -2X^T Y + 2X^T X \beta$$

matrix Calculus

$$\frac{\partial a^T b}{\partial b} = \frac{\partial b^T a}{\partial b} = a$$

$$\frac{\partial b^T A b}{\partial b} = 2Ab = 2b^T A$$

$$\frac{\partial \epsilon^T \epsilon}{\partial \beta} := 0$$

$$-2X^T Y + 2X^T X \beta = 0$$

$$\boxed{X^T Y = X^T X \beta}$$

Normal form

$$X^T X \beta = X^T Y$$

$$\underbrace{(X^T X)^{-1} (X^T X)}\beta = (X^T X)^{-1} X^T Y$$

$$\boxed{\beta = (X^T X)^{-1} X^T Y}$$

Ordinary least squares estimate of β.

How to remember

$$Y = X\beta \rightarrow \beta = \frac{Y}{X}$$

$$= \frac{X^T Y}{\underbrace{(X^T X)}_{\text{square matrix}}} = (X^T X)^{-1} X^T Y$$

Some Question Answers

Hypothesis Testing

> 18 yr, M, India → avg. age

Hypothesis testing or **significance testing** is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample.

Step 1: State the hypotheses. The **null hypothesis (H_0)**, stated as the **null**, is a statement about a population parameter, such as the population mean, that is assumed to be true. The null hypothesis is a starting point. We will test whether the value stated in the null hypothesis is likely to be true. Remember, only reason we are testing the null hypothesis is because we think it is wrong. An **alternative hypothesis (H_1)** is a statement that directly contradicts a null hypothesis by stating that that the actual value of a population parameter is less than, greater than, or not equal to the value stated in the null hypothesis.

Step 2: Set the criteria for a decision. To set the criteria for a decision, we state the **level of significance** for a test. ~~**Level of significance**~~, or significance level, refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of obtaining a statistic measured in a sample if the value stated in the null hypothesis were true. In experimental science, the criterion or level of significance is typically set at ~~5%~~. When the probability of obtaining a sample mean is less than ~~5%~~ if the null hypothesis were true, then we reject the value stated in the null hypothesis.

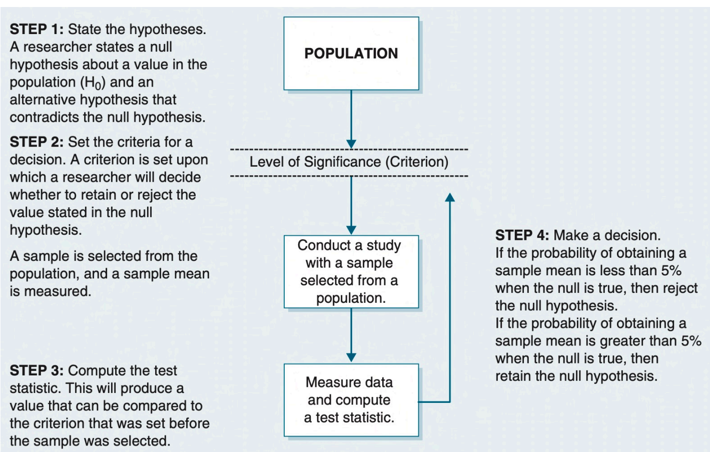
→ t stat, z stat, F stat → p value

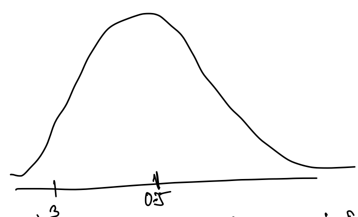
Step 3: Compute the test statistic. The **test statistic** is a mathematical formula that allows researchers to determine the likelihood of obtaining sample outcomes if the null hypothesis were true. The value of the test statistic is used to make a decision regarding the null hypothesis.

Step 4: Make a decision. We use the value of the test statistic to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean, given that the value stated in the null hypothesis is true.

- If the probability of obtaining a sample mean is less than 5% when the null hypothesis is true, then the decision is to reject the null hypothesis.
- If the probability of obtaining a sample mean is greater than 5% when the null hypothesis is true, then the decision is to retain the null hypothesis.

The **p-value** is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.





H_0: prop. of smoker in India = 0.5

H_a: " " " ≠ 0.5

reject H_0 | not reject H_0

1000 — $\frac{300}{1000} = .30$

Decisions are made about the **null hypothesis**. Using the courtroom analogy, a judge decides whether a defendant is guilty or not guilty. The judge does not make a decision of guilty or *innocent* because the defendant is assumed to be innocent. All evidence presented in a trial is to show that a defendant is guilty. The evidence either shows guilt (decision: guilty) or does not (decision: not guilty). In a similar way, the null hypothesis is assumed to be correct. A researcher conducts a study showing evidence that this *assumption is unlikely* (we reject the null hypothesis) or *fails to do so* (we retain the null hypothesis).

H_0: person not guilty

reject H_0

The following are the steps followed in the performance of the t-test:

1. Set the significance level for the test.
2. Formulate the null and the alternative hypotheses.
3. Calculate the t-statistic using the formula below:

$$t = \frac{\widehat{b_1} - b_1}{s_{\widehat{b_1}}}$$

Where:

b_1 = True slope coefficient.

$\widehat{b_1}$ = Point estimate for b_1

$b_1 s_{\widehat{b_1}}$ = Standard error of the regression coefficient.

4. Compare the absolute value of the t-statistic to the critical t-value (t_c). Reject the null hypothesis if the absolute value of the t-statistic is greater than the critical t-value i.e., $t > +t_{critical}$ *or* $t < -t_{\text{critical}}$.

(table)

Method for finding estimates of co-eff in linear regression → Ordinary least square.

$$\boxed{\beta = (X^TX)^{-1}X^Ty}$$

Regression Statistics	
Multiple R	0.9971
R Square	0.9941
Adjusted R Square	0.9922
Standard Error	3.6515
Observations	5

	Coefficients	Standard Error	t Stat	P-value
Intercept	−159	10.520	(15.114)	0.001
Slope	0.26	0.012	22.517	0.000

The t-statistic is calculated using the formula:

$H_0: B_1 = 0$

$H_a: B_1 \neq 0$

Testing of beta₁:

$$\boxed{t = \frac{\widehat{b_1} - b_1}{\widehat{S_{b_1}}}}$$

Where:

- b_1 = True slope coefficient
- $\widehat{b_1}$ = Point estimator for b_1
- $\widehat{S_{b_1}}$ = Standard error of the regression coefficient

$$t = \frac{0.26 - 0}{0.012}$$

$$3.182 < \;= 21.67$$

The critical two-tail t-values from the t-table with $n - 2 = 3$ degrees freedom are:

$$t_c = \pm 3.18$$

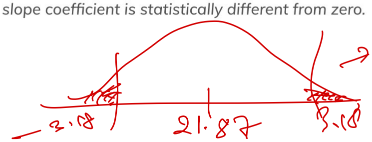
cf: https://analystprep.com/cfa-level-1-exam/quantitative-methods/hypothesis-testing-in-regression-analysis/

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	[illegible]	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	[illegible]	6.965	9.925	22.327	31.599
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Notice that $|t| > t_c$ (i.e $21.67 > 3.18$).

Therefore, the null hypothesis can be rejected. Further, we can conclude that the estimated slope coefficient is statistically different from zero.



→ 0%

QR Decomposition for beta estimation

$$\boxed{X^T X\beta = X^T y}$$

Normal form.

$X = QR$ (QR decomposition)

~~Q~~ Q is orthogonal, $Q^T Q = I$

R is upper triangular.

$$X^T X \beta = X^T y$$

$$(QR)^T QR\beta = (QR)^T y$$

$$(AB)^T = B^T A^T$$

$$R^T Q^T Q R\beta = R^T Q^T y$$

$$R^T R\beta = R^T Q^T y$$

$$(R^T)^{-1} R^T \cdot R\beta = (R^T)^{-1} R^T Q^T y$$

$$\boxed{R\beta = Q^T y}$$

R is upper triangular matrix

Benefits

- stable Estimate
- No need of matrix inversion

Drawback is QR decomp. is costly.

Gram Schmidt orthogonalization

Gradient Descent

loss function: $X^T X\beta = X^T y$

$$\mathcal{L} = X^T X\beta - X^T y$$

$$\min_{\beta} \mathcal{L} = \boxed{X^T(X\beta - y)}$$

$$\operatorname*{argmin}_{\beta} X^T(X\beta - y)$$

$$\beta^{new} = \beta^{old} - \eta \nabla \ell(\alpha)$$

If loss function is convex then global minima is guaranteed.

Maximum Likelihood Estimation

Prob. distribution

$N(\mu, \sigma^2)$ — $\boxed{\text{Population}}$ $\xrightarrow{\text{sampling}}$ $\boxed{\text{Samples}}$

unobserved (RCT/ RS)

Estimation

estimate: $\boxed{\text{mean}}$ → statistic] → NHST → Inference

val. of mean → estimate

Coin ($n=10$) $\boxed{\text{HTHHTTTHTHTT}}$

prob. of Head = ? $\frac{4}{10} = 40\% = \boxed{0.4}$

let prob. of Head $= p$ & prob. tail $= (1-p)$

able to write this because we know distribution

Likelihood: $\mathcal{L} = p\cdot(1-p)\, p\cdot(1-p)(1-p)\, p\cdot(1-p)\, p\,(1-p)(1-p)$

$= p^4(1-p)^6$

log likelihood $\log \mathcal{L} = \log\left(p^4 (1-p)^6\right)$

$= \log p^4 + \log (1-p)^6$

$\log \mathcal{L} = 4\log p + 6\log(1-p)$

$\log(a\cdot b) = \log a + \log b$

$\log a^m = m\log a$

$\frac{\partial \log \mathcal{L}}{\partial p} = \frac{4}{p} + \frac{6}{(1-p)}\cdot(-1)$ [chain rule] $\overset{\text{set}}{=} 0$

$\frac{4}{p} = \frac{6}{1-p}$

$4 - 4p = 6p$

$10p = 4$

$\boxed{p = \frac{4}{10}}$

Example Suppose that X is a discrete random variable with the following probability mass function: where $0 \le \theta \le 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

Solution: Since the sample is (3,0,2,1,3,2,1,0,2,1), the likelihood is

$$L(\theta) = P(X=3)P(X=0)P(X=2)P(X=1)P(X=3) \times P(X=2)P(X=1)P(X=0)P(X=2)P(X=1)$$

Substituting from the probability distribution given above, we have

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

Let us look at the log likelihood function

$$\begin{aligned} l(\theta) &= \log L(\theta) = \sum_{i=1}^{n} \log P(X_i|\theta) \\ &= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + 2\left(\log\frac{1}{3} + \log(1-\theta)\right) \\ &= C + 5\log\theta + 5\log(1-\theta) \end{aligned}$$

where C is a constant which does not depend on θ. It can be seen that the log likelihood function is easier to maximize compared to the likelihood function.

Let the derivative of $l(\theta)$ with respect to θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

and the solution gives us the MLE, which is $\hat{\theta} = 0.5$.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal distribution with unknown mean μ and variance σ^2. Find maximum likelihood estimators of mean μ and variance σ^2.

Answer

In finding the estimators, the first thing we'll do is write the probability density function as a function of $\theta_1 = \mu$ and $\theta_2 = \sigma^2$:

$$f(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}} \exp\left[-\frac{(x_i - \theta_1)^2}{2\theta_2}\right]$$

for $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. We do this so as not to cause confusion when taking the derivative of the likelihood with respect to σ^2. Now, that makes the likelihood function:

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2) = \theta_2^{-n/2}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right]$$

and therefore the log of the likelihood function:

$$\log L(\theta_1, \theta_2) = -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i - \theta_1)^2}{2\theta_2}$$

Now, upon taking the partial derivative of the log likelihood with respect to θ_1, and setting to 0, we see that a few things cancel each other out, leaving us with:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \frac{-\cancel{2}\sum(x_i - \theta_1)\cancel{(-1)}}{\cancel{2}\,\theta_2} \overset{\text{SET}}{\equiv} 0$$

Now, multiplying through by θ_2, and distributing the summation, we get:

$$\sum x_i - n\theta_1 = 0$$

Now, solving for θ_1, and putting on its hat, we have shown that the maximum likelihood estimate of θ_1 is:

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Now for θ_2. Taking the partial derivative of the log likelihood with respect to θ_2, and setting to 0, we get:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{\sum(x_i - \theta_1)^2}{2\theta_2^2} \overset{\text{SET}}{\equiv} 0$$

Multiplying through by $2\theta_2^2$:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \left[-\frac{n}{2\theta_2} + \frac{\sum(x_i - \theta_1)^2}{2\theta_2^2} \overset{set}{\equiv} 0\right] \times 2\theta_2^2$$

we get:

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

And, solving for θ_2, and putting on its hat, we have shown that the maximum likelihood estimate of θ_2 is:

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

(I'll again leave it to you to verify, in each case, that the second partial derivative of the log likelihood is negative, and therefore that we did indeed find maxima.) In summary, we have shown that the maximum likelihood estimators of μ and variance σ^2 for the normal model are:

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

respectively.

Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights (in pounds):

```
115 122 130 127 149 160 152 138 149 180
```

Based on the definitions given above, identify the likelihood function and the maximum likelihood estimator of μ, the mean weight of all American female college students. Using the given sample, find a maximum likelihood estimate of μ as well.

Answer

The probability density function of X_i is:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

for $-\infty < x < \infty$. The parameter space is $\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty\}$. Therefore, (you might want to convince yourself that) the likelihood function is:

$$L(\mu, \sigma) = \sigma^{-n}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

for $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. It can be shown (we'll do so in the next example!), upon maximizing the likelihood function with respect to μ, that the maximum likelihood estimator of μ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Based on the given sample, a maximum likelihood estimate of μ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(115 + \cdots + 180) = 142.2$$

$(\hat{\theta} - \theta) \longrightarrow 0$ as $n \longrightarrow \infty$

Properties of Estimator:

- UNBIASEDNESS: An estimator is said to be unbiased if in the long run it takes on the value of the population parameter. That is, if you were to draw a sample, compute the statistic, repeat this many, many times, then the average over all of the sample statistics would equal the population parameter.
- EFFICIENCY: An estimator is said to be efficient if in the class of unbiased estimators it has minimum variance.
- SUFFICIENCY: We say that an estimator is sufficient if it uses all the sample information. The median, because it considers only rank, is not sufficient. The sample mean considers each member of the sample as well as its size, so is a sufficient statistic.
- CONSISTENCY: If an estimator, say θ, approaches the parameter θ closer and closer as the sample size n increases, θ is said to be a consistent estimator of θ.

SLR

$\hat{\beta} =$

$(X^TX)^{-1}X^Ty$

unbiased esti

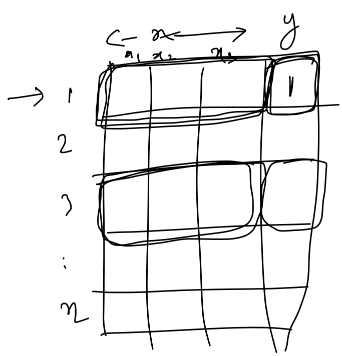
BLUE

best uncae

Derivation of Logistic Regression

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right]$$
$$= \frac{d}{dx}(1+\mathrm{e}^{-x})^{-1}$$
$$= -(1+e^{-x})^{-2}(-e^{-x})$$
$$= \frac{e^{-x}}{(1+e^{-x})^2}$$
$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$
$$= \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}}$$
$$= \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$
$$= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right)$$
$$= \sigma(x) \cdot (1-\sigma(x))$$



need to build LR model such that if $y_i = 1$ then p_i = ~~p_i~~ or $(1-p_i)$

$y = 1$

$$\mathcal{L} = p^y \cdot (1-p)^{1-y}$$
$$= p^1 \, (1-p)^{\cancel{1-1}}$$
$$p^0 \cdot (1-p)^{1-0} = (1-p)$$

For each training data-point, we have a features x_i and an observed class, y_i. The probability of the class is p, if $y_i = 1$, or $1-p$ if $y_i = 0$

$$P(Y|X) = p(x_i)^{y_i} \cdot (1-p(x_i))^{1-y_i}$$

$$\mathcal{L} = P(Y|X) = \hat{y}^{y_i} \cdot (1-\hat{y})^{(1-y)}$$

$$\hat{y} = \sigma(\beta_0 + \beta_1 \cdot x)$$
$$\log l = y \log \hat{y} + (1-y)\log(1-\hat{y})$$
$$= y \log \sigma(\beta_0 + \beta_1 \cdot x) + (1-y)\log(1-\sigma(\beta_0 + \beta_1 \cdot x))$$

$$\frac{\partial \log l}{\partial \beta_j} = \frac{y}{\sigma(\beta_0+\beta_1\cdot x)}\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j} + \frac{(1-y)}{(1-\sigma(\beta_0+\beta_1\cdot x))}\frac{\partial(1-\sigma(\beta_0+\beta_1\cdot x))}{\partial\beta_j}$$

$$= \frac{y}{\sigma(\beta_0+\beta_1\cdot x)}\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j} - \frac{(1-y)}{1-\sigma(\beta_0+\beta_1\cdot x)}\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j}$$

$$= \left[\frac{y}{\sigma(\beta_0+\beta_1\cdot x)} - \frac{1-y}{1-\sigma(\beta_0+\beta_1\cdot x)}\right]\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j}$$

$$\boxed{\frac{\partial\sigma(\beta)}{\partial\beta} = \sigma(\beta)(1-\sigma(\beta))}$$

$$= \frac{y - y\sigma(\beta_0+\beta_1\cdot x) - \sigma(\beta_0+\beta_1\cdot x) + y\sigma(\beta_0+\beta_1\cdot x)}{\sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))} \cdot \sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))\frac{\partial(\beta_0+\beta_1\cdot x)}{\partial\beta_j}$$

$$= \frac{y - \cancel{y\sigma(\beta_0+\beta_1\cdot x)} - \sigma(\beta_0+\beta_1\cdot x) + \cancel{y\sigma(\beta_0+\beta_1\cdot x)}}{\cancel{\sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))}} \cdot \cancel{\sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))} \cdot x$$

$$= (y - \sigma(\beta_0+\beta_1\cdot x)) \cdot x$$

$$-\partial L = -(y-\hat{y})\,x$$
$$= (\hat{y}-y)\,x$$

$$\boxed{-\frac{\partial \log l}{\partial \beta_j} \implies (\hat{y}-y)x}$$

Stochastic Gradient Descent

$$\beta = \beta - \eta \cdot \nabla l(\beta)$$

$$\frac{\text{Sigmoid}}{\text{Logistic}} = \frac{1}{1+e^{-x}}$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}\left(\frac{1}{x}\right) = -\frac{1}{x^2}$$

$$\frac{d}{dx}(e^{-x}) = -e^{-x}$$

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$$= -\frac{1}{(1+e^{-x})^2}(-e^{-x})$$

$$= \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$

$$= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right)$$

$$\sigma'(3) = \sigma(3)(1-\sigma(3))$$

$$\boxed{\frac{d}{dx}(\sigma(x)) = \sigma(x)(1-\sigma(x))}$$

		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum TP}{\sum TP + FP}$$

$$\text{F1 measure} = \frac{2 \cdot P \times R}{P + R}$$

(Harmonic mean)

$$\text{Recall} = \frac{\sum TP}{\sum TP + FN}$$

$$\text{Accuracy} = \frac{\sum TP + TN}{\sum TP + FP + FN + TN}$$

Dataset not balanced

Labels is not balanced

no. of positive classes $\neq$ no. of neg. classes.

$(n = 1000)$ $(+ \Rightarrow 400)$ $(- \Rightarrow 600)$

100 Question $\rightarrow$ T | F

90 10

a, b & c are in A.P.

$\frac{1}{a}, \frac{1}{b}$ & $\frac{1}{c}$ in H.P.

$$\frac{2}{b} = \frac{1}{a} + \frac{1}{c}$$

$= \cdot$

Clustering

$S = \{x_1, x_2 \ldots x_n\}$ Given a dataset and number of cluster 'k', group your dataset in 'k' groups.

Goal:

Divide dataset into 'k' subsets / cluster / groups / partitions.

$$\underbrace{A_1, A_2, A_3 \ldots A_k}_{k \text{ partitions}}$$

Properties of partition:

① $A_i \neq \phi \quad \forall i \in k$ (no partition should be empty)

② $A_i \cap A_j = \phi \quad \forall i, j \in k$ (there should not be a single pt in two cluster)

③ $\bigcup_{i=1}^{K} A_i = S$ (All points are assigned some or other cluster).

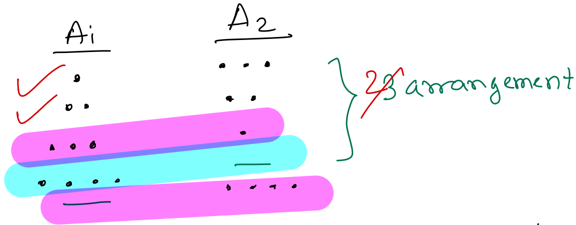
$$\underline{\underline{\mathcal{L}\, P(A_1, A_2 \ldots A_k)}}_{\min} = \underbrace{\sum_{i=1}^{K} \sum_{x_i \in A_i} d(x_i, \bar{A}_i)}_{\text{Sum over all such points}}$$

datapoint

cluster center

distance → [eucledian, minkowski distance, taxicab, Jaccard, L₁ norm]

How bad the cost function is?



$P(n = 4, k = 2) = 2 \ldots\ldots \longrightarrow$

$$\frac{2^{n-2}}{2} = \frac{2^{4-2}}{2} = \underline{\underline{2}} \checkmark$$

n points into k clusters

$S(n, k)$ ← Stirling's number of 2nd kind

$$= \frac{1}{k!} \sum_{i=0}^{k} (-1)^k \binom{k}{i} (k-i)^n$$

Partition

$P(2) = \begin{matrix} 1+1 \\ 2+0 \end{matrix} = 2$

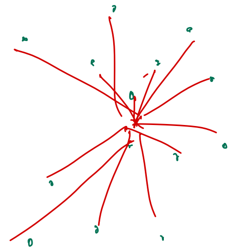
$P(3) = \begin{matrix} 1+1+1 \\ 2+1 \\ 3+0 \end{matrix} \quad 3$

$P(5) = \begin{matrix} 1+1+1+1+1 \\ 2+3 \\ 4+1 \\ 5+0 \end{matrix}$

$P(100)$

Cluster Analysis Proof

Property 1: The best choice for the centroids $c_1, \ldots, c_k$ are the n-tuples which are the means of the $C_1, \ldots, C_k$. By best choice, we mean the choices that minimize SS_E.



Proof: By calculus, the minimum of SS_E is achieved when $\frac{\partial SS_E}{\partial c_{ij}} = 0$ for all i and j where c_{ij} is the ith element in the n-tuple for c_j. Now

$$0 = \frac{\partial SS_E}{\partial c_{ij}} = \frac{\partial}{\partial c_{ij}} \sum_{j=1}^{k} \sum_{x \in C_j} \sum_{i=1}^{n} (x_i - c_{ij})^2 = 2 \sum_{x \in C_j} (x_i - c_{ij}) = 0$$

Thus

$\sum_{x \in C_j} x_i = \sum_{x \in C_j} c_{ij}$

$$\sum_{x \in C_j} x_i = \sum_{x \in C_j} c_{ij} = c_{ij} m_j$$

m

γ_{d_1}

$\gamma_{d_2} > \gamma_{d_1}$

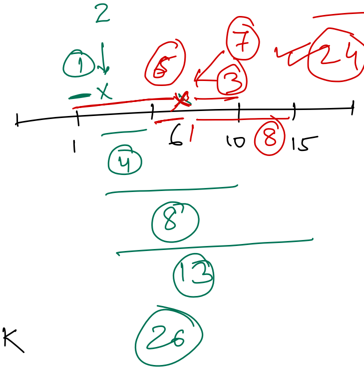
and so

$$\boxed{c_{ij} = \frac{1}{m_j} \sum_{x \in C_j} x_i}$$

for all i, which means that

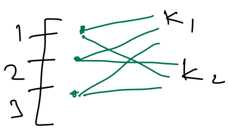
$$c_j = \frac{1}{m_j} \sum_{x \in C_j} x$$

where operations on n-tuples are defined element by element (as for vectors).



Lloyd's Algorithm for K means:

1. Initialize K centres (Select K datapoints as K center's)
2. (S) Compute distance of each datapoint with 'k' center's, → Store info a distance matrix

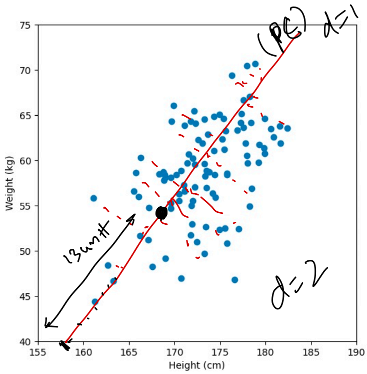


$$\begin{array}{c|ccc} & k_1 & k_2 & k_3 \\ \hline pt1 & \circ & \circ & \circ \\ pt2 & \circ & \circ & \circ \\ pt3 & \circ & \circ & \circ \end{array}$$

3. Assign Cluster label to each pt wrt min. distance from center (use S)
4. Compute new center location for each labeled data pts.
5. Compute WCSS (for convergence).
6. Check for convergence and halt.

Dimensionality Reduction

PCA is dimensionality Reduction Technique.
The PCAs of your data are the eigenvectors of your data's covariance matrix



var (Height) = V_h

var (weight) = V_w

✓ if $V_h > V_w$ then: $(x, 0)$

✓ if $V_w > V_h$ then: $(0, y)$

$\left(\frac{\sigma_x}{\bar{x}}\right)$ $\left(\frac{\sigma_x}{\bar{y}}\right) \rightarrow \left(\frac{cm}{cm}\right)$ = Co-eff of variance.

Rank: # of independent rows / cols.

$$A = \begin{array}{c} \\ \\ \\ \end{array}\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 1 & 3 & 5 \end{bmatrix}\begin{array}{l} R_1 \\ R_2 \\ R_3 \end{array} \qquad R_3 := R_1 + R_2$$

$$R(A) = 2$$

independent rows in your matrix will form **Basis.**

Basis of A = $[1\ 2\ 3]$ and $[0\ 1\ 2]$

Hence our new co-ordinate system will be $[1\ \ 0]$ $[0\ 1]$ $[1\ \ 1]$

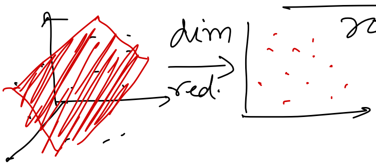
$R_1 + 0R_2$ $0 \cdot R_1 + R_2$ $R_1 + R_2$

old co-ord system : $\underbrace{(1,0,0)}_{x}$ $\underbrace{(0,1,0)}_{y}$ $\underbrace{(0,0,1)}_{z}$

✓ $[1, 2, 3]$ $[0, 1, 2]$

so, A can be represented as

$\underbrace{[1\ \ 0]}_{\text{row 1}}$ $\underbrace{[0\ \ 1]}_{\text{row 2}}$ $\underbrace{[1\ \ 1]}_{\text{row 3}}$



input Data → SVD → PCA → Output

Dim. Reduction (k)

$105 = 3 \times 5 \times 7$

SVD: Singular Value Decomposition.

$$\underset{(m \times n)}{X} = \underset{(m \times m)}{U} \cdot \underset{(m \times r)}{\Sigma} \cdot \underset{(r \times n)}{V^T}$$

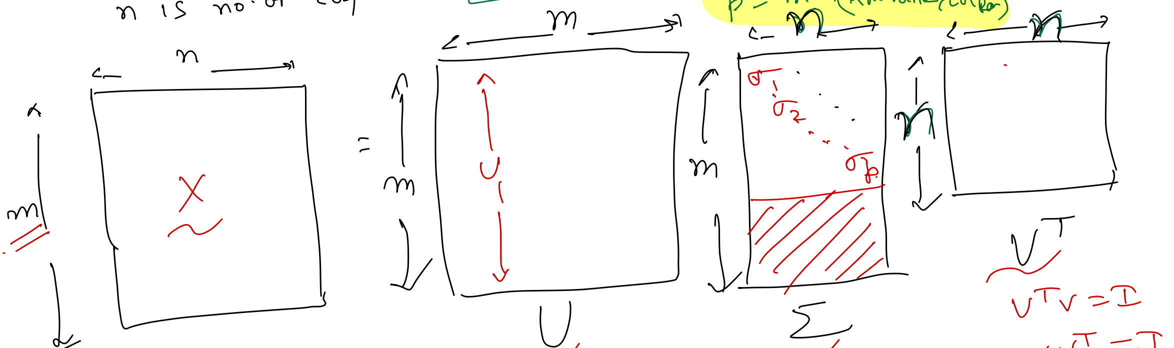
m is no. of examples

n is no. of col/ dim.

$m \times n = m \times m \times m \times n \times n \times n$

$p = \min(\text{Row rank}, \text{Col}_{\text{Rank}})$

Row and Columns are orthonormal



U

Colums are orthonormal

$U^T U = I$

Σ

Singular matrix (eigen value)

$\boxed{\sigma_1 > \sigma_2 > \sigma_3}, \ldots > \sigma_p$

V^T

$V^T V = I$

$V V^T = I$

$(AB)^T = B^T A^T$

$(A^T)^T = A$

$$X = U \Sigma V^T$$

$$X^T X = V \Sigma^T U^T U \Sigma V^T$$

$$= V \Sigma^T \Sigma V^T$$

$$X^T X = V \underline{D} V^T$$

(D is diag. matrix with squared of singular values)

$$(X^T X) V = V D V^T V$$

$$\boxed{X^T X V = V D}$$

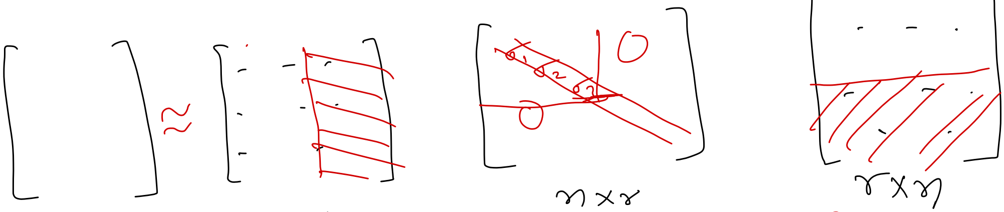
eigenvector, eigenvalue

$\boxed{Ax = \lambda x}$

$\frac{1}{N}\left(X^T X\right)$ covariance matrix.

$105 = 3 \times 5 \times 7 \approx 3 \times 5 \times 5$

$A = U \Sigma V^T$



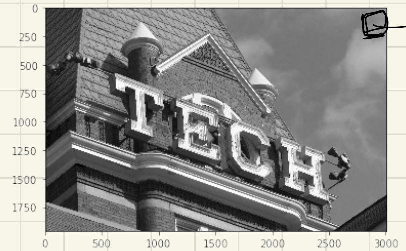
$m \times n \approx m \times n \quad n \times r \quad r \times n$

$= m \times 3 \quad 3 \times 3 \quad 3 \times n$

r = rank of matrix (A).

$r \leq \min(m, n)$

r can be atleast min. of no. of rows or cols in original matrix.



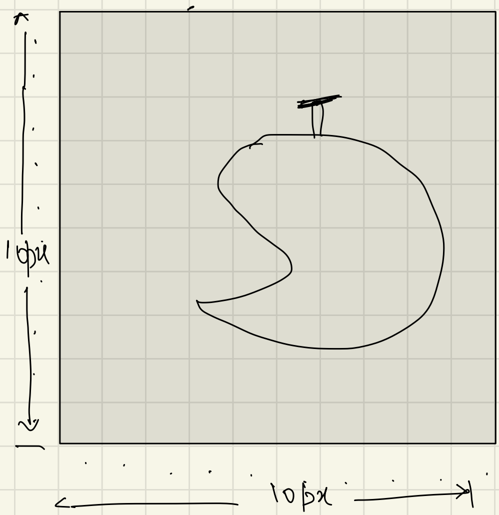
→ 200

b/w image

0 ↑ 200 255

↑ 100 ↑ ↑

white black



2D array = (10, 10)

0 or 1

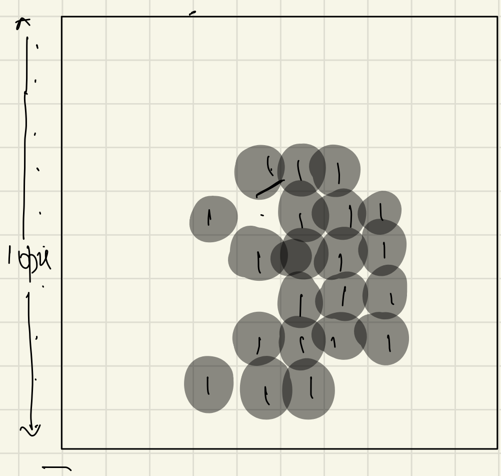
0 0 0 . . . 0 0

B/W

0 0 1 1 1 0 0

0 0 1 0 1 1

(0, 255)



0, 1

Bytes = 8 bits = $2^8 = 256$

Topics to Prepare for Final Exam:

① Regular Expression for Text Cleaning Processing.

things to learn:

strip — removing whitespace from left/right.

re.sub — Find and replace in text/string.

Inplace ?

② Pandas filtering of records using masking.

(dt)

③ Working with [datetimens64] datatype and accessor function

- diff()

④ Pandas differencing operator

⑤ Pandas Multiple Aggregator Groupings.