$E[X+Y] = E[X] + E[Y]$

even if X and Y are dependent

E.g. 5 card hand (# of aces)

$X_i = 1$ if i^{th} card is an ace

$E[X] = E[X_1 + \dots X_5]$

$= E[X_1] + E[X_2] + \dots E[X_5]$

$= 5 E[X_1]$

$= 5 \times \frac{1}{13} = \frac{5}{13}$

Geom(p) : independent Bern(p) trials

$(1-p)^{n-1} \times p$

PMF $(P(X=k)) = (1-p)^{k} \times p$

k failures before success

Scanned by CamScanner

$E[X] = \sum_{k=0}^{\infty} k p q^k$

$= p \sum_{k=0}^{\infty} k q^k$ — (1)

$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$.

$\sum_{k=0}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2}$.

or $\sum_{k=1}^{\infty} q^{k-1} \times k = \frac{1}{(1-q)^2}$.

or $\sum_{k=1}^{\infty} q^k \times K = \frac{q}{p^2}$. — (2)

Using (1) and (2), we get

$E[X] = p \times \frac{q}{p^2} = \frac{q}{p}$.

Proof of linearity.

let $T = X+Y$

proove $E(T) = E[X] + E[Y]$.

$\sum_{\forall t} t P(T=t) \overset{?}{=} \sum_{\forall x} x P(X=x) + \sum_{\forall y} y P(Y=y)$

$P(T=t) = \sum_{\forall x} P(T=t \mid X=x) \times P(X=x)$.

$=$ (stuck

$P(T=t) = \sum_{\forall s} (X+Y)(s)\, p\{s\}$

$= \sum_{\forall s} \big(X(s) + Y(s)\big)\, p\{s\}$

$= \sum_{\forall s} X(s)\, p\{s\} + \sum_{\forall s} y(s)\, p\{s\}$

$= E[X] + E[Y]$.

$r = 5$ (success)
$n = 11$ (failure)

1000100~~0~~10000100~~1~~

Negative binomial r, p

$$\binom{n+r-1}{r-1} p^r (1-p)^n = P(X=n)$$

Here X = # of failure before r success.

Random Variable → One particular choice.

Distribution → Maps for different choices

Poisson Distribution
(Important)

$P(X=k) = e^{-\lambda} \lambda^k / k!$

$k \geq 0; k \in \mathbb{I}$ (Integer)

λ is the scale parameter. $\lambda > 0$

$$PMF = \sum_{x \geq k} P(x=k) = 1$$

$$= \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

$$= e^{-\lambda} (e^{\lambda})$$ → from Taylor series

Taylor series of a function $f(x)$ at a is

$$f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(x-a)^2}{2!} + \ldots$$

if $f(x) = e^x$; $a=0$, then

$$\frac{1}{1} + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots = e^x$$

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$$

$$E[X] = \sum k \frac{e^{-\lambda} \lambda^k}{k!}$$

$$= e^{-\lambda} \sum_{k=1}^{\infty} \frac{k \lambda^k}{k!}$$

$$= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}$$

$$= e^{-\lambda} \times \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \qquad \boxed{\begin{array}{c} k-1 = j \\ j = 0 \\ \infty \end{array}}$$

$$= e^{-\lambda} \times \lambda \times e^{\lambda} = \lambda$$

This distribution is used to count # of "success" where each success has a small probability of success.

e.g. no. of emails

no. of insects in your food.

· no. of earthquakes.

no. of hacking attacks

Say we have j events

$A_1, A_2 \ldots A_n$ $P(A_j) = p_j$ and n is very large

of A_j's that occur is approximately $Pois(\lambda)$

$$\lambda = E[X] \quad X \sim Pois(\lambda)$$
$$= \sum p_j$$

Example

↳ Hospital 1.8 births/hour

1) $P(X=4) = \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-1.8}(1.8)^4}{4!} = 0.07$

2) $P(X \geq 2) = P(X=2) + P(X=3) \ldots$

$= 1 - P(X < 2)$

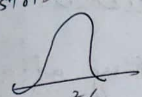
$= 1 - P(X=0) - P(X=1)$

$= 1 - \frac{e^{-1.8} \times (1.8)^0}{0!} - \frac{e^{-1.8} (1.8)}{1}$

$= 0.537$

Scanned by CamScanner

Normal Distribution.

$N(0,1)$ 

$$f(z) = c\, e^{-z^2/2}.$$

$$= \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

↳ Normalizing constant.

$$\int_{-\infty}^{\infty} e^{-z^2/2}\, dz = \int_{-\infty}^{\infty} f(z)\, dz = I.$$

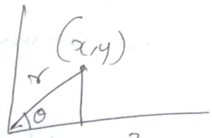
$$I^2 = \int_{-\infty}^{\infty} e^{-z^2/2}\, dz \int_{-\infty}^{\infty} e^{-z^2/2}\, dz$$

$$= \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \int_{-\infty}^{\infty} e^{-y^2/2}\, dy$$ (Notation change)

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2} e^{-y^2/2}\, dx\, dy.$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2/2)}\, dx\, dy.$$

$$= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} \times \underbrace{r}_{\text{Jacobian}}\, dr\, d\theta$$



$r^2 = x^2 + y^2$

$$= 2\pi$$

$$\Rightarrow I = \sqrt{2\pi}.$$

CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$

Easy to compute on computer systems.

Exponential Distribution.

→ One parameter λ. (Rate at which some events occur)

$X \sim Exp(\lambda)$ has PDF $\lambda e^{-\lambda x}, x > 0$.
otherwise 0.

PDF
$$P(X=x) = \lambda e^{-\lambda x}.$$

CDF
$$F(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}.$$

We know $\frac{dF(x)}{dx} = f(x)$.

One good thing

if $Y = \lambda X$

$Y \sim Exp(1)$

$$P(Y \le y) = P(\lambda X \le y) = P(X \le y/\lambda)$$
$$= 1 - e^{-\lambda \cdot y/\lambda}$$
$$= 1 - e^{-y}.$$
$$\Rightarrow \lambda = 1.$$

Say $Y \sim Exp(1)$.

$$E[Y] = \int_0^\infty y e^{-y} dy.$$
$$= y(-e^{-y})\Big|_0^\infty + \int_0^\infty e^{-y} dy.$$
$$= 0 + (-e^{-y})\Big|_0^\infty$$
$$= 0 + 1 = 1.$$

$$Var(Y) = E[Y^2] - (E[Y])^2$$
$$= \int_0^\infty y^2 e^{-y} dy - (1)^2$$
$$= y^2(-e^{-y})\Big|_0^\infty + 2\left[\int_0^\infty y e^{-y} dy\right] - 1.$$
$$= 0 + 2 - 1 = 1.$$

$E[X] = \frac{1}{\lambda}$

$Var(X) = \frac{1}{\lambda^2}$

since $X = \frac{Y}{\lambda}$
and $E[Y] = 1$
$Var[Y] = 1$

∴ using properties of Expectation and Variance, we get the result.

Memoryless property.

$P(X \geq t+s / \underbrace{X \geq s}_{\text{we have like waited } s \text{ minutes}}) = P(X \geq t)$.

$P(X \geq s) = 1 - P(X \leq s)$.
$= e^{-\lambda s}$

$P(X \geq t+s / X \geq s) = \dfrac{P(X \geq t+s, \underbrace{X \geq s}_{\text{redundant}})}{P(X \geq s)}$.

coz if we r waiting for $t+s$, we have already waited for s minute

$= \dfrac{e^{-\lambda(t+s)}}{e^{-\lambda s}}$

$= e^{-\lambda t}$

$= P(X \geq t)$

$\Rightarrow P(X \geq s+t / X \geq s) = P(X \geq t)$.

Standard Normal

$Z \sim N(0,1)$.

$-Z \sim N(0,1)$. (Symmetry).

$E[Z] = 0$

$E[Z^2] = 1$

$E[Z^k] = \begin{cases} 0 & k = \text{odd} \\ 1 & k = \text{even} \end{cases}$

let $X = \mu + \sigma Z$, $\mu \in \mathbb{R}$ (mean)
$\sigma > 0$ (S.D.) $\Rightarrow$ This is the scale.

then $X \sim N(\mu, \sigma^2)$.

$E[X] = E[\mu + \sigma Z]$
$= E\mu + \sigma E[Z]$
$= \mu$

$Var(\mu + \sigma Z) =$

Note

$Var[X+Y] \neq Var[X] + Var[Y]$.

if X, Y are independent

$Var[X+Y] = Var[X] + Var[Y]$.

Scanned by CamScanner

$Var(\mu+\sigma Z) = Var(\sigma Z)$
$= \sigma^2 Var[Z]$.

or $Z = \frac{X-\mu}{\sigma}$. [standardization].

Find PDF of $N(\mu, \sigma^2)$.

CDF: $P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le x\right)$.
$= P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right)$.
$= \Phi\left(\frac{x-\mu}{\sigma}\right)$.
$= F\left(\frac{x-\mu}{\sigma}\right)$

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

$f(x) = \frac{dF(x)}{dx}$.

$-X = -\mu + \sigma(-Z) \sim N(-\mu, \sigma^2)$.

---

if $X_j \sim N(\mu_j, \sigma_j^2)$ independent
then $X_1 + X_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$
$X_1 - X_2 \sim N(\mu_1-\mu_2, \sigma_1^2+\sigma_2^2)$

Rule of thumb => We cannot evaluate $F(x)$ for normal.

$P(|X-\mu| \le \sigma) \approx 0.68$
$P(|X-\mu| \le 2\sigma) \approx 0.95$
$P(|X-\mu| \le 3\sigma) \approx 0.97$.

$Var[X]$ if $X \sim pois(\lambda)$.

$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$

Diff. b/s

$\sum_{k=0}^{\infty} \frac{k\lambda^{k-1}}{k!} = e^{\lambda}$

Mult. b/s by λ

$\sum_{k=0}^{\infty} \frac{k \times \lambda^k}{k!} = \lambda e^{\lambda}$

Diff. b/s

$\sum_{k=1}^{\infty} \frac{k^2 \lambda^{k-1}}{k!} = \lambda e^{\lambda} + e^{\lambda}$
$= e^{\lambda}(\lambda+1)$.

$Var[X] = E[X^2] - (E[X])^2$

$= \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} - \lambda^2$

$= \cancel{\lambda^2} + \lambda - \cancel{\lambda^2}$

$= \lambda$.

$$\sum_{k=0}^{\infty} \frac{k^2 \lambda^k}{k!} = \lambda e^{\lambda}(\lambda+1)$$

$$\sum_{k=0}^{\infty} \frac{k^2 e^{-\lambda} \lambda^k}{k!} = \lambda e^{-\lambda} \times e^{\lambda}(\lambda+1)$$

$$\Rightarrow \sum_{k=0}^{\infty} \frac{k^2 e^{-\lambda} \lambda^k}{k!} = \lambda^2 + \lambda.$$

Strong Law of large numbers

Assume we have $X_1, X_2, X_3 \dots X_j$ iid samples with mean μ and Var σ^2.

$$X_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

SLLN says as $n \to \infty$

$X_n \to \mu$

with prob 1.

E.g. $X_j \sim \text{Bern}(p)$

$$\frac{X_1 + X_2 \dots X_n}{n} \to p \text{ with prob 1.}$$

E.g.

Application: Coin toss.

say biased coin (p) don't know p

If we toss the coin many times, we can even estimate p.

Weak law of large number (WLLN)

For any $c > 0$ $P(|\bar{X} - \mu| > c) \to 0$ as $n \to \infty$.

Proof: Using Chebyshev's inequality

$$P(|\bar{X}_n - \mu| > c) \le \frac{\text{Var}(\bar{X}_n)}{c^2}$$

$$= \frac{\frac{1}{n^2} n\sigma^2}{c^2}$$

$$= \frac{\sigma^2}{c^2 n}$$

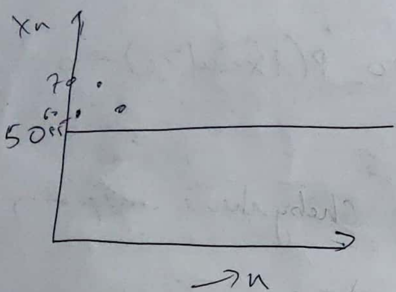
$$= 0 \quad n \to \infty.$$

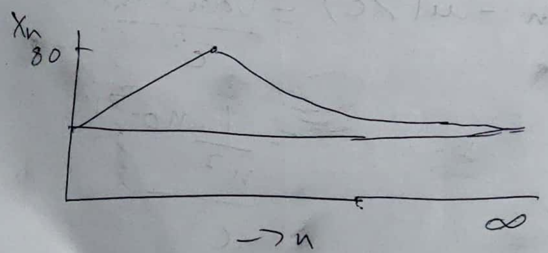
X = # of heads in 100 tosses

$$X_n = \frac{60 + 70 + 55 + \dots X_n}{n}$$

as $n \to \infty$ $X_n \to 50$.

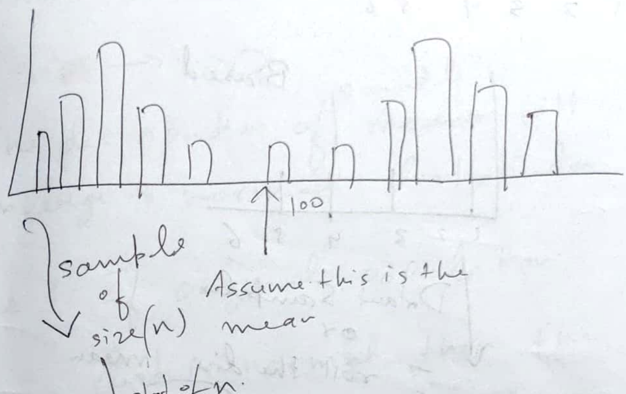
$E[X] \equiv 50$

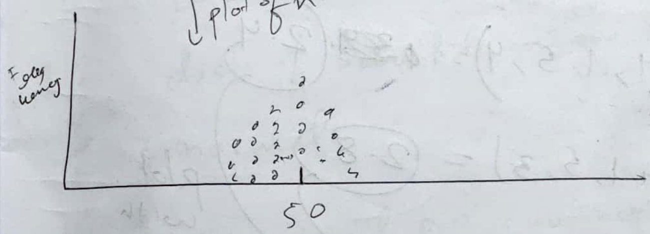


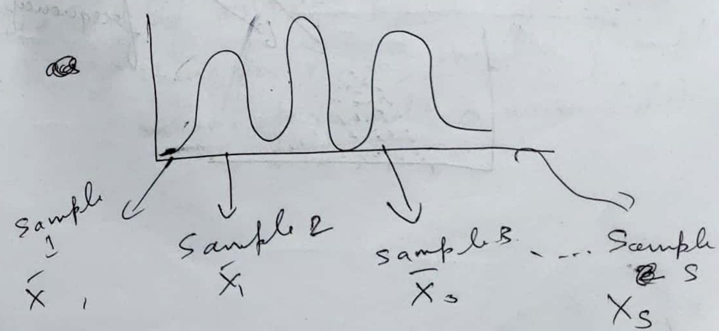


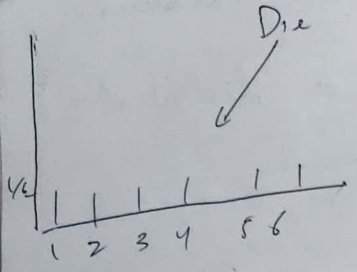
Central limit theorem.

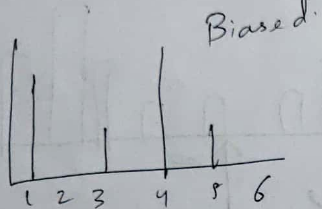
Say we have a distribution of data as







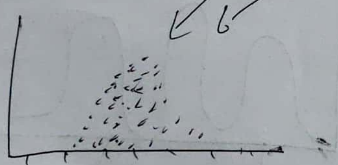




Draw Samples
or
I roll the die mean

(1, 1, 1, 5, 4) = ~~3~~ 2.4
⋮
(1, 4, 1, 5, 3) = 2.8
⋮

plot
with
frequency



As the number of samples increase
the plot will approximate the
↳ frequency
normal distribution.

$S \to \infty$
the distribution of sample means will
converge to normal distribution.

Note. I: If original population
dist is normal, then the
dist. of sample mean is
normal.

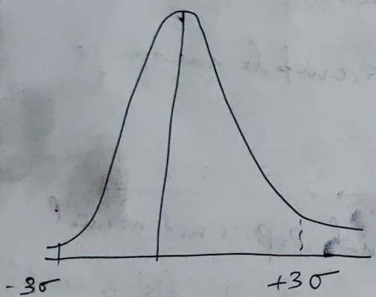
Note 2: If original pop is not normal,
then for $n > 30$, the dist. of sample
means approximate a normal dist.

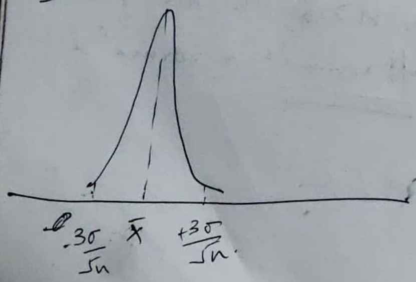
$\bar{X}_1, \bar{X}_2, \bar{X}_3 \dots \bar{X}_{z}$

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 \dots \bar{X}_z}{n}$$

$\bar{x} \approx \mu$.

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ → sample size.



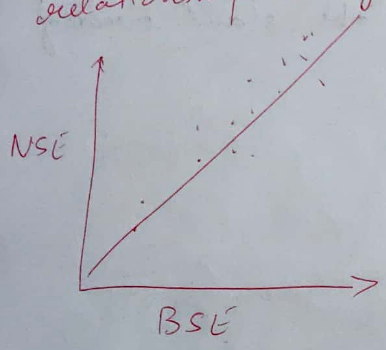


$\frac{\sigma}{\sqrt{n}}$ is the std. dev of the samples distribution of the sample ~~the~~ mean

Also called as the stand. error of means.

Covariance and correlation

→ They form a part of bivariate relationship analysis.



→ Look at the data points
↳ pattern looks like linear relationship

→ When one rises, other rises in the same way.

→ In this case, relationship is strong.

In other words, there is a strong positive relationship.

We can quantify this relationship via covariance i.e how they co-vary.

Note: Covariance is only one measure to quantify this idea.

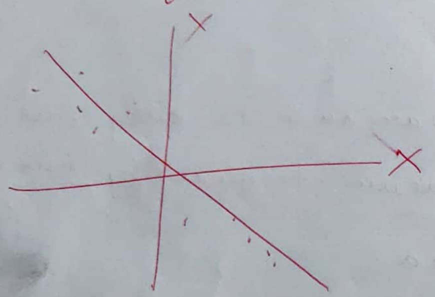
Others include: correlation

Covariance: A descriptive measure of the relationship b/w two variables.

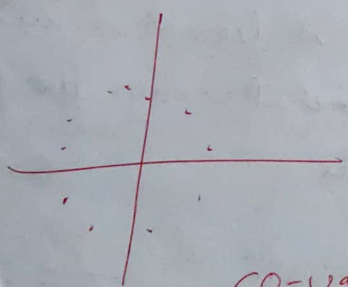
Emphasis is on direction.

+ve indicates increasing R/P
-ve " decreasing R/P

Note Covariance is used to identify linear relationship



-ve covariance



CO-variance = 0

$$\mathrm{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Co-variance:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

for sample

$$\sigma_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \cancel{\mu_x})(y_i - \mu_y)}{n}$$

For population

Example: x : workers — 10 random samples from new building

y : no. of walls painted

$\bar{x} = 21.3$ $\bar{y} = 41.2$

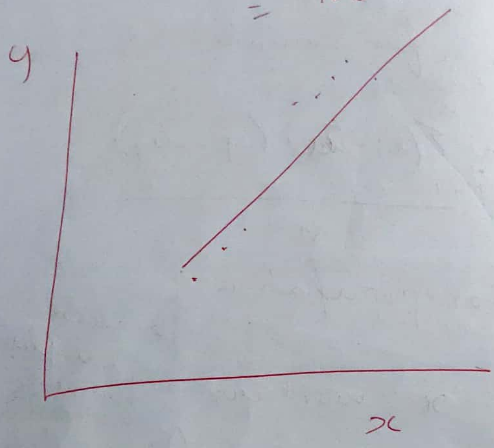
$x\ (s=6.48)$	$y\ (s=16.69)$	$x_i - \bar{x}$	$y - \bar{y}$	$\sum(x_i-\bar{x})(y_i-\bar{y})$
12	20	-9.3	-21.2	197.16
30	60	8.7	18.8	
15	27	-6.3	-14.2	
24	50	2.7	8.8	
14	21	-7.3	-20.2	
18	30	-3.3	-11.2	
28	61	6.7	19.8	
26	54	4.7	12.8	
19	32	-2.3	-9.2	
27	57	5.7	15.8	

$\Sigma = 962.4$

Scanned by CamScanner

$$Cov(x,y) = S_{xy} = \frac{962.4}{10-1}$$

$$= 106.93$$



Positive linear relationship.

Focus only on sign

Correlation

Covariance provide the direction

Correlation provide direction and strength.

Covariance → No upper and lower bound.

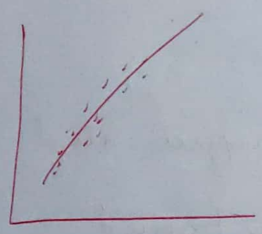
Correlation → Between (-1, 1).

Covariance is dependent on scale

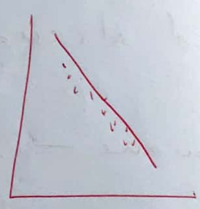
Correlation is independent of scale.

Scanned by CamScanner

Drawback: Covariance and correlation are applicable to linear relationships.



Near +1



Close to -1

Close to 0.

Correlation Coefficient
(r)

$$r_{xy} \text{ (or } \rho_{xy}) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

$$cov(X,Y) = E[(X-\mu)(Y-\mu)]$$

From wikipedia -

If (x_i, y_i) (for $i = 1$ to n) can be realized with probability p_i, then

$$cov(X,Y) = \sum_{i=1}^{n} p_i \left(x_i - E[X]\right)\left(y_i - E[Y]\right)$$

$$cov(X,a) = 0$$

$$cov(X,X) = var(X)$$

$$cov(X,Y) = cov(Y,X)$$

$$cov(aX,bY) = ab\, cov(X,Y)$$

$$cov(X,Y) = E[XY] - E[X]E[Y]$$

$$\rho_{xy} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

$$= \frac{106.93}{6.48 \times 16.69} = 0.989$$

Rule of thumb

if $|r| \geq \frac{2}{\sqrt{n}}$

then relation exists

Not Universal

Important point:

Correlation does not imply causation.

$X = x_1 \; x_2 \; x_3$

$cov(X) \rightarrow$ covariance matrix

	X_1	X_2	X_3
X_1	$var(X_1)$	$cov(X_1, X_2)$	$cov(X_1, X_3)$
X_2	$cov(X_2, X_1)$	$Var(X_1)$	$cov(X_2, X_3)$
X_3	$cov(X_3, X_1)$	$cov(X_3, X_2)$	$Var(X_1)$

Suppose we have a random variable $X \sim N(\mu, \sigma^2)$.

To convert this normal distribution into standard normal, we follow the standardizing procedure. i.e.

$$Z = \frac{X-\mu}{\sigma}$$

Here $Z \sim N(0,1)$.

E.g. $X \sim N(162.2, \sigma = 6.8)$

Say what is the prob. that an average girl is taller than 170.5

$P(X > 170.5)$

$$P\left(\frac{X-162.5}{6.8} > \frac{170-162.5}{6.8}\right)$$

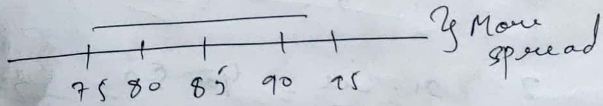
or, $P(Z > 1.22)$.

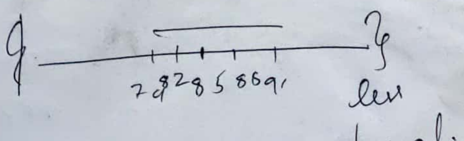
Discussion on reading z-table

z-score.

Suppose, we have the following data

85	88
95	79
75	91
80	85
90	82
$\mu = 85$	$\mu = 85$





Z-score is a statistical measure that describes how far a given data point is from the mean.
↳ It describes the distance in terms of standard deviation.

Q) Suppose a data sample has mean 50 and σ of 5. A value of 55 will have z-score of 1. and 40 will have z-score of -2

$$z = \frac{x-\mu}{\sigma}$$

$$= \frac{55-50}{5} = 1$$

$$z = \frac{x-\mu}{\sigma} = \frac{40-50}{5} = -2.$$

Q) Suppose in a test a person scored ~~500~~ 630 marks. Mean marks is 500, sd is 150.

$$z = \frac{x-\mu}{\sigma} = 630$$

$$= \frac{630-500}{150} = 0.83$$

Q) ~~US~~ ICSE board a person ~~had~~ got 87% with others having an average 80% with S.D. of 5.

CBSE board a person got 82% with class avg. 73 and s.d. of 6

$$z_1 = \frac{87-80}{5} \quad ; \quad z_2 = \frac{82-73}{6}$$

$$z_1 = 1.4 \quad ; \quad z_2 = 1.5$$

Scanned by CamScanner

Suppose IQ of people follow normal distribution with $\mu = 100$ and $\sigma = 15$

What % of people are i) stupid
ii) smart
iii) average

stupid: 50 or less

smart: 140 or more

avg: 50-140

Stupid $= \dfrac{50-100}{15} = -3.333$

~~.005~~ % 0.004

= ~~0.5~~ % 0.4%

smart: $\dfrac{140-100}{15} = \dfrac{40}{15} = 2.66$

$= .9961$

$1 - 0.9961$

.0039

0.4%

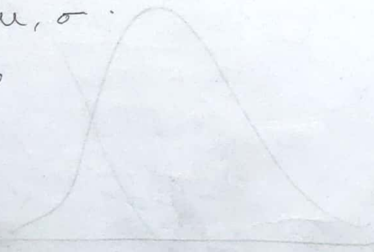
avg = 99.2%

Inferential Statistics: We draw statistical inference when complete information is not available.

Population μ, σ

sample

$\bar{x}, s$



Goal is maintain 32 points all time time (for students)

Student	Points
1	31.93
2	31.24
3	31.53
4	31.45
5	30.93
6	34.66
7	33.55
8	29.79
9	31.82
10	32.27
11	32.56
12	33.32
13	32.04
14	32.82
15	31.70

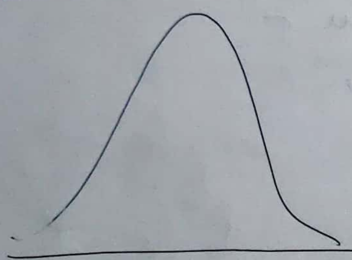
Scanned by CamScanner

$\bar{x} = 32.10$

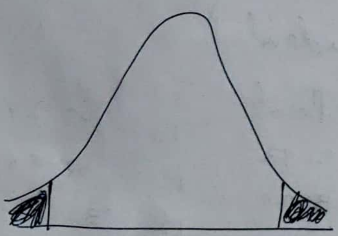
$s = 1.17$

Confidence Interval:

Point estimate $\pm$ margin of error



sampling distribution of mean



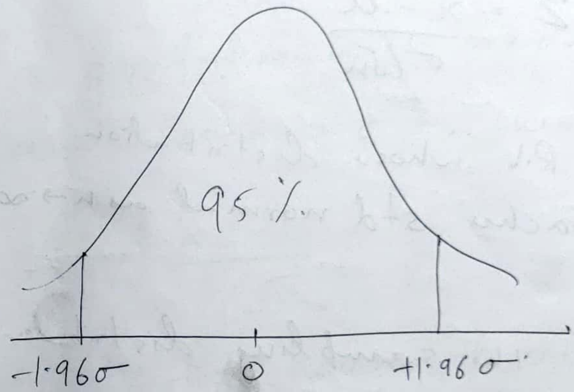
95% probability interval

Point estimate is a single number calculated from sample data for which we have some assurance that it is close to the parameter it is supposed to estimate.

95% of all sample means will be in the unshaded region

α - area in the tails is called as alpha.

α - area of shaded region.



We know standard deviation of the population is $\frac{\sigma}{\sqrt{n}}$ $(\sigma_{\bar{x}})$

So the interval is

$\bar{x} \pm 1.96\,\sigma_{\bar{x}}$. $\left(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$.

Scanned by CamScanner

Sampling Distribution.

1) If $\bar{x}$ is the mean of a random sample of size n taken from a population having mean μ and variance σ^2, then

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

is a R.V. whose distribution approaches std normal as $n \to \infty$.

This is the sampling distribution of the mean with known σ

2) Sampling distribution of mean with unknown σ.

If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and variance σ^2, then

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

is the R.V. having t distribution with parameter $V = n - 1$.

here $s = \text{var}(\text{samples})$

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

Scanned by CamScanner

3) Sampling distribution of variance.

If S^2 is the variance of a random sample of size n taken from normal population with variance σ^2, then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

It has chi-square distribution with $v = n-1$.

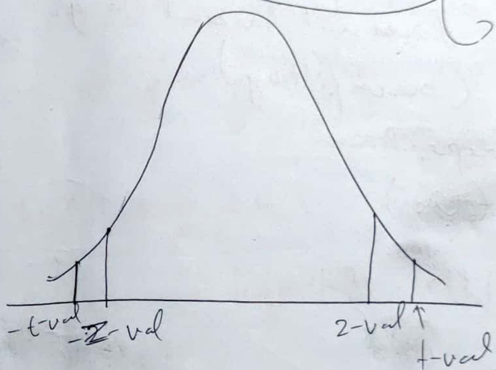
$$\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$ = t-distribution

$$\frac{x^{v/2-1}\, e^{-x/2}}{2^{v/2}\,\Gamma\left(\frac{v}{2}\right)},$$ $x > 0$; 0 otherwise

Confidence interval estimation of the variance of the population is not known. Margin of error

$$\bar{x} \pm t\text{-value} \times \frac{s}{\sqrt{n}}$$

std. error.



At 95% confidence
z-value = ±1.96
t-value

16	9	± 2.262
30	29	± 2.045
75	74	± 1.993
100	99	± 1.984

Scanned by CamScanner

Hypothesis Testing:

Example: Railway IRCTC Neer bottled water 1ltr.

IRCTC does not want to upset its customer (underfilling)

IRCTC does not want to go bankrupt (overfilling).

Two perspectives

Customer (Us).

Vol $\geq$ 1ltr

IRCTC.

Vol = 1ltr

Us: Are we getting (on an average) atleast 1ltr of water.

IRCTC: Are we supplying exactly 1ltr of water

∴ lets ~~the~~ test this using hypothesis

Definition: A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

Null hypothesis

Alternate hypothesis.

By definition Null and Alternate hypothesis are opposite to each other.

We can start either with null or alternate hypothesis.

Scanned by CamScanner

Null
↳ Accepted as true (so test it)
↳ Assumed one.

Alternate
↳ Unknown as true
↳ It is a claim.

Null : H_0

Alternate : H_a.

Logically speaking (common sense)

$H_0 =$	$H_0 \leq$	$H_0 \geq$
$H_a \neq$	$H_a >$	$H_a <$

Some points

We do not reject the null hypothesis

We say the data supports the alternate hypothesis

Examples : IRCTC example

Null $H_0 =$ 1 ltr

$H_a \neq$ 1 ltr

Errors

	Null Hypothesis	
	True	False
Fail of reject	True positive ✓	False Negative × Type II error
Reject	False positive × Type I error	True Negative ✓

Scanned by CamScanner

Actual Reality

Decision ↓	$\mu = 1\,ltr$	$\mu \neq 1\,ltr$
Don't reject	correct	Type II error
Reject	Type I error	correct

$H_0: \mu = 1\,ltr$ $\quad H_a: \mu \neq 1\,ltr$

Say, we take 100 bottles

$\bar{X} \approx 0.999\,ltr$

Don't reject

Sample 2: 200 bottles
(ridiculous sample)

$\bar{X} = 900\,ml$

reject the null hypothesis

Say, in case 2, we get a weird sample.

$\bar{X} = 899\,ml$

You will reject the null.

However, if this sample is flawed you rejected the null. The reality however is opposite. Bottles are filled correctly.

$\Downarrow$

Classic Type I error.

$\Downarrow$

Sample had the problem.

Scanned by CamScanner

Say the average IQ of a population is 100 with std deviation of 15. We now believe the value has changed. We took a sample of 75 people and avg IQ was 105. Do you think the IQ has changed?

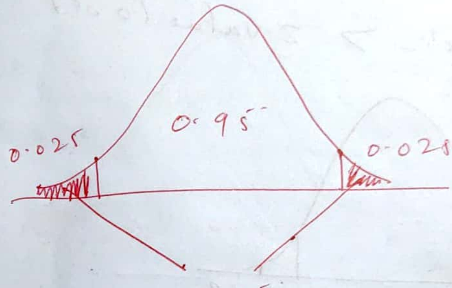
Steps:

1) Null and Alternate hypothesis
2) Choose significance level (α)
3) Find critical values
4) Find test statistic
5) Conclusion

$H_0: \mu = 100$

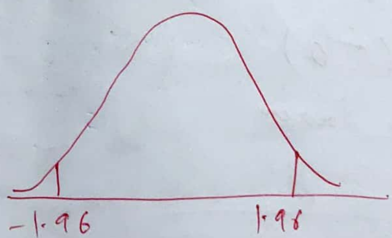
$H_a: \mu \neq 100$

It is a two tailed test.



$\sigma = 15$

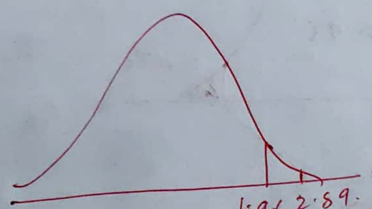
Z-value(0.05) = 1.96.



Test statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{105 - 100}{15/\sqrt{75}} = 2.89.$$

Test statistic > z-value (0.05)

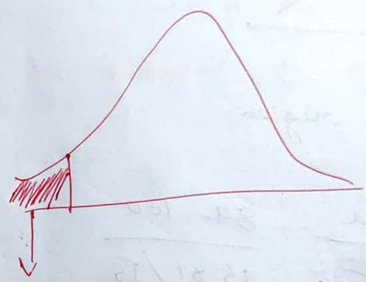


⇒ Conclusion : The data supports the alternate hypothesis

Q2. $\mu = 100$ (IQ)
IQ is now lower
Sample IQ
69, 79, 89, 99, 109.
s.d = 15.81.

1) $H_0 : \mu = 100$ (1 tailed test)
$H_a : \mu < 100$

2) $\alpha = 0.05$ (suppose)

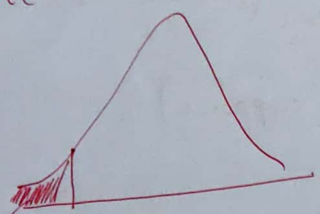


We have chosen this side as we are testing for lower values. i.e lower than the mean.

3) This time we don't know population mean
⇒ We will use t-distribution thus the t-test.

Scanned by CamScanner

$t(0.05, 4) = 2.132$

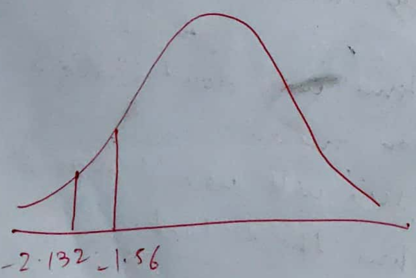


-2.132

→ rejection region.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{89 - 100}{15.81/\sqrt{5}}$$

$$= -1.56$$



~~The~~

Conclusion: The data supports the null hypothesis.

Say we took a survey of one of the subjects in 2^{nd} year.

1 - 2 - - - 5

poor excellent

previous study said

$\mu = 3$

$\sigma = 1.5$

say we want to test the teachers capability now.

Sample size = 225

$\bar{x} = 3.25$

1) $H_0: \mu \leq 3$

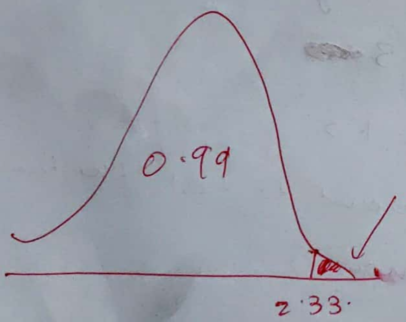
$H_a: \mu > 3$

(We are interested in better than average making of leaders)

2) Confidence level now 99%

$\Rightarrow \alpha = 0.01$

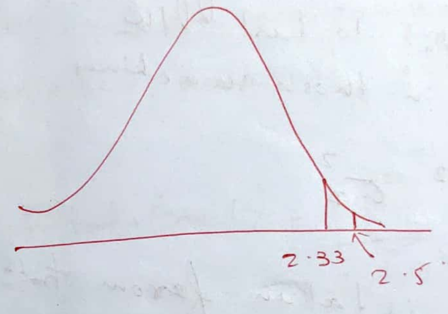
$z(\alpha) = 2.33$.



3) Test - statistic

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{3.25 - 3}{1.5/\sqrt{225}}$$

$$= 2.5$$



test statistic is in the rejection region.

$\Rightarrow$ The data supports the alternate hypothesis.

Scanned by CamScanner

Analysis of variance

F-test

Say we have a bottle filling machine (5ltr).

Say, we have two different machines and we want to test the variation in those machines

i.e. $\sigma_1^2 = \sigma_2^2$

A sample is taken from both the machines

S_1

$n = 25$, $\bar{x} = 5.0392$; $s^2 = 0.1130$, $s = 0.338$

$n = 22$; $\bar{x} = 4.9808$ $s^2 = 0.0137$, $s = 0.1171$

Clearly, there is diff is std dev.

However we are dealing with samples, hence we have to test them.

Here, we are comparing two sample variances, hence we use F-test

$$\text{F-ratio (or F)} = \frac{s_1^2}{s_2^2}$$

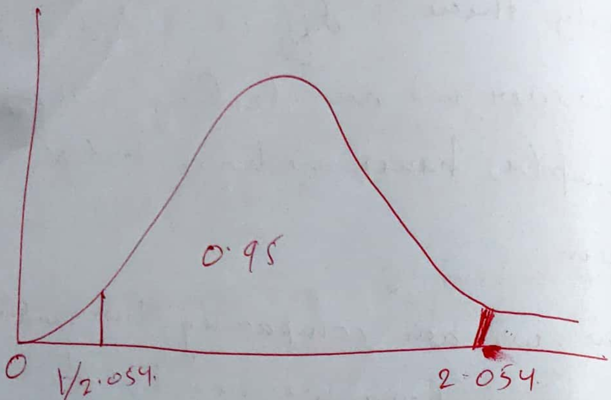
S_1 = larger sample variance

S_2 = smaller "

N_r has $n_1 - 1$ degree of freedom

D_r " $n_2 - 1$ " " "

$df_1 = 24$; $df_2 = 21$



$Ho: \sigma_1^2 = \sigma_2^2$

$Ha: \sigma_1^2 \neq \sigma_2^2$

$F = \frac{0.1130}{0.0137} = 8.248$

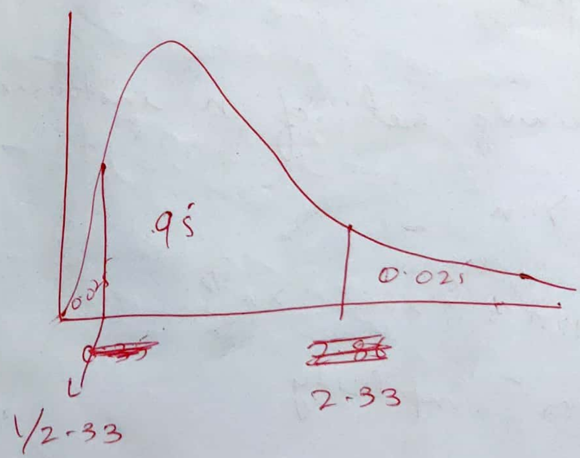
One more example

A	B	
21	16	n
2.5	1.8	σ
100	95	$\bar{x}$

$Ho = \sigma_1^2 = \sigma_2^2$

$Ha = \sigma_1^2 \neq \sigma_2^2$

Two tailed



Scanned by CamScanner

$F = \frac{6.25}{3.24} = 1.92$

F falls in the acceptance region.

Linear Regression.

→ Supervised learning

→ Assumption: There is a linear relationship between input & output.

Training set of n instances

$X_{1:n}, Y_{1:n}$

$X_{1:n} = \{x_1 \dots x_n\}$.

For simplicity

$x_i \in R$; $y_i \in R$.

Wind	People inside BU	Energy
		1
100	2	5
50	42	25
45	31	22
60	35	18

In general

$x_i \in R^d$

1 Training

Data

$[X_{1:n}, Y_{1:n}] \longrightarrow \boxed{\text{Learner}}$

↓

Model parameters

(θ)

Scanned by CamScanner

(2) Testing

Prediction

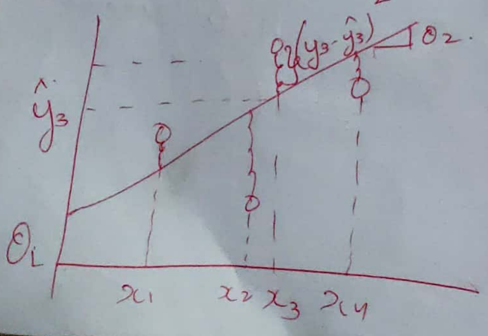
$x_{n+1}, \theta \longrightarrow$ Prediction Engine $\longrightarrow \hat{Y}_{n+1}$

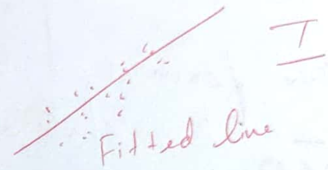
For simplicity.

$$y(x_i) = \theta_1 + x_i\theta_2$$

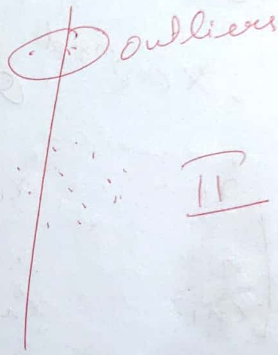
$$\underbrace{J(\theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Objective function}}$$

$$= \sum_{i=1}^{n}(y_i - \theta_1 - x_i\theta_2)^2$$





However, outliers



The outliers in case II destroyed the fitted line

Multiple Regression (generalized)

$$\hat{y}_i = \sum_{j=1}^{d} x_{ij}\theta_j$$

$$= x_{i1}\theta_1 + x_{i2}\theta_2 + x_{i3}\theta_3 \cdots$$

$$\cdots \theta_{d+1}$$

Scanned by CamScanner

θ_{d+1} is the bias

In matrix form.

$$\hat{y} = \bar{X}\theta$$

$$\hat{y} = R^{n\times 1} \;;\; X \in R^{n\times d} \quad \theta \in R^{d\times 1}$$

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

For the BU example

$$y = \begin{bmatrix} 5 \\ 25 \\ 22 \\ 18 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix}$$

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

Hypothetically,

$$\theta = \begin{bmatrix} 1 & 0 & 0.5 \end{bmatrix}^T$$

$$\Rightarrow Y = X\theta$$

$$= \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 22 \\ 16.5 \\ 18.5 \end{bmatrix}$$

Say $x = [50 \quad 20]$

$$\hat{y}(x) = [1 \quad 50 \quad 20] \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}$$

$$= 1 + 0 + 10 = 11$$

Find the parameters.

$$J(\theta) = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Partially differentiate $J(\theta)$ w.r.t. ~~b~~ all parameters

$$\frac{\partial J(\theta)}{\partial \theta_1} = g_1(\cdot)$$

$$\frac{\partial J(\theta)}{\partial \theta_2} = g_2(\cdot).$$

You will normal equations.

You then solve the normal equations.

Using matrix notation.

$$J(\theta) = (Y - X\theta)^T (Y - X\theta).$$

$$\frac{\partial J(\theta)}{\partial \theta} = 0$$

$$\Rightarrow \theta = (X^T X)^{-1} X^T Y.$$

Scanned by CamScanner

Polynomial regression.

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \theta_3 x_i^3 \dots$$

$$\dots \theta_m x_i^m$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} X$$

$$[\theta_0 \;\; \theta_1 \;\; \theta_2 \;\dots\; \theta_m]^T.$$

$$Y = \theta^T X$$

$$\hat{\theta} = (X^T X)^{-1} X^T Y.$$

$$m < n.$$

Maximum likelihood:

Suppose $X \sim N(\mu, \sigma^2)$

$$\| \; \mu + N(0, \sigma^2).$$

Suppose we have 3 data points

$y_1 = 1$

$y_2 = 0.5$

$y_3 = 1.5$

Assume, all y_i's are iid and follow gaussian distribution.

$y \sim N(\theta, 1) = \theta + N(0, 1)$.

Find $\hat{\theta}$

Likelihood : It is a function of the parameters of a statistical model given some data.

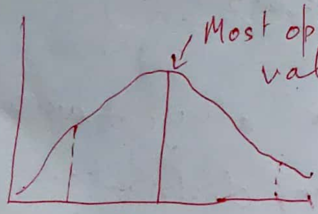
Let $X_1, X_2 \ldots X_n$ be random sample from PMF/PDF

$f(x|\theta)$.

and all X_i's are i.i.d.

Likelihood function

$L(\theta | x_1 \ldots x_n)$

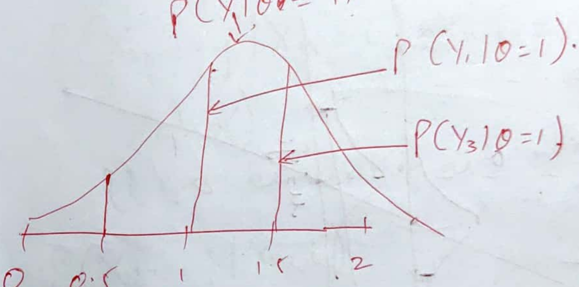


$L(\theta | x_1 \ldots x_n) = \prod_{i=1}^{n} f(x_i|\theta)$.

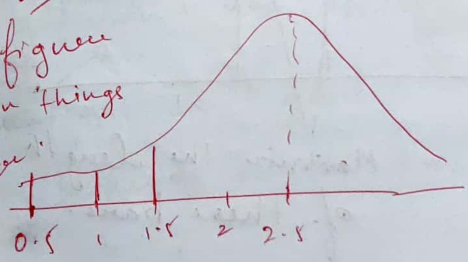
Going back to the example

$y_1 = 1 \quad y_2 = 0.5 \; ; \; y_3 = 1.5$

Guess $\theta = 1, 2.5$



This figure explain things better.

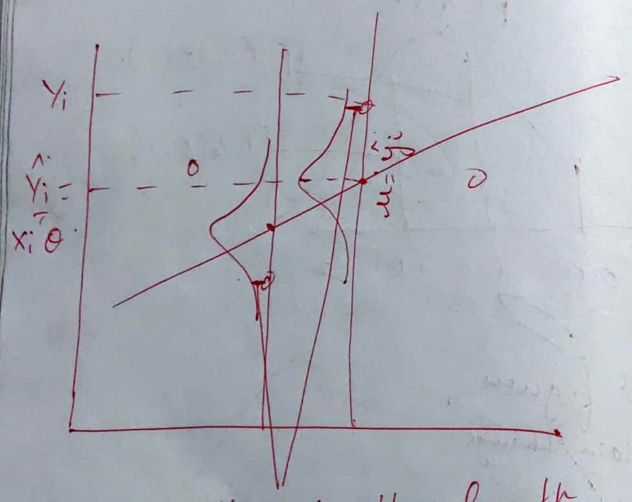


Assumption linear regression

$$y_i \sim N(x^T\theta, \sigma^2)$$

$$= x^T\theta + N(0, \sigma^2)$$

$$\cancel{p(Y|X,\theta,\sigma) = \prod_{i=1}^{n} p(y_i|x_i,\theta,\sigma)}$$



Maximize the length of these bars.

Markov Chains

$X_0 \dots X_n$: state of a system.

Suppose X_n is now, we want to predict X_{n+1}

$$P(X_{n+1}=j \mid X_n = i, X_{n-1}=z, X_{n-2}=a, \dots X_0=\alpha)$$

|| can be written as

$$P(X_{n+1}=j \mid X_n=i)$$

All old values are obsolete

Assumption: Is it good?

Depends on the problem.

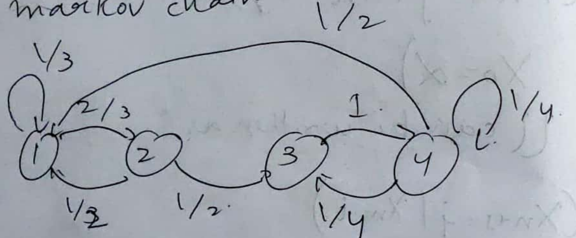
Scanned by CamScanner

$$\underbrace{P(X_{n+1}=j \mid X_n = i) = q_{ij}}$$

⇓

Transition probability.

If q_{ij} is constant with time, the chain is called as homogeneous markov chain.



TM =

$$\begin{array}{c|cccc} Q & 1 & 2 & 3 & 4 \\ \hline 1 & 1/3 & 2/3 & 0 & 0 \\ 2 & 1/2 & 0 & 1/2 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & 1/2 & 0 & 1/4 & 1/4 \end{array}$$

Note each row of Q sum to 1.

$q_{ij} \geq 0$.

$\forall j \sum q_{ij} = 1$.

First order Markov chain → Depends upon the current state.

⇓

Generalize to n^{th} order.

$$P(X_i \mid X_{i-1}, X_{i-2} \dots X_0) = P(X_i \mid X_{i-1} \dots X_{i-n})$$

Note: There is a debate on first order markov chain.

⇓

Applicable in real world?

Question: What's the probability ~~from~~ of going i to j in 2 steps?
or 1 step
or n steps.

We use Q to answer these questions

Current time : n.
" states: X_n.
Total states : m.

~~Xn~~ Distribution of $X_n = S$

$$S = \begin{array}{ccc} X_1 & \cdots & X_m \\ \left[\; \right. & & \left. \; \right] \end{array}$$

$$\begin{array}{ccc} X_1 & & X_m \\ \left[\; \right. & & \left. \; \right] \end{array}$$

$P(X_{n+1} = j) = ?$

Using LTP,

$$P(X_{n+1} = j) = \sum_i P(X_{n+1} = j \mid X_n = i)\, P(X_n = i)$$

$$= \sum_i q_{ij}\, s_i$$

$$= \sum_i s_i\, q_{ij}$$

$SQ \leftarrow$ i,j^{th} entry in SQ.

Or

~~for~~ $\forall j\ P(X_{n+1} = j) = SQ$.

Generalizing,

SQ^2 is the distribution at time $n+2$

SQ^k is the distribution at time $n+k$.

Scanned by CamScanner

Similarly,

obselete

$P(X_{n+1}=j \mid X_n=i) = q_{ij}$

$P(X_{n+2}=j \mid X_n=i) = \sum_k P(X_{n+2}=j \mid X_{n+1}=k, X_n=i)$

$\times P(X_{n+1}=k \mid X_n=i)$

$= \sum_k q_{kj} \times q_{ik}$

$\downarrow$ i,j entry in Q^2.

To generalize,

$P(X_{n+m}=j \mid X_n=i) \rightarrow i,j^{th}$ entry in Q^m.

Stationarity : or stationary distribution

$\Downarrow$

Idea

$\Downarrow$

You run something for a long time. Will the chain converge?

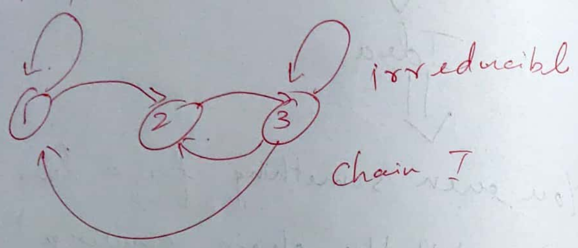
If $\underbrace{SQ}_{\text{Distribution 1 step ahead in time}} = S$, then S is stationary.

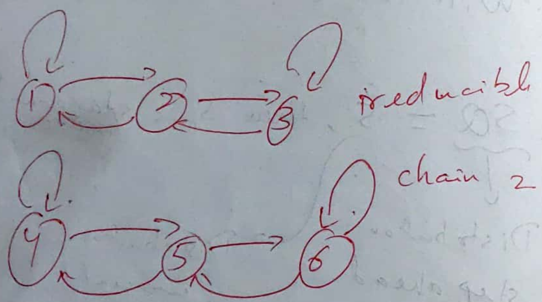
$S \rightarrow$ Distribution current!!

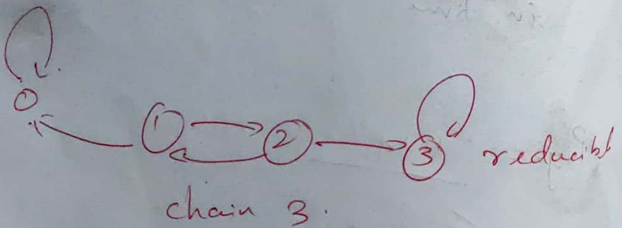
$$P_{ij} = \frac{n_{ij}}{\sum_{\forall k} n_{ik}}$$

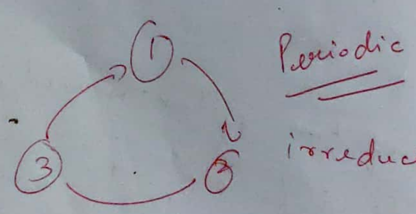
n_{ij} = no of time system moved from state "i to state j"

Markov chains: Similar to memoryless property.





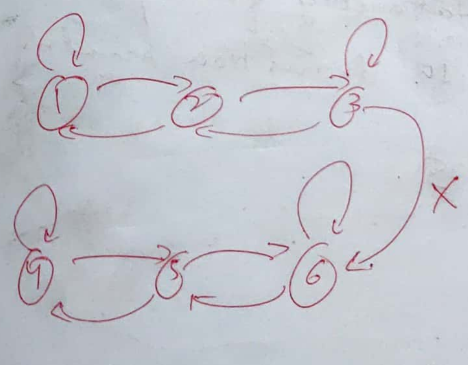




Chain is irreducible if you can get to any state from any state in finite steps with +ve probability.

A state is recurrent if ~~if~~ you start from a particular, then with ~~for~~ prob. 1, we will revisit that state.

Otherwise, the state is transient.

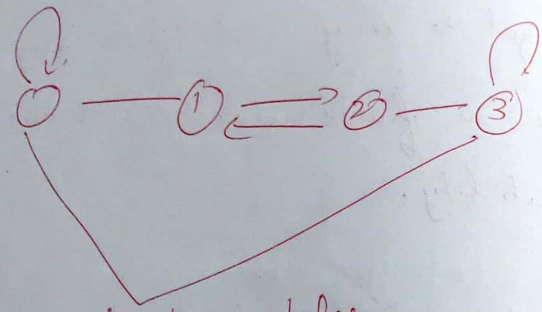


In the figure, 1,2,3 are recurrent if the system does not take X.

If X is taken, 1,2,3 will become

Scanned by CamScanner

~~[illegible]~~ transient.



absorbing states

This example is similar to two people gambling.

s is stationary for a particular chain with transition matrix Q if $sQ = s$.

Note I:

→ For any irreducible Markov chain with k states, there is always a stationary distribution.

→ The stationary distribution is unique.

→ $s_i = \frac{1}{r_i}$

where, r_i is the avg time to return i starting from i.

$P(X_n = i) \to s_i$

as $n \to \infty$

and

$tQ^n \to s$ — It can be randomly chosen distribution.

as $n \to \infty$

Scanned by CamScanner

Reversible Markov chains (w.r.t. time)

Markov chain with TM Q is reversible, if there is a S such that

$$s_i q_{ij} = s_j q_{ji} \quad \text{— (A)}$$

$\forall i, j$

If (A) holds, then S is stationary.

It is like playing a video. You play it forward or backward it gives ~~the~~ only one kind of picture.

Proof

Let $s_i q_{ij} = s_j q_{ji}$ $\forall i, j$

$$\sum_{\forall i} s_i q_{ij} = \sum_i s_j q_{ji}$$

$$= s_j \sum_{\forall i} q_{ji} \; - 1.$$

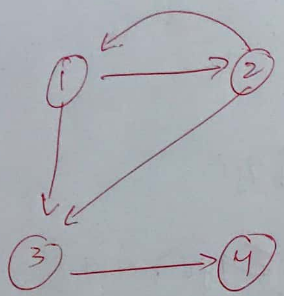
$$= s_j$$

$$SQ = S$$

$\Rightarrow$ Reversibility implies stationarity.

Scanned by CamScanner

Google Pagerank



we use the same idea

$$s_j = \sum_{\forall i} s_i q_{ij}$$

$$Q = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1/2 & 1/2 & 0 \\ 2 & 1/2 & 0 & 1/2 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & 1/4 & 1/4 & 1/4 & 1/4 \end{array}$$

$$G = \alpha Q + (1-\alpha) \frac{J}{M} \quad \text{— (A)}$$

M = no. of pages

$$J = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & & & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

Eq. (A) says that with prob α you follow the chain and $(1-\alpha)$ you randomly select any other page.

$\alpha = 0.85$

Scanned by CamScanner

Eqn. (A) guarantees reducibility.

No zeros in TM.

$tG, tG^2 \ldots tG^n$ → some random distribution

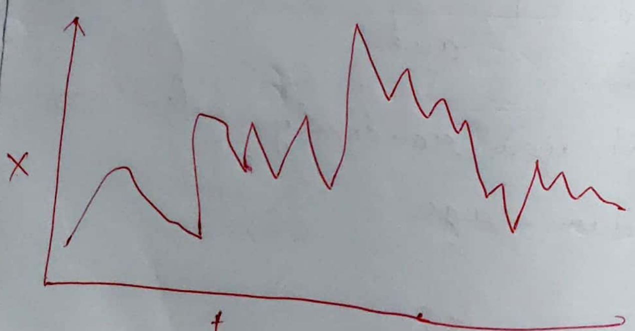
$\underbrace{tG^n}_{\text{Stationary distribution achieved.}}$

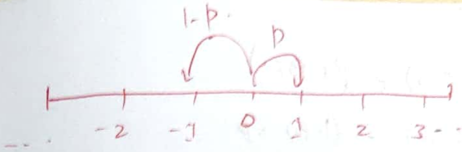
**

Random Walk

↓

Describes a sequence of successive random steps in some mathematical space.





P for $i = 0, \pm 1, \pm 2 \ldots$

and for number $p \in [0,1]$,

the transition probabilities are given by:

$P_{i,i+1} = p$

$P_{i,i-1} = 1-p$

Suppose a person starts at 0.

i) Calculate the prob. the person ends at same position in two steps.

$A = \{\overset{\text{Forward}}{F}\underset{}{\overset{\text{Backward}}{B}},\ BF\}$

$$P(A) = P(FB) + P(BF)$$
$$= p \times (1-p) + (1-p) \times p$$
$$= 2p(1-p).$$

2) Calculate the prob that the person ends one step forward in exactly three steps

$$A = \{FFB, BFF, FBF\}$$

$$\Rightarrow P(A) = 3p^2(1-p)$$

** $$tG = \alpha \underbrace{tQ}_{\text{Sparse matrix}} + (1-\alpha)\underbrace{(tI)}_{M}.$$

It is always $I^{1\times m}$

Scanned by CamScanner