Prob & Stats

Maths for ML.

$$\text{Probability } (P) = \frac{\text{Total possible outcomes}}{\text{Total population / sample space}}$$

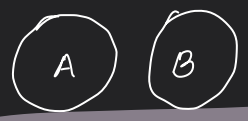
Prob of not $\leftarrow P(A') = 1 - P(A)$

Disjoint Events :

eg. kids can play only one sport.

so.

$$P(A \cup B) = P(A) + P(B)$$



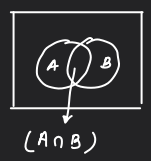
no intersection

Mutually Exclusive

Joint Events :

eg. kids can play as many sports as they want.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Inclusion Exclusion Principle

Independent Events: when two prob. / events are independent of each other.

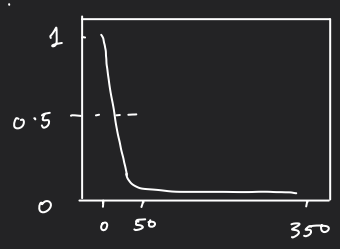
$$P(A \cap B) = P(A) \cdot P(B)$$ (Product Rule)

Birthday problem

From a group of x people, prob. that no two people have the same b'day.

1. 2. 3. 4. ...

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdot \ldots$$



and for
1 person : 1
3 people : 0.992
5 people : 0.973
10 people : 0.883
30 people : 0.294
50 people : 0.030
100 people : 0.0000003

Q. What is the prob. of landing on heads twice GIVEN that the first one is head.

$$P(HH \mid 1st \text{ is } H) = \frac{1}{2}$$

(HH

(HH and HT)

This is **Conditional Probability**

$$P(A \cap B) = P(A) \cdot P(B \mid A)$$

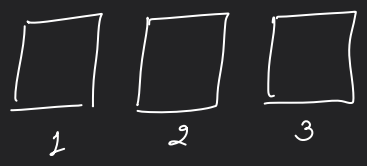
when independent

$P(B \mid A) = P(B)$

Bayes Theorem:

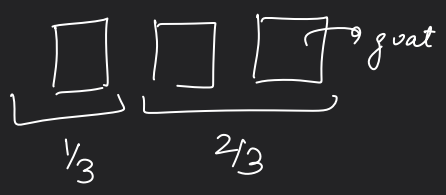
$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(A) \cdot P(B \mid A) + P(A') \cdot P(B \mid A')} \quad \left[\frac{P(A \cap B)}{P(B)} \right]$$

Monty Hall Problem:



as you choose any one door, the host shows you any other door which have goat, should you switch? **YES**

Sol^n:



so it has higher chance to get car in switching (67%)

advanced version $\longrightarrow$ if n doors and host opens k doors.

Sol^n YES. as choosen is having $1/n$ prob.

$$P(\text{win} \mid \text{switch}) = \frac{n-1}{n} \cdot \frac{1}{n-k-1}$$

prob if choosen door does not have car

no. of remaining doors

$$\geq \frac{1}{n} = P(\text{win} \mid \text{not switch})$$

Prior		Event		Posterior
$P(A)$	$\longrightarrow$	E	$\longrightarrow$	$P(A \mid E)$

What if more than 2 events?

ex.

$$P(\text{spam} \mid \underbrace{\text{lottery \& winning}}_{\text{words in email}}) = \frac{P(S) \cdot P(L \& W \mid \text{spam})}{P(S) \cdot P(L \& W \mid S) + P(\text{ham}) \cdot P(L \& W \mid \text{ham})}$$

$$= \frac{P(S) \cdot P(L \mid S) \cdot P(W \mid S)}{P(S) \cdot P(L \mid S) \cdot P(W \mid S) + P(H) \cdot P(L \mid H) \cdot P(W \mid H)}$$

$$P(\text{spam} \mid w_1, \ldots, w_n) = \frac{P(\text{spam}) \cdot P(w_1 \mid \text{spam}) \ldots P(w_n \mid \text{spam})}{P(\text{spam}) \cdot P(w_1 \mid \text{spam}) \ldots P(w_n \mid \text{spam}) + P(\text{ham}) \cdot P(w_1 \mid \text{ham}) \ldots P(w_n \mid \text{ham})}$$

Naive Bayes

Random Variable : (X)

- Random variables allow you to model the whole experiment at once.

eg. X = no. of heads

1 heads	0 heads
$X = 1$	$X = 0$
$P(X=1) = 0.5$	$P(X=0) = 0.5$

Discrete RV	Continuous RV
Can take only countable number of values.	Infinite no. of values.
could be infinite too	Take values on an interval.

Deterministic

$x = 2$, $f(x) = x^2$

Fixed Outcome

Random

X = no. of defective items in a shipment

uncertain outcome

Probability Mass Function (PMF) for discrete distribution

$$p_X(x) = P(X = x)$$

value of R.V. X.

- $p_X(x) \geq 0$
- $\sum p_x(x) = 1$

Binomial Distribution :

(2 possibilities eg coin)

$\binom{n}{k}$ counts all the combinations for landing k heads in n coin tosses.

$$\binom{n}{k} = \binom{n}{n-k}$$

eg: Event : $X = x$: x heads in 5 tosses

$$p_X(x) = \binom{5}{x} p^x (1-p)^{5-x}, \quad x = 0, 1, 2, 3, 4, 5$$

$$X \sim Bin(\overset{(n)}{5}, p)$$

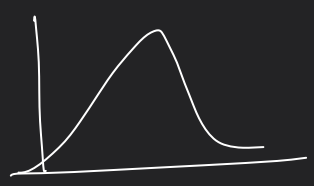
No. of flips

P(H)

Bernoulli Distribution : when success and failure.

$$X \sim Bernoulli(p)$$

prob. of success



Conti. dist → Area = 1

Discrete → sum of heights = 1

Prob. Density Function : $f_X(x)$

(Only defined for continuous variables)

Tell you the rate you accumulate probability around each point.

$P(a < X < b)$ = area under $f_X(x)$. (defined in a range)

$f_X(x)$ needs to satisfy :

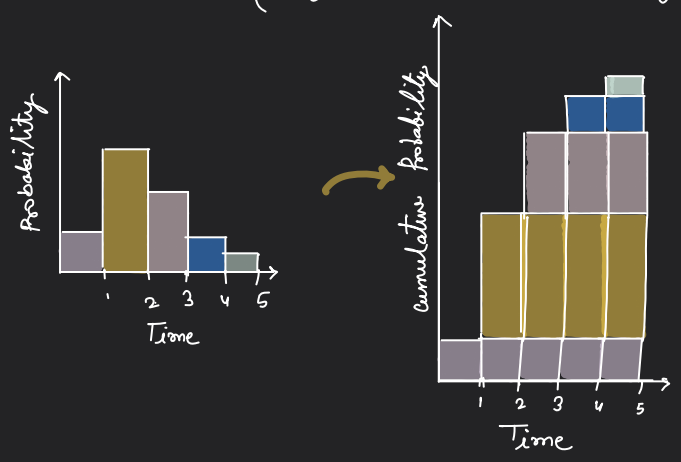
- It is defined for all numbers
- $f_X(x) \geq 0$
- Area under $f_X(x) = 1$.

$P(X = x) = 0 \ \forall x$

- Always positive

Cumulative Distribution :

(cumulative sum type)



The cdf shows how much prob. the variable has accumulated until a certain value.

$F_X(x) = CDF(x) = P(X \leq x)$

- $0 \leq F_X(x) \leq 1$
- left end point is 0
- Right end point is 1.
- Never decreases
- Always positive and decreasing

Uniform Distribution :

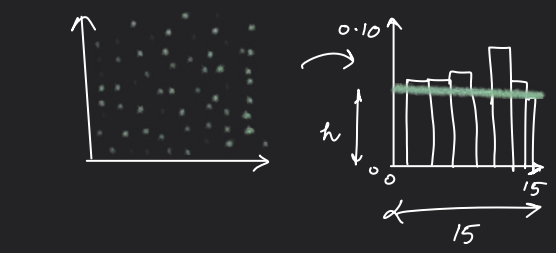
- if all possible values lie in an interval and have same frequency of occurrence.

Parameters :

a : beginning of interval

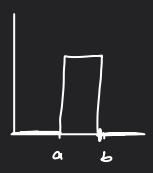
b : end of interval

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \notin (a,b) \end{cases}$$



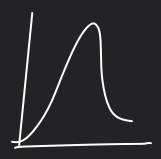
$15 \times h = 1$

$h = \frac{1}{15} = 0.06$



Gaussian / Normal Distribution :

bell shaped data



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

$$f_X(x) : \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Parameters :

μ : center of the bell

σ : spread of the bell

$$Z = \frac{X-\mu}{\sigma}$$: to compare variables of different magnitudes

Chi - squared Distribution :

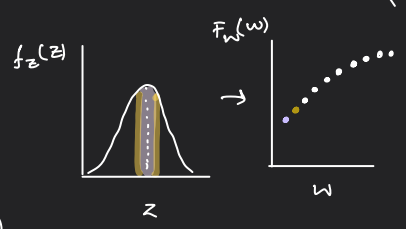
Let's say we have a communication chanel and comm^n channel has noise with a standard normal distribution.

What is the power of the noise in the channel ?

$$W = Z^2$$

What is the distribution of W? (1 degree of freedom)

$$F_W(w) = P(W \le w)$$
$$= P(Z^2 \le w)$$
$$= P(|Z| \le \sqrt{w})$$
$$= P(-\sqrt{w} \le Z \le \sqrt{w})$$



$$f_W(w) = F_W'(w)$$

Accumulated power over 2 transmissions ?

$W_2 = Z_1^2 + Z_2^2$ (chi - sq with 2df)

Accumulated power over 5 transmissions ?

$W_5 = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2$ (5 df)

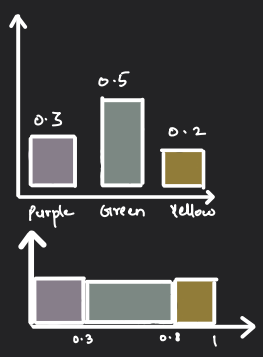
Accumulated power over k transmissions ?

$$W_K = \sum_{i=1}^{K} Z_i^2 \quad (k \text{ df})$$

(As k increases, graph is more spread and symmetrical)

Sampling from a Distribution:

Method 1



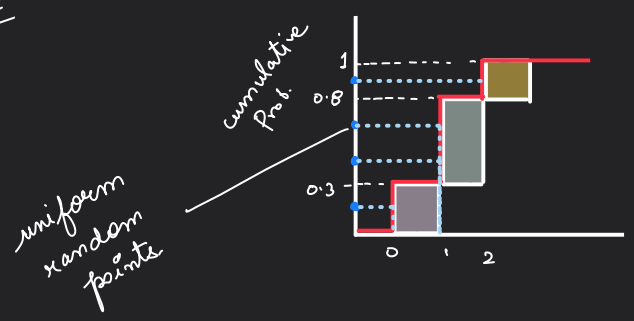
Step 1: Generate a random number between 0 and 1.

Step 2: Find out which interval the number belongs to

- $[0, 0.3)$
- $[0.3, 0.8)$
- $[0.8, 1]$

Step 3: Assign an outcome based on the interval.

Method 2



Expected Value [E(V)]:

(or Weighted avg) $= P(A)\cdot A + P(B)\cdot B + P(C)\cdot C \ldots$

$$E(X) = \sum_x x\, p_X(x)$$ (Discrete Random Variables) [Weighted using PMF]

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$ (Continuous R.V.) [Weighted using PDF]

For uniform distribution, $E(X)$ is in middle or

$$E(X) = \frac{a+b}{2}$$

Median - centermost value

Mode - most frequent. (multimodal distribution)

$$E[aX+b] = aE[X] + b$$

$$E[b] = b$$

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

If we have 8 billion people and give them their names. $E(X) = 1$ which means any 1 person will get correct name assigned.

how

Let's say we have 3 people. then there are 6 possibilities.

and if we solve we get 1

or

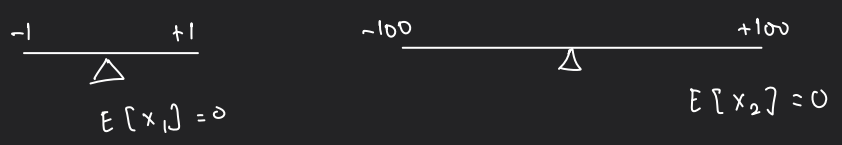
$$E[\text{matches}] = \overset{1^{st}\text{ person}}{\frac{1}{3}} + \frac{1}{3} + \frac{1}{3} = 1$$

$$E[\text{matches}] = \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} = 1$$

But expected values misses a lot of things.

Two functions can have same expected value but one can be narrow and other could be more spread out.

Variance - measuring spread.



$$\text{Variance} = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2$$

↳ Average squared deviation

$$= E[X^2] - E[2E[X]X] + E[E[X]^2]$$

$$= E[X^2] - 2E[X]E[X] + E[X]^2$$

$$= E[X^2] - 2E[X]^2 + E[X]^2$$

$$= E[X^2] - E[X]^2$$

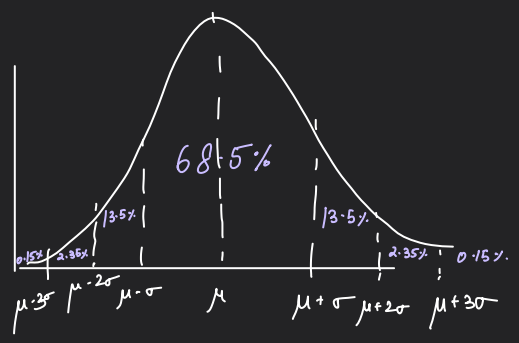
1. Find X's mean
2. Find deviation from that mean for every value of X.
3. Square those deviations
4. Average those squared deviations

$$Var(aX+b) = a^2 Var(X)$$

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

$$\text{Standard deviation} = \sqrt{Var(X)}$$

For normal distribution:



Let's say $R = T + L$ and T an L are gaussian distribution.

Then R is also Gaussian.

$T \Rightarrow N(10, 2^2)$ $\qquad L \Rightarrow N(5, 1)$

Now, $\mu_R = \mu_T + \mu_L = 10 + 5 = 15$

$\sigma_R^2 = \sigma_T^2 + \sigma_L^2 = 4 + 1 = 5$

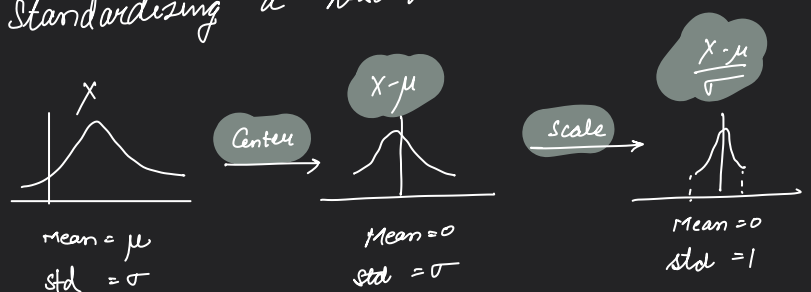
Thus $R \sim N(\underbrace{15}_{\mu}, \underbrace{5}_{\sigma^2})$

In general,

$$W = aX + bY$$

$$\Rightarrow W \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

Standardizing a distribution: to make $\mu = 0$ and $\sigma^2 = 1$.



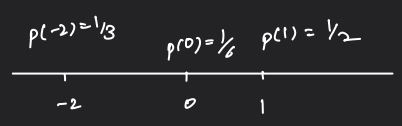
Moments of a distribution:

1st moment $\quad E[X] = \frac{1}{3}(-2) + \frac{1}{6}(0) + \frac{1}{2}(1)$

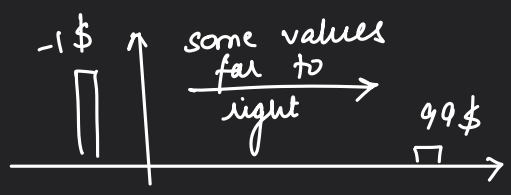
2nd moment $\quad E[X^2] = \frac{1}{3}(-2)^2 + \frac{1}{6}(0)^2 + \frac{1}{2}(1)^2$

$E[X^3] = \frac{1}{3}(-2)^3 + \frac{1}{6}(0)^3 + \frac{1}{2}(1)^3$

$E[X^k] = \frac{1}{3}(-2)^k + \frac{1}{6}(0)^k + \frac{1}{2}(1)^k$



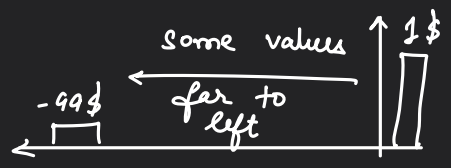
Skewness :



$E[x_1] = 0$

$E[x_1^2] = 99$

$E[x_1^3] = 9702$

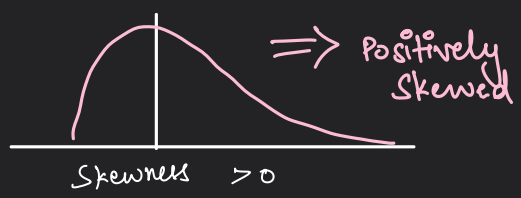


$E[x_1] = 0$

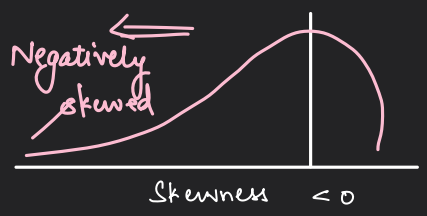
$E[x_1^2] = 99$

$E[x_1^3] = -9702$

3rd moment tells about the skewness



Not skewed
Skewness = 0



$$\text{Skewness} = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

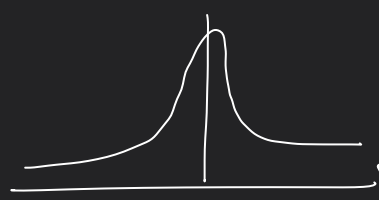
Kurtosis :

$E[x^4]$ almost.

$$\text{Kurtosis} = E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]$$

Thin tails

$E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]$ = small

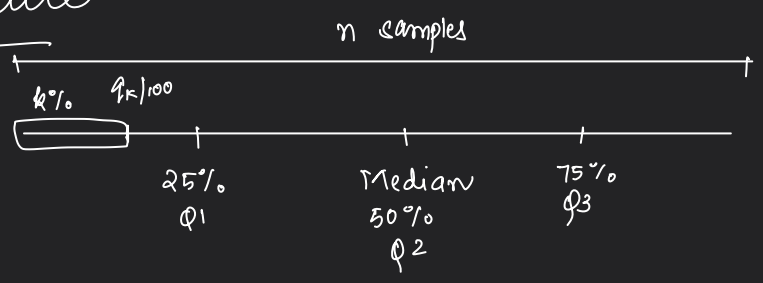


thick tails

$E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]$ = large

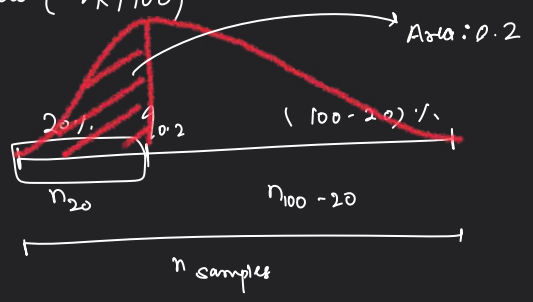
Even if they have same variance and 1st, 2nd and 3rd moments.

Quantile :



k % quantile ($q_k/100$)

for 20 %



$$\frac{20}{100} = \frac{n_{20}}{n} \approx P(X \le q_{0.2})$$

> k % quantile ($q_k/100$) is the value such that $P(X \le q_{k/100}) = \frac{k}{100}$

ex.

Newspaper Ad (X)

8.7	14.2	18.3	18.4
23.2	25.9	29.7	35.2
51.2	54.7	65.9	75

$$Q_1 = q_{0.25} = \frac{18.3 + 18.4}{2} = 18.35$$

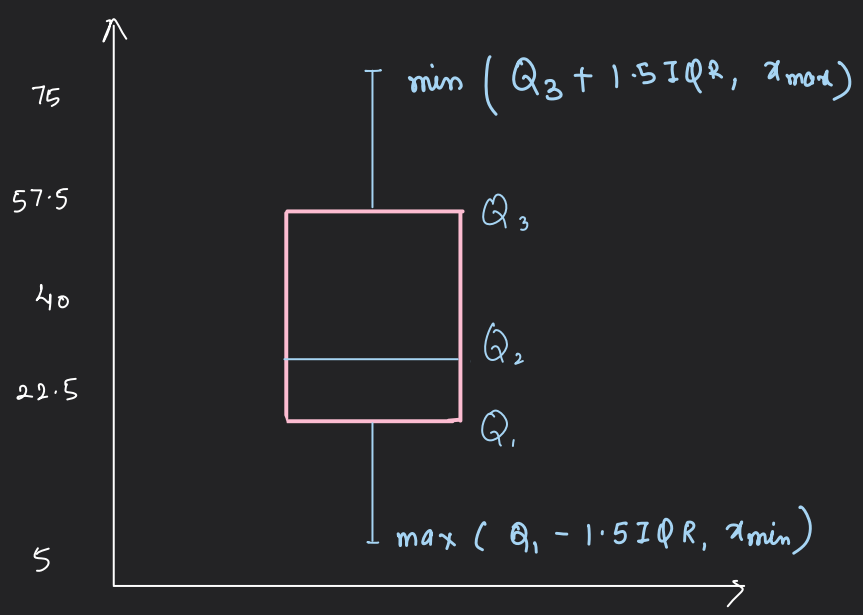
$$Q_2 = q_{0.50} = \frac{25.9 + 29.7}{2} = 27.8$$

$$Q_3 = q_{0.75} = \frac{51.2 + 54.7}{2} = 52.95$$

$x_{min} = 8.7$

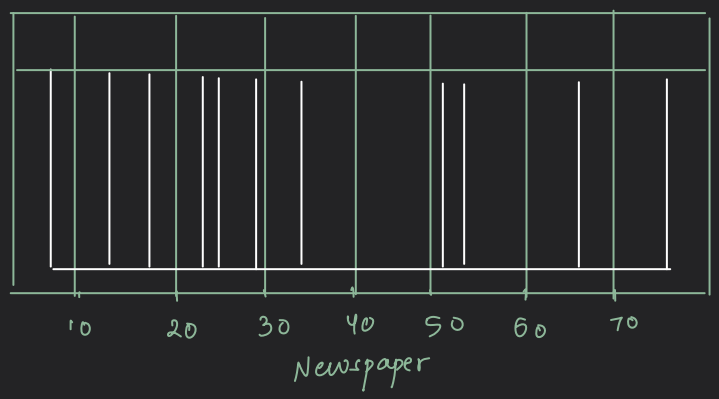
$x_{max} = 75$

$$\underset{\text{(Interquartile Range)}}{IQR} = Q_3 - Q_1 = 52.95 - 18.35 = 34.6$$

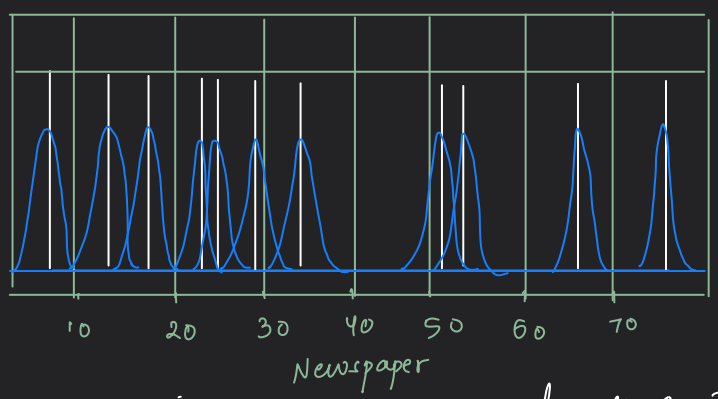


Kernel Density Estimation :

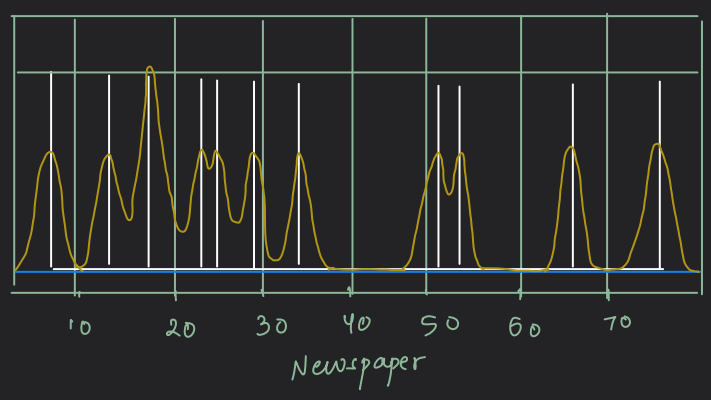
1. Draw your observations along the x axis.



2. Draw a gaussian centered at each observation.

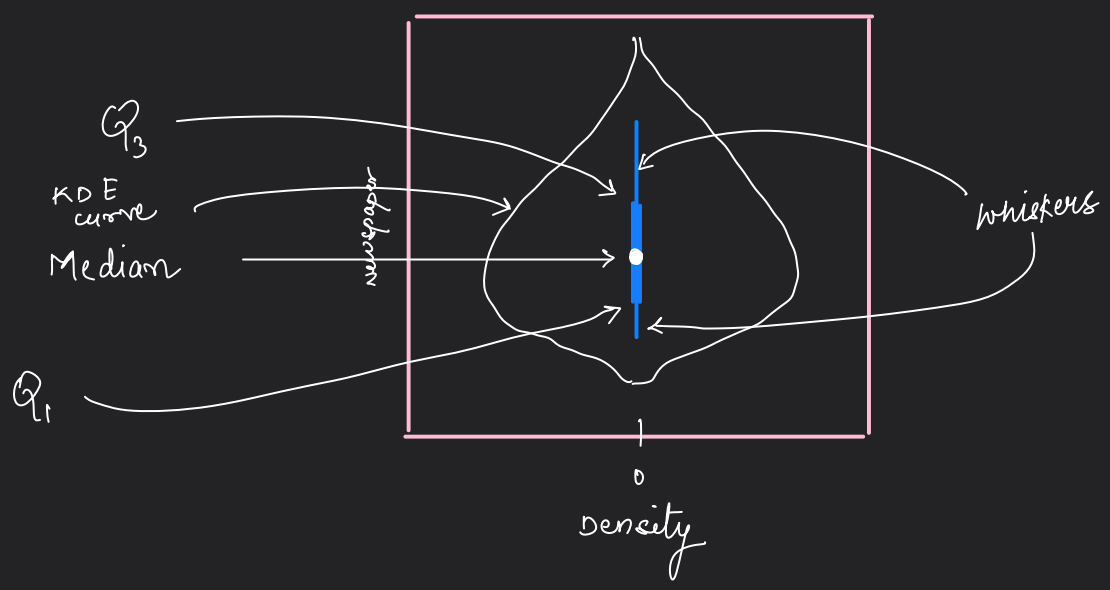


3. Multiply everything by 1/n and sum the curves.



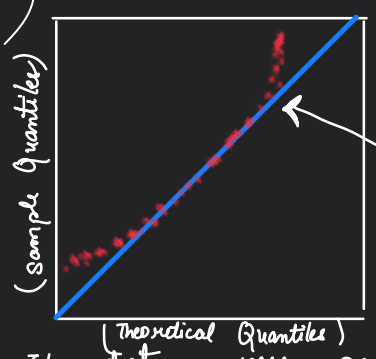
Violin Plots :

- contains info of KDE and box plots



Quantile - Quantile Plots (QQ Plots) :

- standardize your data $\left(\frac{x-\mu}{\sigma}\right)$
- compute quantiles
- compare to gaussian quantiles.



If data were gaussian, the dots should align with this line.

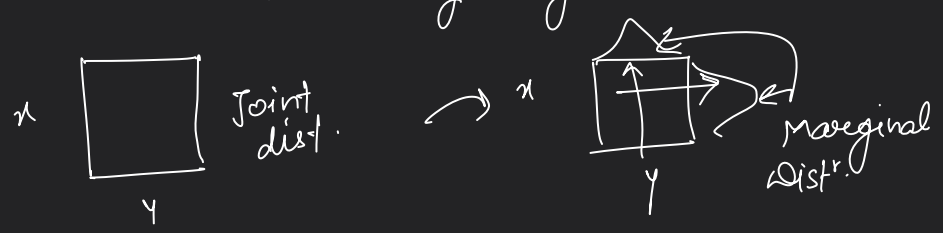
Joint Distributions :

$$p_{XY}(x,y) = P(X=x, Y=y)$$

For independent discrete random variables :

$$p_{XY}(x,y) = P(X=x, Y=y) = P(x) \cdot P(y)$$

Marginal Distribution : Distribution of one variable while ignoring the other



Conditional Distribution : (just like cond^n prob.)

$$P_{Y|X=x}(y) = P(Y=y \mid X=x)$$

$x \to (q)$

Discrete Conditional Distribution :

$$P_{Y|X=x}(y) = \frac{P_{XY}(x, y)}{P_X(x)}$$

Cond^n PDF of Y

Joint PDF of X and Y

Continuous Cond^n dist^n :

$$f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Conditional PDF of Y

Joint PDF of X and Y

Marginal dist^n of X

Covariance : (how one variable is related to other variables when variance is same.)

$$Cov(X,Y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$$

($\sum xy$ Almost ...)

$Cov(X,Y) > 0$ $\quad$ $Cov(X,Y) \simeq 0$ $\quad$ $Cov(X,Y) < 0$

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

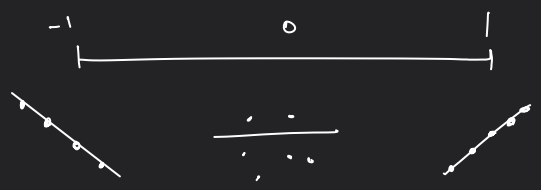
Covariance Matrix :

	X	Y
X	Var(X)	Cov(X, Y)
Y	Cov(Y, X)	Var(Y)

	X	Y	Z
X	Var(X)	Cov(X,Y)	Cov(X, Z)
Y	Cov(X,Y)	Var(Y)	Cov(Y,Z)
Z	Cov(X,Z)	Cov(Y,Z)	Var(Z)

can be diff for same dist^n due to large values so corr. coeff^n.

Correlation Coefficient: (lies between -1 and 1)



$$\text{Corr. Coeff}^n = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y}$$

Multivariate Gaussian Distribution:

For single variable, x

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

μ: center of bell
σ: spread of bell

If more than one variable:

If independent →

$$f_{HW}(h,w) = f_H(h)\, f_W(w)$$

$$= \frac{1}{2\pi\sigma_H\sigma_W} e^{-\frac{1}{2}\left[\boxed{\frac{(h-\mu_H)^2}{\sigma_H^2} + \frac{(w-\mu_W)^2}{\sigma_W^2}}\right]}$$

$$\left\|\begin{bmatrix}\frac{h-\mu_H}{\sigma_H}\\ \frac{w-\mu_W}{\sigma_W}\end{bmatrix}\right\|_2^2 = \begin{bmatrix}\frac{h-\mu_H}{\sigma_H} & \frac{w-\mu_W}{\sigma_W}\end{bmatrix}\begin{bmatrix}\frac{h-\mu_H}{\sigma_H}\\ \frac{w-\mu_W}{\sigma_W}\end{bmatrix}$$

Multiply by diagonal matrix

$$\begin{bmatrix}h-\mu_h\\ w-\mu_W\end{bmatrix} = \begin{bmatrix}h\\ w\end{bmatrix} - \begin{bmatrix}\mu_h\\ \mu_W\end{bmatrix}$$

$$= \left(\begin{bmatrix}h & w\end{bmatrix} - \begin{bmatrix}\mu_H & \mu_W\end{bmatrix}\right)\begin{bmatrix}\frac{1}{\sigma_H^2} & 0\\ 0 & \frac{1}{\sigma_W^2}\end{bmatrix}\left(\begin{bmatrix}h\\ w\end{bmatrix} - \begin{bmatrix}\mu_H\\ \mu_W\end{bmatrix}\right)$$

$$= \frac{1}{2\pi\sigma_H\sigma_W} \exp\left(\begin{bmatrix}h\\ w\end{bmatrix} - \begin{bmatrix}\mu_H\\ \mu_W\end{bmatrix}\right)^T \underbrace{\begin{bmatrix}\sigma_H^2 & 0\\ 0 & \sigma_W^2\end{bmatrix}^{-1}} \left(\begin{bmatrix}h\\ w\end{bmatrix} - \begin{bmatrix}\mu_H\\ \mu_W\end{bmatrix}\right)$$

$$= \frac{1}{2\pi \det \Sigma^{1/2}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}h\\ w\end{bmatrix} - \mu\right)^T \boxed{\Sigma^{-1}} \left(\begin{bmatrix}h\\ w\end{bmatrix} - \mu\right)\right) \quad \text{(for dependent)}$$

$$\boxed{f_X(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}}$$

Population (N) : is the entire group or population that we want to study.

(Population Mean $\rightarrow \mu$)

Sample (n) : is the smaller subset that we actually observe or measure.

(sample mean $\rightarrow \bar{x}$)

While selecting sample set 1, then 2. make sure the selection in 2 or 3 or so on does not depend on previous one or 1. So, keep random selection, rather than avoiding repetition during selection.

→ Samples must be identically distributed.

→ Bigger the sample size, better the estimate.

Population Proportion : $p = \dfrac{\text{no. of items with a given characteristic } (x)}{\text{population } (N)}$

Similarily Sample proportion $(\hat{p})$

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

μ → Pop mean

N → Pop size

But with sample

$$\widehat{Var}(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2$$

Law Of Large Numbers :

As the sample size increases, the average of the sample will tend to get closer to the average of the entire population.

Under certain conditions :

- Sample is randomly drawn.
- Sample size must be sufficiently large.
- Independent observations (must be independent of each other)

Center Limit Theorem :

When you average a large enough no. of variables, the distribution will approximately follow a normal distribution.

(min 30 if possible)

Y_n = mean of random samples

$\mu_{Y_n} = \mu$ ← population mean

$\sigma_{Y_n}^2 = \frac{\sigma^2}{n}$ ← population variance

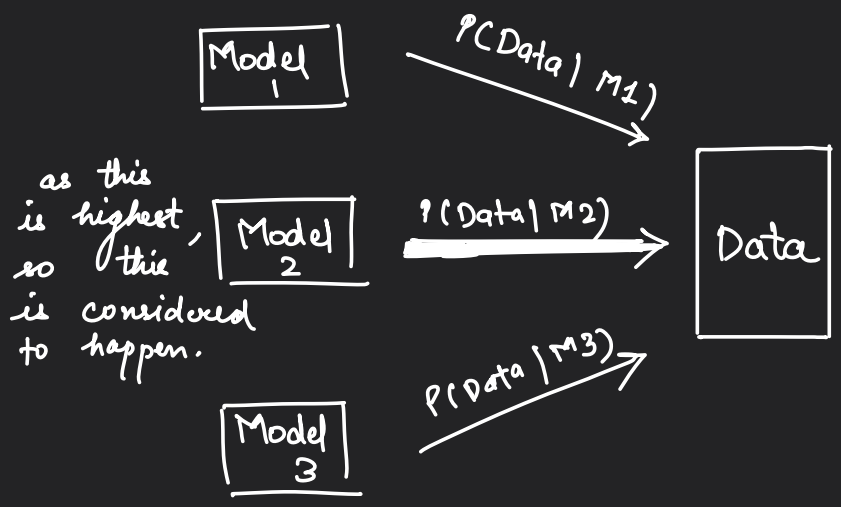
As $n \to \infty$ $\quad \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]}{\sigma_X} \sqrt{n} \sim N(0,1^2)$

or

As $n \to \infty$ $\quad \frac{\sum_{i=1}^{n} X_i - nE[X]}{\sqrt{n}\,\sigma_X} \sim N(0,1^2)$

If the sample is sufficiently large (usually $n > 30$), the the sample means' distributed will be normally distributed, regardless of the underlying population distribution whether it is normal, skewed or otherwise

Maximum Likelihood:



Find model that most likely produced the data.

Bernoulli example →

Let's say we have 3 coins –

Find the coin that most likely produced 8 heads & 2 Tails.

1. $P(H) = 0.7$
2. $P(H) = 0.5$
3. $P(H) = 0.3$

To get 8H 2T

maximise $P(8H, 2T \mid \text{coin})$

here coin 1 is most likely to produce 8H and 2T.

$$P(8H\,2T, \text{coin } 1) = (0.7)^8 (0.3)^2 = 0.0051$$

$$P(8H\,2T, \text{coin } 2) = (0.5)^8 (0.5)^2 = 0.0010$$

$$P(8H\,2T, \text{coin } 3) = (0.3)^8 (0.7)^2 = 0.00003$$

But what could be the best coin →

let's say for head → p
tail → $1-p$

$$L(p; 8H) = p^8 (1-p)^2$$

We want p that maximizes the chances of seeing 8H.

log likelihood

$$\ell(p; 8H) = \log\left(p^8 (1-p)^2\right)$$
$$= 8 \log p + 2 \log(1-p)$$

so we maximise log likelihood.

$$\frac{d}{dp}\left(8 \log(p) + 2 \log(1-p)\right) = \frac{8}{p} + \frac{2}{1-p}(-1) = 0 \Rightarrow \hat{p} = \frac{8}{10}$$

General Case :

n coins

k heads

$x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n$

$x_i \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$

Likelihood

$$L(p;x) = P_p(X=x) = \prod_{i=1}^{n} P_{x_i}(x_i) = \prod_{i=1}^{n} \underbrace{p^{x_i}(1-p)^{1-x_i}}$$

If $x_i = 1$, $p^{x_i}(1-p)^{[1-x_i]} = p$

If $x_i = 0$, $p^{[x_i]}(1-p)^{[1-x_i]} = (1-p)$

$$\sum_{i=1}^{n} x_i = \#\text{ heads} \qquad n - \sum_{i=1}^{n} x_i = \#\text{ tails}$$

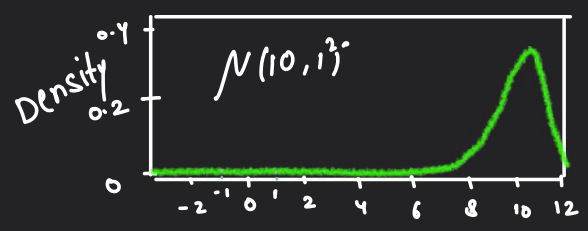
log likelihood :

$$l(p;x) = \log\left(\left(\sum_{i=1}^{n} x_i\right)\log(p) + \left(n - \sum_{i=1}^{n} x\right)\log(1-p)\right)$$

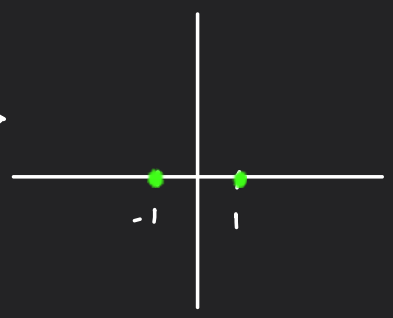
$$\frac{d}{dp} l(p;x) = \frac{\sum_{i=1}^{n} x_i}{p} + \frac{n - \sum_{i=1}^{n} x_i}{1-p}(-1) = 0$$

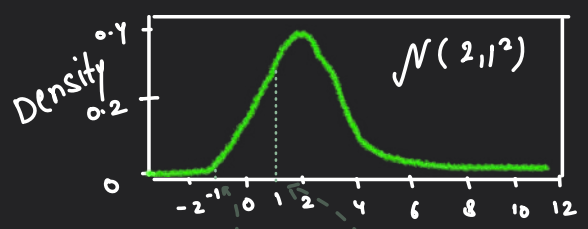
$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Maximum Likelihood : (Gaussian example)



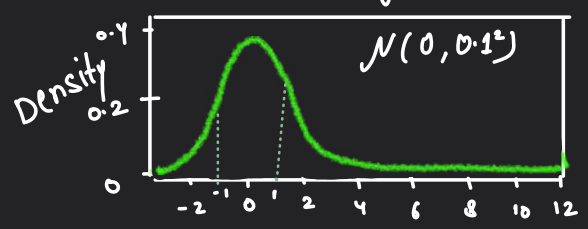
Low Likelihood





Medium likelihood

likelihood = product of heights



Highest Likelihood

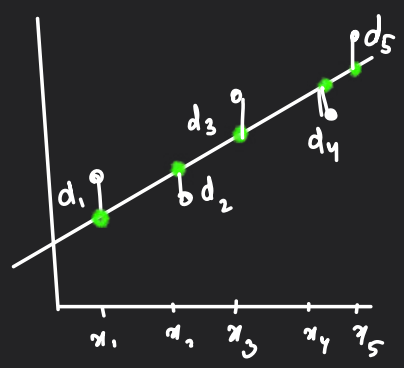
The best dist^n is the one where the mean of the dist^n is the mean of the sample

and same for variance.

For linear regression – The model which generates points close to the data or observations is the most likely one

Likelihood :

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_2^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_3^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_4^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_5^2}$$



Maximise :

$$e^{-\frac{1}{2}d_1^2} \cdot e^{-\frac{1}{2}d_2^2} \cdot e^{-\frac{1}{2}d_3^2} \cdot e^{-\frac{1}{2}d_4^2} \cdot e^{-\frac{1}{2}d_5^2}$$

$$e^{-\frac{1}{2}(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2)}$$

Minimize

$(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2)$ (Least squares error)

But its not always good, as it may overfit so we add Penalty term to loss

(Regularisation)

$\mathcal{L}_2 = a_n^2 + a_{n-1}^2 + \dots + a_1^2$

Reg. parameter $\Rightarrow \lambda$

Regularized error : $\text{log loss} + \lambda(a_n^2 + a_{n-1}^2 + \dots + a_1^2)$

Frequentist	Bayesian
• Prob represent long term frequency of events	• Prob represent the degree of belief (or certainty)
• Concept of likelihood	• Concept of prior
• Goal : Find the model that most likely generated the observed data	• Goal : update prior belief based on observations.

Maximum a Posterior (MAP)

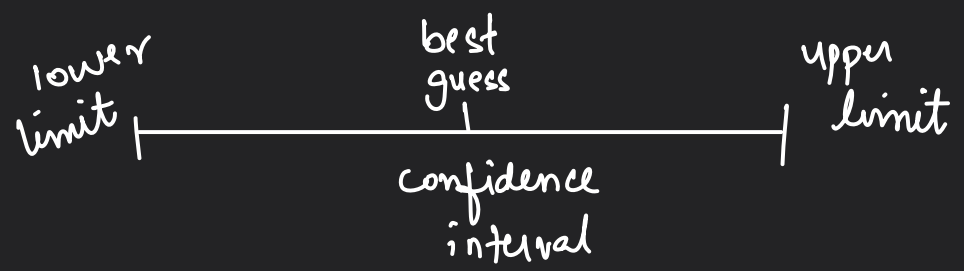
we need to choose 1 value for parameter $\Rightarrow$ Choose the one with highest prob. $\Rightarrow$ Mode of the $\boxed{\text{updated belief}}$ $\Downarrow$ posterior

Posterior $\rightarrow$ $$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$ $\leftarrow$ Prior

Confidence Interval :

↳ prob. that interval contains μ.

Confidence level - 50% of guesses are within a search distance



α : significance Level - prob. that your sample mean falls outside your margin of error.
$(\alpha = 0.05)$

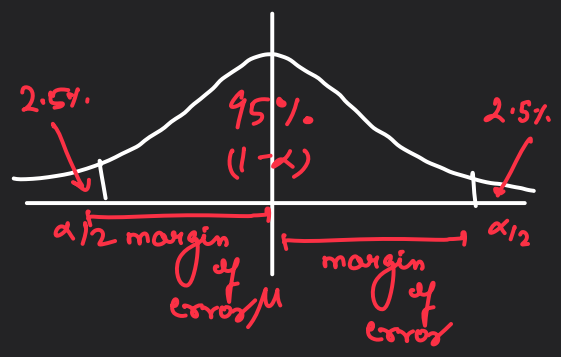
Conf. Level $= (1-\alpha)$: prob. that a randomly generated sample mean is within the margin of error.

if $\alpha = 0.05$

$1 - \alpha = 0.95$ so 95% Confidence level.

Confidence interval:

$\bar{x} \pm$ margin of error

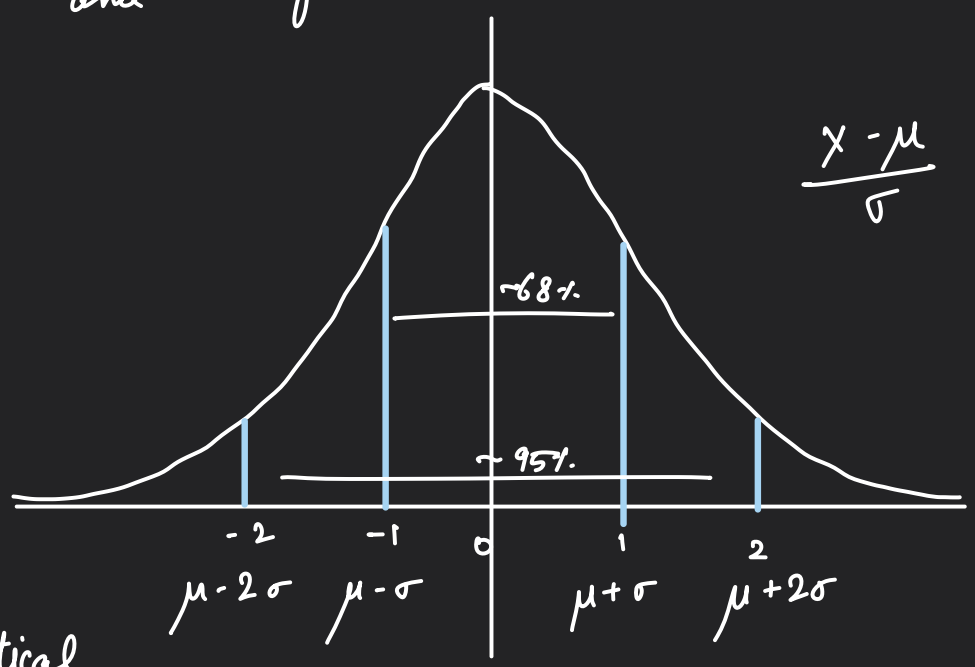


mean of the sample means $\rightarrow$ $\boxed{\mu_{\bar{x}} = \mu}$ $\leftarrow$ population mean

Std dev of sample means $\leftarrow$ $\boxed{\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}}$

- as sample size increase, the confidence interval decreases.

- Margin of error is gathered by sample size and confidence interval.



$\dfrac{x-\mu}{\sigma}$ = Z score or dist^n. $\sim N(\mu, \sigma^2)$

(from table) critical values $\leftarrow$ CV: -1.96 (z)

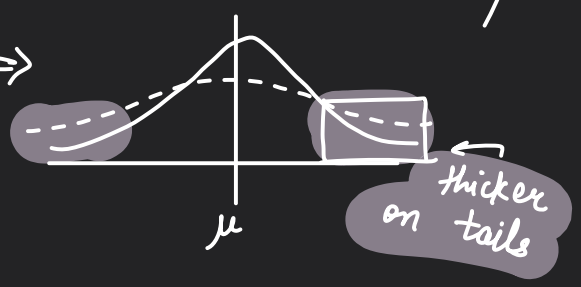
CV: 1.96

Margin of error : $= Z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Confidence Interval $= \bar{x} \pm$ margin of error

if std dev. (σ) unknown, we can use s. ← std dev. of sample.

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \Longrightarrow \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \Longrightarrow$$



not a normal dist^n

student's t dist^n

$$CI. = \bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

t-score

Degrees of freedom of t-dist^n = (no of samples − 1)

↳ defines how high and tight t-dist^n is.

- Larger the d.o.f., more close to Normal dist^n (as $s \to \sigma$)

CI. for proportions : ($\hat{p}$ = proportion) eg 80%

$$CI = \hat{p} \pm ME$$

$$ME = Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Hypothesis Testing :

null hypothesis → H_0 → baseline
alternate hypothesis → H_1 → what you want to prove

Mutually Exclusive
True / False answer

If plenty of evidence against H_0 ⟶ Reject H_0 and accept H_1

Type I error (False Positive) : rejects H_0 when actually true.

Type II error (False Negative) : when don't reject H_0 and hypothesis was actually false.

Decision	Reality	
	H_0 True (Not spam)	H_0 False (Spam)
Reject H_0 (Decide spam)	Type I error	Correct
Don't reject H_0 (Decide not spam)	Correct	Type II error

more severe.

⇒ Significance level →

Sending a regular email to spam is worse than not sending a spam email to the regular inbox.

$\alpha = 0.05$ (You want small α, but not zero)

0 ——— α ——— 1

Emails are always ham (No Type I error)

But (↓ Type I error ↑ Type II error)

Every email is considered spam. (Every time an email is ham you make a Type I error)

3 types of hypothesis :

Right tailed test $\longrightarrow$ $H_0: \mu = 66.7$ vs $H_1: \mu > 66.7$

Left tailed test $\longrightarrow$ $H_0: \mu = 66.7$ vs $H_1: \mu < 66.7$

Two tailed test $\longrightarrow$ $H_0: \mu = 66.7$ vs $H_1: \mu \neq 66.7$

Reject H_0 if $\underbrace{P(H_1 \mid H_0)}_{\text{(p - value)}} < \alpha$ ← 0.05

Type I error prob.

p - value : is the prob., assuming H_0 is true, that the test statistic takes on a value as or more extreme than the value observed

If p - value $< \alpha \Rightarrow$ reject H_0 (and accept H_1 as true)

p - value $> \alpha \Rightarrow$ don't reject H_0

What is the least extreme sample you could get that you would still reject H_0?

or Sample that has p - value $= \alpha \Rightarrow$ Critical Values (k_α)

Reject H_0 if $t > k_\alpha$ for right tailed

Reject H_0 if $t < k_\alpha$ for left tailed

Reject H_0 if $t > k_{\alpha_1}$ or $t < k_{\alpha_2}$ for two - tailed

T- distribution :

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$ follows a t - distribution

Parameters :

Degrees of freedom (ν) $(n-1)$: controls how heavy tails are $X \sim t_\nu$

(Similiar, left, right and two tailed tests)

Independent two sample t-test →

$X \sim N(\mu_x, \sigma_x)$ $\quad$ $Y \sim N(\mu_y, \sigma_y)$

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}\right) \Rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}} \sim N(0,1^2)$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{N_x} + \frac{S_y^2}{N_y}}} \sim t_\nu$$

Degrees of freedom: $= \dfrac{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)^2}{\frac{\left(\frac{S_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{S_y^2}{n_y}\right)^2}{n_y - 1}}$

now similiar - right, left, two-tailed

Paired t-test: dependent)

$$\bar{D} = \frac{\sum D_i}{n} \qquad D_i = X_i - Y_i$$

$$\frac{\bar{D} - \mu_D}{\frac{\sigma_D}{\sqrt{N}}} \sim N(0,1^2)$$

But σ_D is unknown $\Rightarrow \sigma_D \longrightarrow S_D = \sqrt{\dfrac{\sum_{i=1}^{10} (D_i - \bar{D})^2}{N-1}}$

$$\Rightarrow \underbrace{T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{N}}} \sim t_{N-1}}_{\text{Test statistic}}$$

A/B Testing : is a methodology for comparing two variations.

Propose Variations (A/B) → Randomly split sample → Measure outcomes of each group and determine a metric → Statistical analysis to make a decision (t-test)