Statistics

Types of data

Categorical or Qualitative data

- **Nominal** — male/female — No hierarchy
- **Ordinal** — bad / avg / good — order

Numerical or Quantitative data

- **Discrete** — age/rank — 35, 23
- **Continuous** — weight/height — 69.5, 72.7

Population mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	[N = no. of items in population]
Sample mean	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	[n = no. of items in sample]

↳ Mean is prone to outliers, median can solve outliers issue. Mode can work best for categorical columns to find highest frequency. Weighted mean we assign some weights & trimmed mean we remove fraction of values from both the sides to remove outliers.

↳ Variance $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ used in finding dispersion of data

Mean absolute deviation | $MAD = \frac{\Sigma |x_i - \bar{x}|}{n}$

Standard deviation | $SD = \sqrt{\sigma}$ | this holds the same unit as data

Coefficient of variation | $CV = \frac{\text{Standard deviat}}{\text{mean}} \times 100\%$ | It gives the idea of spread of colums & compare

↳ Graphs for univariate columns ÷ Categorical & Numerical

a) Categorical - Frequency barcharts, relative frequency (pie chart), cummulative f (line chart)

b) Numerical - Bins histograms, normal, skew, bimodal, uniform, no patter

↳ Graphs for bivariate colums ÷ C&C, N&N, C&N

a) Categorical & C ÷ Contingency table or crosstab

b) Numerical & N ÷ Scatterplots

c) C & N ÷ Bar charts with aggregate f^n or cross tabs

Survived \ Pclass	1	2	3
0	21	40	11
1	12	14	50

Pivot tables

↳ Quantiles :- Divide a set of numeric data into equal sized group

Quartiles :- Q1 (25 percentile), Q2 (50 percentile), Q3 (75 perc)

Deciles :- (10^{th} percentile) D1 ... D9

Percentile :- P1, P2 ... P99

Quintile :- 5 equal parts [minimum, Q1, Q2, Q3, max] df.describe()

↳ IQR = Q3 − Q1 = Interquartile range [Box of boxplot]

↳ $PL = \frac{P}{100}(N+1)$ PL = desired percentile value location

N = no. of obs. P = percentile rank

Ex Find 75 percentile score from data,

78, 82, 84, 88, 91, 93, 94, 96, 98, 99 → sort the values

$PL = \frac{75}{100}(10+1) = \frac{33}{4} = 8.25$ ∴ 96 — 98 : 96 + (98−96) × 0.25 = 96.5

8 + 9 × 4

8.25

↳ Percentile of a value :-

$\text{Percentile rank} = \frac{x + 0.5y}{n}$

x → no. of values below given

y → no. of values equal to given

n → total no. of values

Ex 78, 82, 84, (88), 91, 93, 94, 96, 98, 99

(1, 2, 3)

$\therefore \frac{3 + 0.5(1)}{10} = 0.35 = 35\%$

Ex How to create boxplots ?

6	213	241	260	281	290	314	321	350	1500
1	2	3	4	5	6	7	8	9	10

$Q2 = \frac{50}{100}(11) = 5.5 = 285.5$

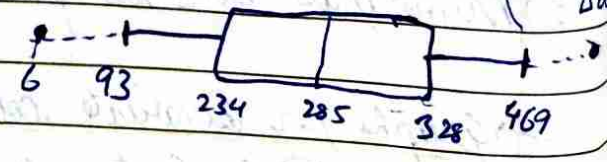
$Q1 = \frac{25}{100}(11) = 2.75$ ∴ 213 + 0.75(241 − 213) = 234

$Q3 = 75 \times 11/100 = 8.25 = 328.25$

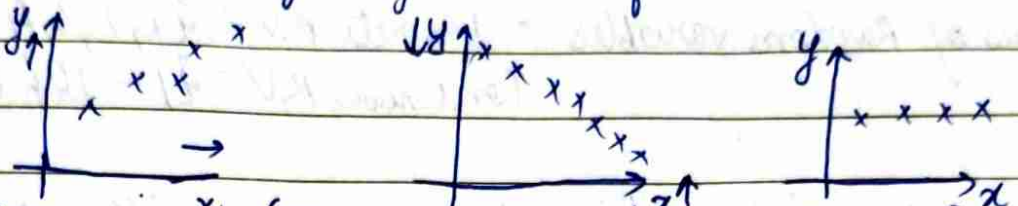
min = Q1 − 1.5(IQR) = 93

max = Q3 + 1.5(IQR) = 469

IQR = Q3 − Q1 = 328 − 234 = 94



↳ Covariance : Statistical measure that describes the degree to which 2 variables are linearly related. Measures how much 2 variables change together like if on ↑ other ↑ or ↓



Covariance formula :

Population	Sample
$\sigma_{xy} = \frac{\Sigma(X-\mu_x)(Y-\mu_y)}{N}$	$s_{xy} = \frac{\Sigma(X-\bar{x})(Y-\bar{y})}{n-1}$
$X, Y \rightarrow$ value of X, Y in populat^n	$X\ Y \rightarrow$ Same
$\mu_x, \mu_y \rightarrow$ population mean	$\bar{x}, \bar{y} \rightarrow$ sample mean
$N \rightarrow$ no. of observation	$n \rightarrow$ Same

↳ Disadvantage of covariance :
Does not tell us about strength of relat^n b/w 2 variable as magnitude of covariance is affected by the scale of variable
Covariance tells us about the spread of data. it gets scaled up

↳ Correlation : Refers to statistical relat^n b/w variables it measures the degree to which 2 variable are related & how they change

$$\text{Correlation} = \frac{\text{Covariance}(x,y)}{\sigma_x * \sigma_y}$$

(σ_x : standard deviat^n of x)

★ If 2 things are correlated does not mean it is causation or causeright
Causation refers to a cause & effect relat^n b/w 2 variables

↳ Graphs for multivariate columns :
Scatterplots 3D, Barplots with Hue, Facetgrids, Jointplots, Pairplots, Bubblecharts

↳ Random variables :- Set of possible values from a random exp.
Ex coin toss H & T ... X = {0, 1} H=0, T=1
Random variable [capital letters]
Types of Random variables :- Discrete RV = {1,2,3,4,5,6} Dice
Continuous RV = {1 ... 10} CGPA

↳ Probability distribut^n :- list of all of the possible outcome of a random variable along with their probability values

Coin toss	H(0)	T(1)
Probability	1/2	1/2

We can derive mathematical f^n to model relat^n b/w outcome & prob
X → outcome 1 2 3 4 5 6
Y → probability 1/6 1/6 1/6 1/6 1/6 1/6
$y = f(x)$ → Probability distribut^n function

Probability mass funct^n (PMF) → Cummulative Distribut^n f^n (CDF)
Probability density funct^n (PDF)

↳ PMF → probability distribution of discrete random variable (Dice (1-6))
Sum of prob is equals to 1

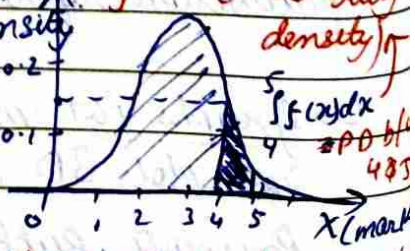
$$\therefore y = \begin{cases} 1/6 & \text{if } x \in \{1,2,3,4,5,6\} \\ 0 & \text{if otherwise} \end{cases}$$

CMF of PMF describes the probability of X found at $\leq x$

$F(x) = P(X \leq x)$

Ex for dice role $P(4) = 1/6$
$CMF(4) = P(X \leq 4) = P(1) + P(2) + P(3) + P(4) = 4/6$
$CMF(6) = P(X \leq 6) = 1$ → np.cumsum(s)

↳ PDF is a mathematical f^n describes prob. distribut^n of continuous Vari
* For ∞ range of value y-axis cannot be probabity prob. y-axis (Probability density)
So we take small step of ∫ to find prob. of density
getting marks b/w 4-4.00...1 that value is on y-axis



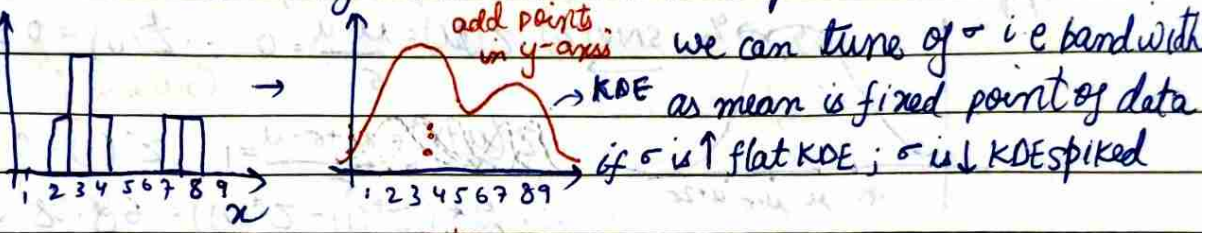
$\int_4 f(x)dx$ = PD b/t 4&5

$P(0 \leq X \leq 10) = 1 = \text{Area}$

↳ Density estimation ÷ Statistical technique used to estimate the prob. density f^n (PDF), estimating the underlying distribut^n of data pt.
2 types ÷ Parametric density estimation & Non-parametric DE
Commonly used techniques ÷ Kernel density estimat^n (KDE), Histogram estimat^n, Gaussian mixture models (GMMs)

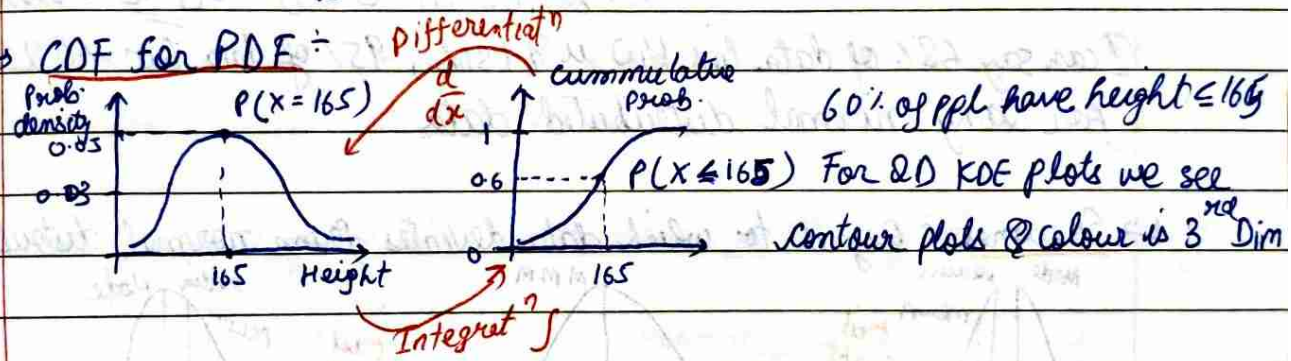
a) Parametric density estimation ÷ Eg we have data of CGPA from 1-10 we see its histplot and observe which kind of distribut^n say normal distribut^n then we use the PDF formula of normal with parameters of μ & σ to create the distribution

b) Non-parametric DE ÷ without making underlying assumptions here we will be using the data rather than parameters



we can tune of σ i.e bandwidth as mean is fixed point of data
if σ is ↑ flat KDE ; σ is ↓ KDE spiked

↳ CDF for PDF ÷

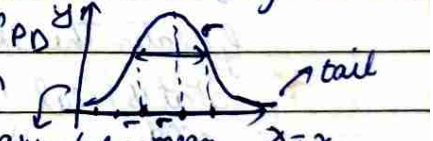


60% of ppl have height ≤ 165
For 2D KDE plots we see contour plots & colour is 3rd Dim

↳ Normal distribution ÷ continuous probability distribut^n symetric around mean and in bell curve (Gaussian)

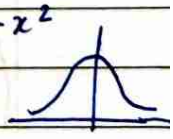
2 parameters ÷ mean (μ) centre of distribut^n PD
Standard deviat^n (σ) spread of distribut^n

$X \sim N(\mu, \sigma)$
belongs ↳ Normal



PDF Eq^n of normal dist

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

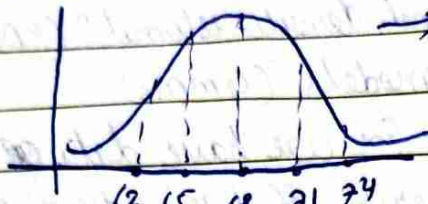


Standard normal variate $Z \sim N(0,1)$ mean $\mu = 0$ Std. $\sigma = 1$
Convert any column to standard normal variate

$N(\sigma, \mu)$ ← Age − Age → $Z \sim N(0,1)$ ↳ Symetrical
27, 26 → μ ∴ $Z_1 = \frac{27-\mu}{\sigma}$ ↳ mean = median = mode
61, 73 → $Z_2 = \frac{61-\mu}{\sigma}$ ↳ 68-95-99.7 Rule
↳ Area is 1

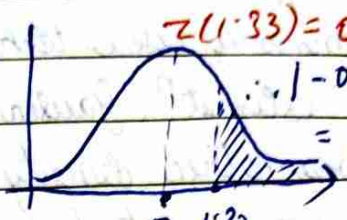
Q) Height of adults follow normal distribution with $\mu = 68$ & $\sigma = 3$ inches. Find probability that a random male is taller than 72 inches?

$\therefore X \sim N(68, 3)$



$\rightarrow Z = \frac{72-68}{3} = 1.33$

from Z-table

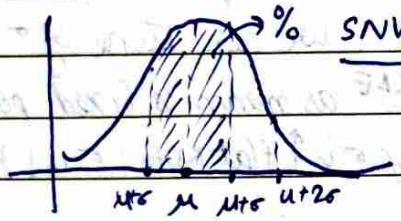


$z(1.33) = 0.908$

$\therefore 1 - 0.908 = 0.0902$

Z-table of std. normal dist. is computed $\therefore$ 9% chance tall > 72 inches

Q) For a normal distribution $X \sim N(\mu, \sigma)$ what percent of populat^n lie b/w mean & 1 std, 2 std, 3 std?



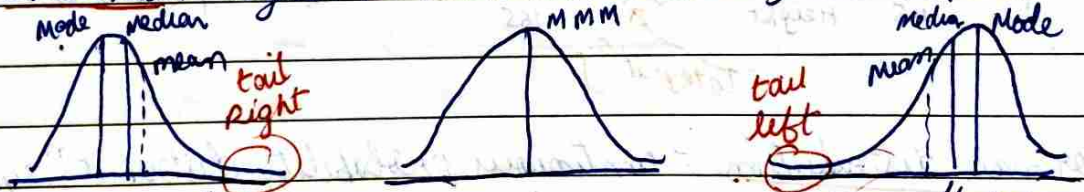
$z(\mu) = \frac{\mu - \mu}{\sigma} = 0 \quad \therefore z^t(0) = 0.5$ ← area $\int_{-\infty}^{x}$

$z(\mu+\sigma) = \frac{\mu+\sigma-\mu}{\sigma} = 1 \quad \therefore z^t(1) = 0.84$

$\therefore 2 * (z^t(1) - z^t(0)) = 68.28\%$ b/w μ & 1σ

I can say 68% of data lies b/w μ & 1 std, 95% of data b/w μ & 2 std for any normal distributed data

↳ Skewness: Degree to which data deviates from normal distribution



Right skew / (+) skew — Left skew / (−) skew

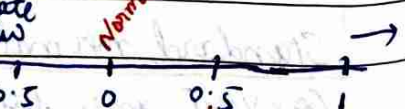
Greater the skew ↑ distance b/w mean, median, mode

* Statistical moment: 1st moment = mean

2nd moment = Variance^2, 3rd moment = Skew^3, 4th mom = Kurtosis^4

$$\text{skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{(x - \bar{x})}{s} \right)^3$$ → Sample skew pandas uses this skew

highly skew ← moderate skew ← Normal

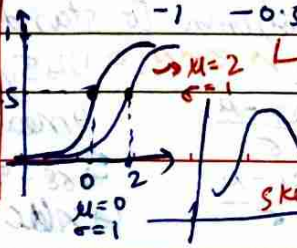


No skew ≈ normal

↳ CDF on normal curve

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t-\mu)^2}{2\sigma^2}} dt$$



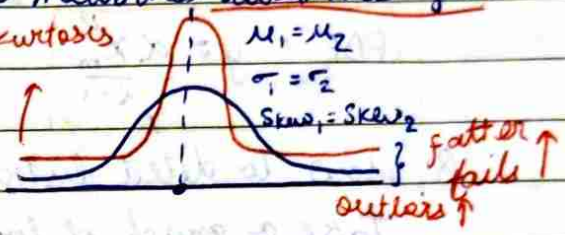
skew ≈ 0 — Not always for normal graph

↳ Used in data science to detect outliers, hypothesis testing, central limit theorem, assumpt^n of ML algo like GMM & linear Reg (errors is normally distributed)

↳ Non-gaussian distribution → Continous non-gaussian
↳ Discrete non-gaussian

Kurtosis is 4th statistical moment, it measures tailedness of PD

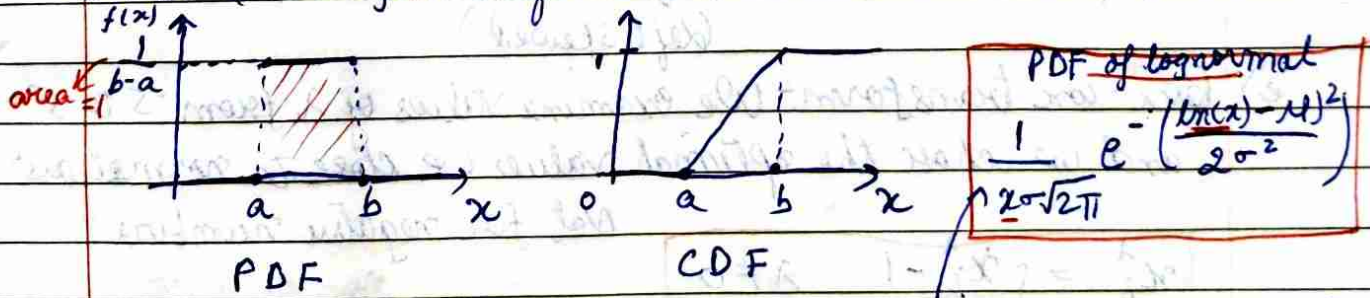
$$\text{Kurtosis} = \frac{n\times(n+1)}{(n-1)(n-2)(n-3)} \times \sum_i^n \left(\frac{x-\bar{x}}{s}\right)^4 - \frac{3\times(n-1)^2}{(n-2)(n-3)}$$



↳ QQ plot : (quantile-quantile) plots checks for distribution of data how similar it is. Take a column X and creates its 1st quantile, 2nd ... 99th quantile and do the same for a normal data. Now draw a scatterplot and if it fits in same line then data is normal and gaussian in nature.
It can use any other distribution and can compare how similar it is to our data.

↳ Uniform distribution : All outcomes are equal within range
Uniform → Continous / Discrete → dice {1, 2, 3, 4, 5, 6}
Denoted as $X \sim U(a, b)$ → parameters: a = lower, b = higher

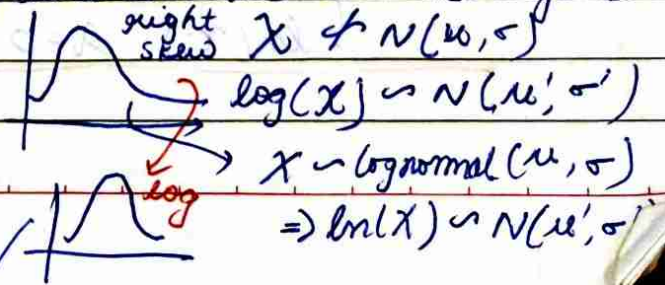
↳ PDF & CDF for uniform distribution (skewness = 0)



PDF of lognormal
$$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\left(\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)}$$

↳ Log normal distribut^n : Heavy tailed continious PD whose logs is normally distributed.

* Distribut^n jiska log is normal is called as log-normal

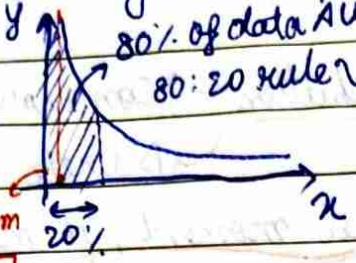


Say if we have right skew data apply log & the value one normal ✓

↳ **Pareto distribution** :- Exhibits power-law behaviour
Power law where x & y are exponentially proportional

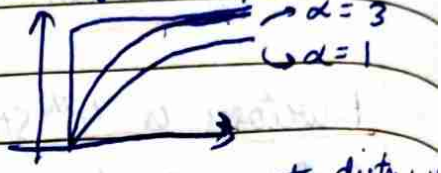
$y = k \cdot x^{a}$

$\alpha \uparrow$ peak ↑ tail ↓
$\alpha \downarrow$ peak ↓ tail ↑

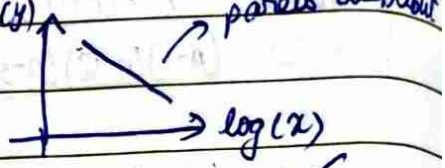


80:20 rule → Not always true

Ex wealth distribution



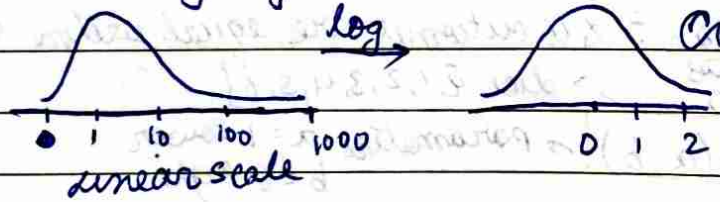
$$PDF = y = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$$



Q) How to detect distribut^{n} is pareto ?
Take a graph of log(y) v/s log(x) is straight line ✓
We can also use QQ plot take our reference as pareto plot

↳ **Transformation** :- To convert data into normal distribut^{n}
f^{n} transf → Log transform, reciprocal transform, power (Sq, Sqrt),
power tx → Box-cox transform, Yeo-johnson transform.

a) **Log transform** :- Take log of values, if data is right skew
On taking log we break the interval into smaller ones.



Can't use log on (-)ve value

b) Reciprocal ($1/x$) (c) Square (x^2) (d) Sqrt ($\sqrt{x}$)
(Left skewed)

e) **Box-cox transform** :- We examine values of λ from -5 to 5 and we chose the optimal value i.e close to normal dist
Not for negative numbers.

$$x_i^{\lambda} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln(x_i) & \lambda = 0 \end{cases}$$

f) Yeo-Johnson transform :- Adjusts box-cox & can apply on (-)ve numb

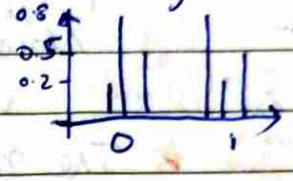
$$x_i^{\lambda} = \begin{cases} [(x_i+1)^{\lambda} - 1]/\lambda & \text{if } \lambda \neq 0,\ x_i \geq 0 \\ \ln(x_i)+1 & \text{if } \lambda = 0,\ x_i \geq 0 \\ -[(-x_i+1)^{2-\lambda} - 1]/(2-\lambda) & \text{if } \lambda \neq 2,\ x_i < 0 \\ -\ln(-x_i+1) & \text{if } \lambda = 2,\ x_i < 0 \end{cases}$$

↳ Bernoulli Distribution :- Binary outcome, either success or failure

PMF = $P(X=x) = p^x(1-p)^{1-x}$; p = prob of success

$P(x=0) = 0.2$ & $P(x=1) = 0.8$; 1-p = prob of failure

$P(x=0) = 0.5$ & $P(x=1) = 0.5$



for binary classificat^n like spam/ham, fraud etc falls in Bernoulli

↳ Binomial distribution :- Number of sucesses in a fixed number of independent bernoulli trials, with 2 possible outcomes (sucess & failure) where probability of sucess is constant for each trial. (tossing coin 3 times (independent events))

It has 2 parameters n → no. of trials & p → probability of sucess.

8) Probability of watching lecture is 0.5 (3 ppl) → $\frac{2}{Y/N}\frac{2}{}\frac{2}{} = 2^3 = 8$ outcomes

No-one watches out of 3 people = 1/8 (NNN)

1 out of 3 watch = 3/8 [YNN, NNY, NYN]

2 out of 3 watch = 3/8 [YYN, N, YY, YNY]

3 out of 3 watch = 1/8 [YYY]

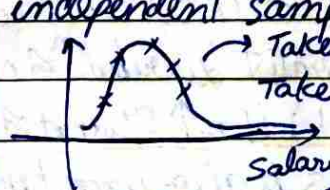
But what if we cant get the sample space we will use the formula

PDF of Binomial = $P(X=x) = {}^nC_x p^x (1-p)^{n-x}$

n = no. of trials
p = prob. of sucess
x = desired result

→ ${}^3C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = \frac{3}{8}$

↳ Sampling distribution :- It is PD that describes the statistical property of a sample statistic (such as mean, variance) for multiple independent samples of same size of proportion. Estimate the variability of a sample statistic



Take 50 samples (Random) → $\bar{x}_1$

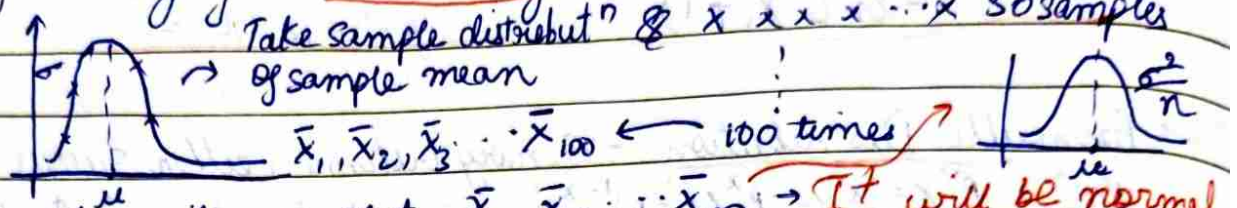
Take another 50 sample → $\bar{x}_2$

Salary 100th time 50 sample → $\bar{x}_{100}$

$\therefore \bar{x}_1, \bar{x}_2, \bar{x}_3 \cdots \bar{x}_{100}$ → Sampling distribut^n of sample mean

Take mean/Variance

★↳ Central limit theorem :- CLT states distribution of sample means of large no. of independent & identical distributed random variable will approach to normal distribution, regardless of underlying distribution of the variables.



Now if we plot $\bar{x}_1, \bar{x}_2 \cdots \bar{x}_{100}$ → It will be normal irrespective of original distribut^n, skew, lognormal, poisson etc

$\bar{x}_1$ → point estimate (single value)

★ The mean will be the same, the variance will reduce by a factor of n i.e no. of samples. σ^2/n [sample $n > 30$] ★

↳ Now we cannot say that μ of sample mean is same or approx to population mean. We can say about our confidence interval. and say we are 95% sure that our mean will lies between $[\mu - 2\sigma, \mu + 2\sigma]$ as this holds 95% of our value so we are sure about mean.

confidence level = 95%

↳ Ways to calculate CI :- → Z-procedure [pop → std available σ]
→ t-procedure [pop → Std X available not

Confidence intervals is created for parameters (population) & not statistics (sample)
Statistics help us get the confidence interval for a parameter

↳ Find confidence interval using Z-procedure :-

Condit^n → Z-procedure formula used where data is randomly selected, Standard deviation is known for a populat^n, data is normal.

$$CI = \text{point estimate } (\bar{x}) \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$Z_{\alpha/2} = 1.96$ for 95% confidence (Critical Value)

$1-\alpha$ = confidence level
$1-\alpha = 95\%$.
σ = Std populat^n
n = sample size => say 100

★↳ t-procedure :- (table at 95% generally)

$$CI = \bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$\bar{x}$ → Point est. sample mean
s → sample Std
n → no. of sample

It follows student's T distribut^n ≈ Normal
parameter is DOF = $n-1$; if $n \uparrow$ Student ≈ Normal
It has fatter tails, as 's' is also uncertain so range increases of CI ; $t_{\frac{\alpha}{2}} > Z_{\frac{\alpha}{2}}$ But $n \uparrow$ $t_{\frac{\alpha}{2}} \approx Z_{\frac{\alpha}{2}}$

↳ Hypothesis testing :- A method of statistical inference used to decide whether the data at hand supports sufficiently support a particular hypothesis. It allows us to make probabilistic statements about population parameter.

Eg Lays has green packet chips if they substitute it with Say purple so will there sales ↑, what is the prob of this hypothesis

H_0 ↳ Null hypothesis (Status quo) :- It is a statement which assumes no significant change. It serves as starting point. It is the assumption of no effect untill proven otherwise i.e lawyer proves that he is the guilty one & null hypothesis (H_0) say he is not guilty / not Committed crime. The purpose of hypothesis testing is to gather evidence to reject or fail null hypothesis and prove in the favour of alternative hypothesis (H_1 or H_a)

(H_1 or H_a) ↳ Alternative hypothesis (Research hypothesis H_1 or H_a) :- It contradicts null hypothesis and claims significant result. We need to contradict and prove the man is accused. Always any one of them is true either null or alternative hypothesis.
We try to gather evidence to reject the null hypothesis.
⇒ Failing to reject null hypothesis does mean null hypothesis is true. We couldn't gather much evidence to prove him accused. That does not mean he is completely not guilty.

↳ Rejection region approach :-

1. Formulate a Null and alternate hypothesis
2. Select significant lvl (0.05 (95%) or 0.01 (99%)) of rejecting null hypothesis
3. Check assumptions (example distribution) → Normal, lognormal, category
4. Decide appropriate test (Z-test, T-test, Chi-square test, ANOVA)
5. State the relevant test statistic → If (-) pop is known; ↳ for categorical columns
6. Conduct the test
7. Reject or not reject null hypothesis.
8. Interpret the result :

Q) Suppose company has training program. Avg. productivity before program is 50 units with a std. of 5 units. After programme we took sample of 30 employees and productivity increased by 53 units. How significantly productivity has increased?

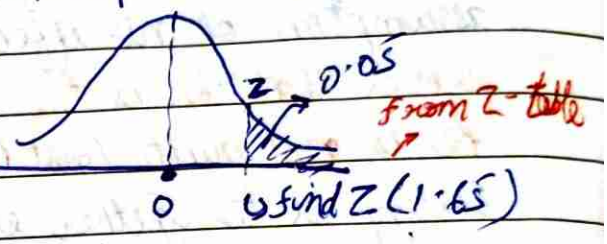
↳ 1) $H_0 = \mu = 50$ → Null hypothesis $\bar{x} = 53$ $\sigma = 50$
$H_a = \mu > 50$ → Alternate hypothesis $\mu = 50$ $n = 30$

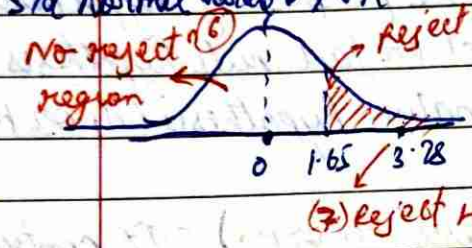
2) $\alpha = 0.05 \rightarrow 5\%$.

3) Normality is valid as sample ≥ 30 / pop std. (σ) is known → Z-test

4) Z-test

5) $Z = \frac{\bar{x} - \mu}{\text{Std normal var } \sigma/\sqrt{n}} = \frac{53 - 50}{5/\sqrt{30}} = 3.28$





As the $\bar{x}$ has fallen in reject region so we can reject the null hypothesis with 95% confidence

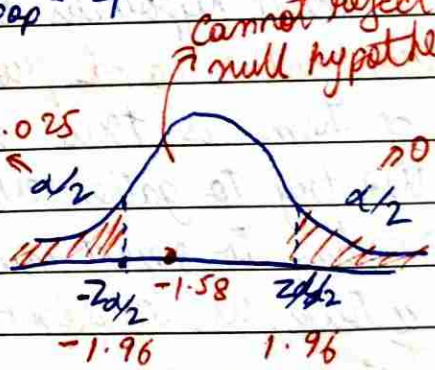
Q) $\mu = 50$, $n = 40$, $\bar{x} = 49$, $\sigma_{pop} = 4$.

1) $H_0 = \mu = 50$, $H_a = \mu \neq 50$

2) $\alpha = 0.05$

3) Normality ✓ σ ✓ Ztest ✓

4) Z test, $z = \frac{49 - 50}{4/\sqrt{40}} = -1.58$



↳ Type I v/s Type 2 Error: Hypothesis H_0 X → ✓

Type I is false positive; rejects H_0 when ✓ H_0 ✓ → X

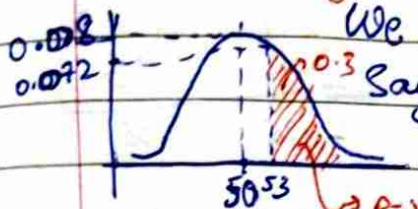
(β) Type 2 is false negative, accepts H_0 when X

↳ Prob. of committing Type II error is β

Truth / Decision	H_0 true	H_a true
Reject H_0	Type I	✓
Accept H_0	✓	Type II

↳ Hypothesis testing is used to compare ML models using CV-folds
Used in feature selection using t-test, Chi^2, ANOVA, hyperparameter

↳ **p-value** - probability of getting a sample as more extreme (having more evidence against H_0) than our own sample given the null hypothesis is true. It measures the strength of evidence against the Null Hypothesis.

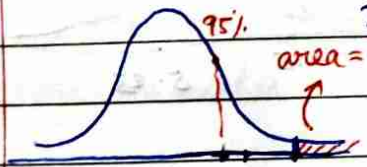


We are flipping a coin 100 times H_0: $P(H) = P(T)$
Say we did exp again and H_a: $P(H) > P(T)$
we get 53 heads
↳ p-value of this experiment is $P(H>53)$
So we can say 30% chance that getting $(H>53)$

★ If we have p-value $\leq \alpha$ we reject your null hypothesis Here α is generally 0.05 so $0.3 \nleq 0.05$ so we dont have enough evidence to reject our null hypothesis. If $P(H>80)$ we get p-value ~ 0 so we can reject the H_0 and say coin is fair.

↳ Without significance value (α):-
Very small p-value ($p < 0.01$) strong evidence against H_0
Small p-value ($0.01 < p < 0.05$) moderate evidence against H_0
Large p-value ($0.05 \leq p < 0.1$) weak evidence against H_0
Very large p-value ($p \geq 0.1$) weak or no evidence against H_0

Q) P-Value in context of Z-test
$\mu = 50$, $n = 30$, $\bar{x} = 53$, $\sigma_{pop} = 4$; find if there's significant ↑?
$\therefore H_0: \mu = 50$ & $H_a: \mu > 50$, $\alpha = 0.05$ (one tailed Test)



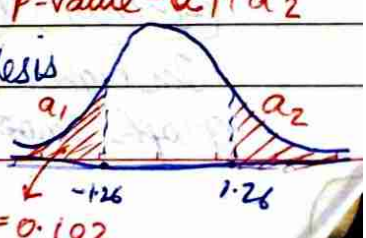
$$Z\text{-stat} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{53-50}{4/\sqrt{30}} = 4.10$$

Z_{table} gives area on left
$\therefore$ p-value is $1 - Z_{table}(4.1) = 0.0001 = $ p-value
$\therefore$ p-value < 0.05 so we can reject our null hypothesis.

Q) $\mu = 50$, $n = 40$, $\bar{x} = 49$, $\sigma = 5$, $\alpha = 0.05$ | $H_0: \mu = 50$ & $H_a: \mu \neq 50$ (2-tailed test)

$$Z = \frac{49-50}{5/\sqrt{40}} = -1.26$$

$\therefore$ p-value $> \alpha$
Can't reject our null hypothesis
$\therefore$ pvalue $= 0.206$

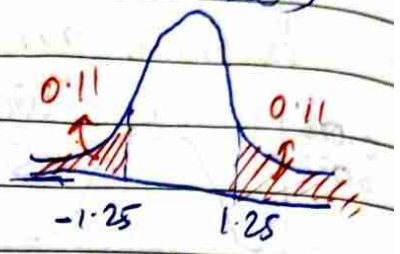


Q) $\mu = 50$, $n = 25$, $\bar{X} = 49.7$, $s = 1.2$, $\alpha = 0.05$ (pop σ ✗)

$H_0 : \mu = 50$, $H_a : \mu \neq 50$ (std of sample)

Assuming sample is normal; we use t-test here as(s)

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{49.7 - 50}{1.2/\sqrt{25}} = -1.25$$



$dof = n - 1 = 24$

t-value = 0.22

$\therefore 0.22 > 0.05$ ∴ Can't reject null hypothesis H_0

★ Shapiro-wilk test says that a sample follows normal or not

from scipy.stats import shapiro → returns statistic & p-value

if p-value < 0.05 not normal | p-value > 0.05 normal

↳ Independent 2 sample t-test

Like we can test if males avg. age > or < females avg age

↳ Independent obs., normally distributed, equal variance (homoscedasticity)

if $\sigma_A^2 \neq \sigma_B^2$ checked by F-test, if $\sigma_A^2 \neq \sigma_B^2$ use Welch's t-test ($\sigma_A^2 = \sigma_B^2$)

↳ Here σ_A, σ_B not required

Q) $H_0 : \mu_d = \mu_m$

$H_a : \mu_d \neq \mu_m$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

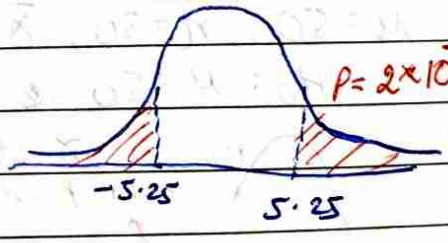
$n_1 = 30$ $n_2 = 30$

$\bar{x}_1 = 18.5$ $\bar{x}_2 = 14.3$

$\sigma_1 = 3.5$ $\sigma_2 = 2.7$

$dof = n_1 + n_2 - 2 = 58$

$$t = \frac{18.5 - 14.3}{\sqrt{(3.5)^2/30 + (2.7)^2/30}} = 5.25$$



Rejecting our null hypothesis as time spend on mobile ≠ desktop

↳ Chi-square distribution : If we take sum of squares of std random normal variable values is χ^2 (chi square), continuous P.D.

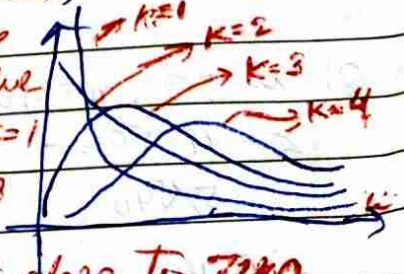
$Z_1 \sim N(0,1)$ ∴ $Q_1 = Z_1^2 = \chi^2_{1=k}$

$Z_2 \sim N(0,1)$ $Q_2 = Z_1^2 + Z_2^2 = \chi^2_{2=k}$

$Z_3 \sim N(0,1)$ $Q_3 = Z_1^2 + Z_2^2 + Z_3^2 = \chi^2_3$

On higher DOF = k-1 the graph smoothens ↳ k

Sum of squares of a normal value will ≈ 0 if k=1 as major values lie b/w -1 & 1 when square it they to zero



No. of categories

↳ Chi-square test ÷ We use this test to check for significant non-parametric association between categorical variables we can check goodness of fit test as well if observed distribution of a single categorical variable matches an expected theoretical distribut^n.

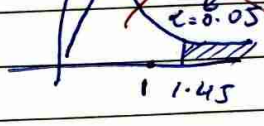
Q) Dice is rolled 60 times and we need to check wether dice is fair

H_0: dice is fair → uniform distribut^n ↳	10	10	10	10	10	10
H_1: dice is unfair → Observed distribut^n ↳	12	8	11	9	10	10

Formula $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ (obs, exp) = Here $k = 6$ as 6 out comes

$\alpha = 0.05$ $\therefore DOF = K - 1 = 5$

$\chi^2 = \frac{4+4+1+1}{10} = 1$



We will check χ^2 values for DOF=5 from chi-table

We cannot reject H_0

Dice is fair

F-distribution ÷ Continuous PD used in statistical hypothesis and analysis of variance (ANOVA). It has 2 parameters dof, & dof2 Its left bound is at zero, (+)ve skewed. Used to compare fits of different statistical models. F-statistic is calculated by dividing the ratio of two sample variance or mean squares from ANOVA table Value is compared from F-distribution for its confidence.

↳ One way ANOVA test ÷

H_0 = all the group means is Equal

H_1 = Atleast one group mean is diff

Now its just a creation of big table which computes b/w sample or categories & within sample or categories

$\chi_1^2 \rightarrow df_1$

$\chi_2^2 \rightarrow df_2$

$F = \frac{\chi_1^2/d_1}{\chi_2^2/d_2}$