Statistics :- science of collecting, organizing and analyzing data.

⇓

Decision Making

Data - Facts or pieces of information ← Measured, collected, Analyzed

Eg: Weight of students , IQ of students

Eg: House Price Dataset

City	Area	No of Rooms	Price
Hamburg	1000	2	450 €
Berlin	1250	2.5	500 €

Price: Analyze

Data scientist → Model → Price

Data Analyst → Report → Visualization - Meaningful Decision

Importance of statistics

1) Data Exploration
2) Model Building and Validation
3) Statistical Analysis → [Sample data ⇒ Conclusions]
4) Hypothesis Testing
5) Optimization and efficiency.
6) Report Making

Types of Statistics

1) Descriptive Statistics :- Descriptive statistics involve methods for summarizing and organizing data to make it understandable. This type of statistics helps to describe the basic features of data in a study.

1. Measure of Central Tendency
   mean, median, mode

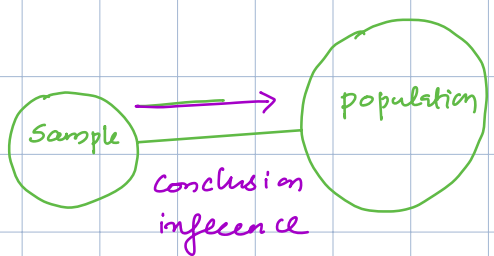
2. Measure of Dispersion
   Variance, Standard Deviation

3. Data Distribution
   Histogram, boxplot, Pie chart, PDF, PMF

4 Summary Statistics

Five number Summary → Q_1, Q_2, Q_3, Max Value

2) Inferential Statistics : Inferential statistics involves methods for making predictions or inferences about a population based on a sample of data. It allows for hypothesis testing, estimation and drawing conclusions.



1) Hypothesis Testing
2) P Value
3) Confidence Interval
4) Statistical analysis Test
   1. z test
   2. t test
   3. ANOVA → F test
   4. Chi Square
   5. Regression Analysis

Example

Let say there are twenty statistics class in your university. You have collected the height of students in the class

Heights are recordeds as [175cm, 135cm, 180cm, 140cm, 120cm]

Descriptive Question - What is the average height of entire class ?

⇓

Mean = measure of central Tendency

Inferential Questions . Are the height of the sample students in class room similar to what you expect from the entire university?

⇓

Inferential

Population and Sample Data

Population | Sample Data

entire set of individuals or objects of interest in a particular study. It includes all members of defined group that we are studying or collecting information on

characteristics

1) complete set : contains all the observation of interest

2) parameter : A numerical value summarizing the entire population.

Eg: [population mean μ, population variance σ^2]

Example 1) population in school study

* All students in school
* Determine the avg height of students = population mean

2) Population in market research

*) All consumers in a city

*) To understand the purchasing behaviour of all consumers

3) Population in a medical study

* All patients with a specific disease
* To study the effectiveness of a drug

subset of the population that is used to represent entire group.

Sampling - selecting a group of individuals / obs from the population to draw conclusion

characteristics

1) subset: Represent a portion of the population

2) statistic : A numerical value summarizing the sample data

[sample mean, sample variance]

3) Random Sampling: samples should be randomly selected to avoid bias

Examples

1) Sample in school study

* group of 50 students from school

used case = estimate the avg height of students in a school

2) Market Research sample

* 500 consumers
* Behavior $\Rightarrow$ population generalization

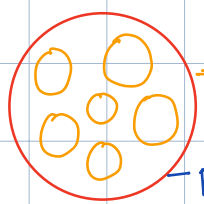
3) Medical Study sample

* 200 patients

* test effectiveness of the drug.

(4) Types of Sampling Techniques

eg:- After election an exit poll is conducted



→ Take samples from each region

A, B, C, D = Media

population in state

Ask for whom do the vote

Tendative study

1) Probability Sampling
2) Non probability sampling

1. Probability Sampling

a) simple random sampling

every member of the population has an equal chance of being selected

eg: selecting sample randomly

drawing names randomly from a class of students

b) systematic sampling

select every n^{th} member of the population after a random starting point.

Airport —— Credit card shops ⟶ Every 5th person they pitch for credit card (5^{th}, 10^{th}, 15^{th}, ...)

Feedback Survey ⟶ selected every 11^{th} member ↓ Droped mail for feedback.

c) Stratified sampling

Divide the population into strata (groups) based on specific characteristics and then randomly samplin from each strata

Eg: ① Divide the employees by department and then randomly select a proportional number from each department to form a survey sample

② Population divided on age group

< 12 | 12 – 18 | > 18 } survey on food

d) Cluster Sampling

Divide the population into clusters, randomly selecting clusters then sampling all the members

from the selected clusters.

Eg: Randomly selected several schools from a district and surveying all teachers within those schools.

e) Multi Stage Sampling

combining several sampling methods,

selecting clusters → randomly sampling within those clusters

Eg:- Randomly selecting cities, each selected city randomly selecting households to survey.

2. Non probability Sampling

Select individuals who are easiest to reach.

Eg: surveying people at mall

ⓐ Convenience sampling

select individuals who are easiest to reach

ⓑ Judgemental (Purposive) Sampling

Select individual based on the researcher's judgement ⟶ Useful or Representative

Eg: choose experts in a field to participate {Data Science}

(c) Snowball sampling

Existing study subjects recruit future subjects from among their acquitances

Eg: survey members of a rare disease

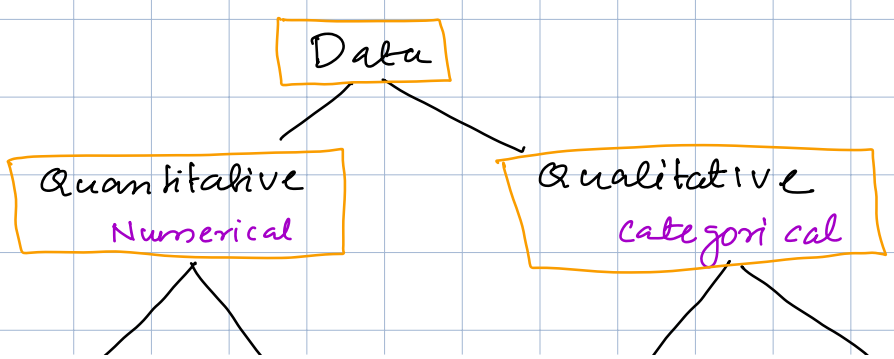
d) Quota Sampling

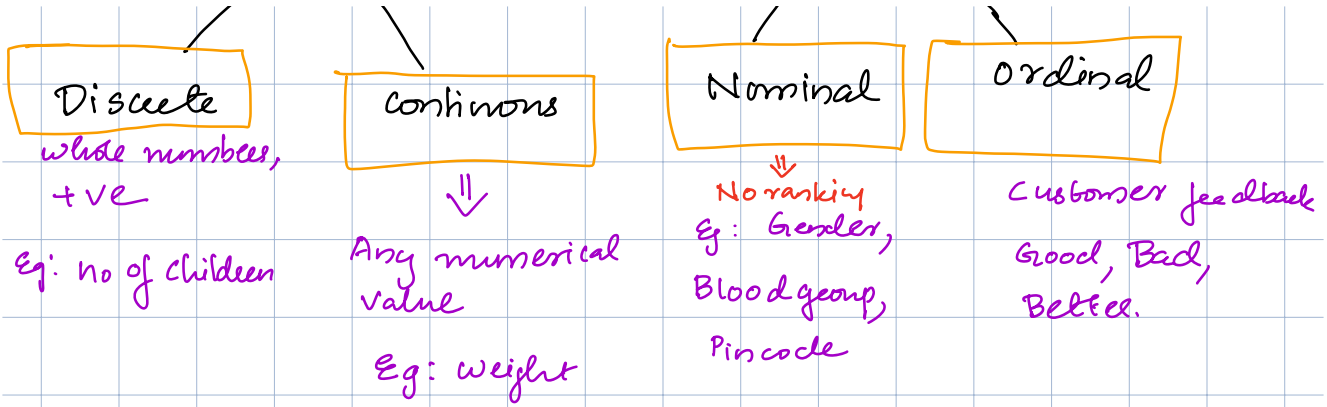
Age, group, gender, caste.

Selecting the sampling technique depends on use cases.

(5) Types of Data

Data is of different type → important while analysis and model selection





⑥ Scales of measurement of Data

The nature of information within the value assigned to variables

4 Primary Scales of Measurements

1) Nominal scale
2) Ordinal scale
3) Interval
4) Ratio

1) Nominal scale - classify data into distinct categories that do not have an intrinsic order.

Qualitative / Categorical data

Characteristics

1) data is categorized based on labels, names or qualities

2) These categories are mutually exclusive

3) No logical order among categories [No rank]

Eg: Gender M, F

Color R, B, P, G

Types of Cuisine Italian, Indian, Mexican

(2) Ordinal Scale

classifies the data into categories that can be ranked or ordered

Characteristics

i) data is categorized and ranked in a specific order

ii) Interval between Ranks are not necessarily equal.

Example :- (1) Education level ← high school (1), Bachelors (2), Masters (3), Doctorate (4) ⇒ Feature engineering ⇒ Ranks

(2) Customer feedback ← satisfied (2), very satisfied (1)

not satisfied (3)

(3) socio economic status
- low 0
- middle 1
- high 2

(3) Interval scale

The interval scale not only categorizes and orders but also specify the exact difference between intervals. It lacks a true zero point

characteristics

1) Data is ordered with consistent interval between values
2) Allows for meaningful comparison of differences
3) No true zero point.

(ratio cannot be measured)

Example: Temperature in Fahrenheit

$10^\circ F$, $20^\circ F$, $30^\circ F$

$20 - 10 = 10$

$30 - 20 = 10$

⇒ differences same

Temp $0^\circ F$ not possible

IQ scores

90, 100, 110

10 10

IQ 0 is not possible

Calender years

2024, 2020, 2016,

0 year is not defined

4) Ratio Scale

① The order matters

② Differences are measurable

③ contains a zero starting point

Eg: student marks in a class

0, 90, 60, 30, 75, 45

ASC = 0, 30, 45, 60, 75, 90

15

30

Ratio = $90/30$ = 3:1

Assignment

① length of different rivers in world ⇒ Interval

② favourite food based on gender ⇒ Ordinal

③ Marital status = Nominal

④ IQ measurement = Ratio

Measure of Central Tendency

Measures of central tendency are statistical metrics that describe the center point or typical value of dataset. They provide a single value that summarizes a set of data by identifying the central position within that dataset.

① Mean or Average

② Median

③ Mode

Eg: Ages = [24, 32, 12, 48, 16, 20]

① Mean

Mean is sum of all values divided by the number of values.

Population mean (μ)

population (N)

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$ N = population size

Here X is a random variable

$$X = \{5, 8, 12, 15, 20\}$$

$$N = 5$$

$$\mu = \frac{60}{5} = 12$$

Sample mean ($\bar{x}$)

Sample (n)

$n \leq N$

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

$n \rightarrow$ sample size

Characteristics

- Mean is usually affected by extreme outliers
- Usually used for interval and ratio data.

$$X = \{1, 2, 3, 4, 5\}$$

$$\mu = \frac{1+2+3+4+5}{5} = 3$$

$$X' = \{1, 2, 3, 4, 5, 100\}$$

$$\mu = \frac{115}{6} = 19.1$$

Omitting outliers cause information loss

(2) Median

Middle value in a dataset when the values are arranged in either ascending or descending order.

$X = \{1, 2, 3, 4, 5\}$

No of elements = 5

5 is odd

Middle element = 3 = Median

$X' = \{1, 2, 3, 4, 5, 100\}$

Median = 3.5

No much change

Characteristics

- Not affected by extreme outliers
- Used specifically for ordinal, interval and ratio data

(3) Mode

Appears most frequently in a dataset

Dataset - 2, 4, 4, 6, 7, 7, 7, 9

	frquency
2	1
4	2
6	1
7	3

some datasets have multiple modes

2 modes - bimodal

more modes - multimodal

Characteristics

① Not affected by outliers

② Used for all the four scales nominal, ordinal, interval, ratio

Choosing the Appropriate Measure

Mean = data when symmetrical and without outliers, provide some mathematical average which is useful for further statistical calculations.

Median - Best used when data is skewed or contains outliers. Provides the middle value, which better represents the center of a skewed dataset

Mode :- Best used for categorical data to identify the most common category. Also

Useful for identifying the most frequent value in ordinal, interval or ratio data.

Real World Applications

Feature Engineering

Age,	Weight,	Salary	Gender	Degree
24	70	40k	M	BE
25	80	70k	F	—
27	95	45k	F	—
24	—	50k	M	PHD
32	—	60k	—	BE
—	60	—	—	Master
—	65	55k	—	BSC
40	72	—	M	B.E

① Handling Missing Values

↓

if Mean ⇒ outliers create problem

Median ⇒ more efficient

Interval scale, Ratio Data

Categorical ⇒ Mode Nominal

Mode Nominal & Ordinal

Measure of Dispersion

Measure of dispersion describe the spread or variability of dataset. They indicate how much the values in a dataset differ from the central tendency

Common Measure of Dispersion

1. Range
2. Variance
3. Standard Deviation
4. Interquartile Range (IQR)

1. Range

Defⁿ :- Difference between the maximum value and minimum values in a dataset.

Range = Max Value − Min Value

Eg: Ages {14, 13, 10, 20, 25, 75, 15}

Range = 75 − 10 = 65

Characteristics

1) simple to calculate
2) sensitive to outliers
3) Provides only rough measure of dispersion.

Weight = {35, 40, 45, 39, 30 ... 70}

Range without 70 = 40 − 30 = 10

Range with 70 = 70 − 30 = 40

(2) Variance

Defn :- Variance measures the average squared deviation of each value from the mean. It provides sense of how much the values in a dataset vary

Population Variance

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

$x_i \Rightarrow$ DATA POINTS

μ = Population mean

N = population size

Sample Variance

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

x_i = Data points

$\bar{x}$ = Sample mean

n = Sample size

Example → size of a flower petals

$\{5, 8, 12, 15, 20\}$ ⇒ Variance of this distribution

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

$$\mu = \frac{5+8+12+15+20}{5} = 12$$

$$\text{Variance} = \frac{7^2 + 6^2 + 0^2 + 3^2 + 8^2}{5} = \frac{49+36+9+64}{5}$$

$$= 27.6 = \sigma^2$$

Characteristics

⇒ Provides a precise measure of variability

⇒ units are squared of the original data unit

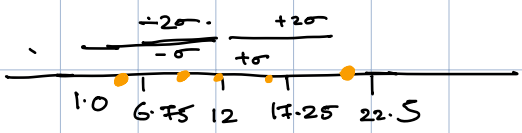
⇒ More sensitive to outliers

(3) Standard Deviation

Defⁿ:- The standard deviation is the square root of the variance

$$\sigma = \sqrt{27.6} \approx 5.25$$

{5, 8, 12, 15, 20



Z-score
↳ standard normal distribution

Characteristics

① provides a clear measure of spread in same units as the data

② Also sensitive to outliers

key Differences and similarities

Relationship

Standard deviation is the square root of variance. If you have the variance, you can find the standard deviation by taking the square root of the variance

Conversely if you have the standard deviation, you can find the variance by squaring the standard deviation.

Units :

Variance :- The unit of variance = square of unit of original data

- provides measure of dispersion of data points in squared units, which is difficult to interpret directly

Standard deviation = same unit as original data.

= provides a measure of dispersion in the same units as the original data, making it easier to interpret and understand.

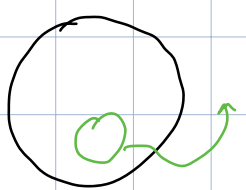
Sample Variance

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

Population Variance

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

$n-1$ = Bassel correction



Inferences

$\bar{x} \approx \mu$
$s^2 \approx \sigma^2$
} Might come up case 1

$\bar{x} <<< \mu$
$s^2 <<< \sigma^2$
} case 2

if $s^2 = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n}$ then in case of case 2, we

are underestimating the true population variance

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

Based correction

degree of freedom $n-1$

Random Variables X

$y = 5x + 2$

$y = 7 \qquad x = 1$

$y = 12 \qquad x = 2$

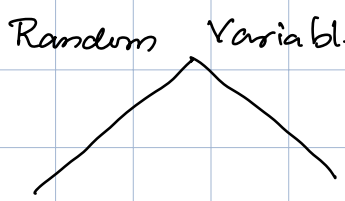
$y = 17 \qquad x = 3$

$X \rightarrow$ function whose values are derived from different process or experiment.

eg: Tossing a coin. $X = \begin{cases} 0 & H \\ 1 & T \end{cases}$

Rolling a fair dice $X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$

Different types of Random Variables



Eg: 1) Tossing coin
rolling dice

Eg: 1) Tomorrow how many inches it is going to rain
[0, 1.1, 5.5, ... 10.75]

X - function $\xrightarrow[\text{process}]{}$ value

Percentiles And Quartiles

Percentage :- {1, 2, 3, 4, 5, 6}

No of odd numbers = 3

percentage of odd numbers in this group = $\frac{3 \times 100}{6}$ = 50%

Percentiles :- A percentile is a value below which a certain percentage of observations lie.

{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10}

Percentile of value x = $\frac{\#\text{ of values below } x \times 100}{n}$

if $x = 9$ $= \frac{11}{14} \times 100$

$= 78.57\%$ of value 9

78.57% of entire distribution is less than 9

$\Rightarrow$ Value $= \frac{\text{Percentile}}{100} \times (n+1)$

if percentile = 25 Value $= \frac{25}{100} \times (15)$

$\Rightarrow \frac{1}{4} \times 15) = 3.75$

3.75 not in the data so 3rd & 4th number

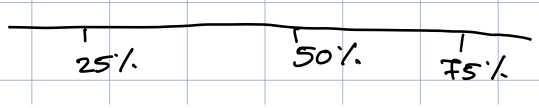
average $= \frac{3+4}{2} = \underline{\underline{3.5}} =$ 25 percentile

② Quartiles

25% = 1st Quartile

50% = 2nd quartile

75% = 3rd quartile



③ 5 Number summary and Boxplot [To find outliers]

- in python using seaborn.

1) Minimum
2) First Quartile (25 percentile) Q_1
3) Median
4) Third Quartile (75 percentile) Q_3
5) Maximum

Removing the outlier

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 6, 7, 8, 9, 29\}$

(3 is the 5th; 29 is outliers)

[Lower Fence ⟷ Higher Fence]

Lower Fence = $Q_1 - 1.5\,(IQR)$ → Inter Quartile Range

Higher Fence = $Q_3 + 1.5\,(IQR)$

Q_1 = 25 percentile $= \frac{25}{100} * (19+1) = \frac{25}{100} \times (n+1)$.

$= 5^{th}$ value

$Q_1 = 3$

$Q_3 = \frac{75}{100} \times (20) = \frac{3}{4} \times 20 = 15^{th}$ value

$= 7$

$IQR = Q_3 - Q_1 = 7 - 3 = 4$

Lower Fence $= Q_1 - (1.5)\,IQR$

$= 3 - 1.5(4) = 3 - 6 = -3$

Higher Fence $= Q_3 + 1.5(IQR)$

$= 7 + 6 = 13$

$[-3 \quad 13]$

Values outside this range is an outlier

so here outlier is 29

How do we create a box plot

Minimum = 1

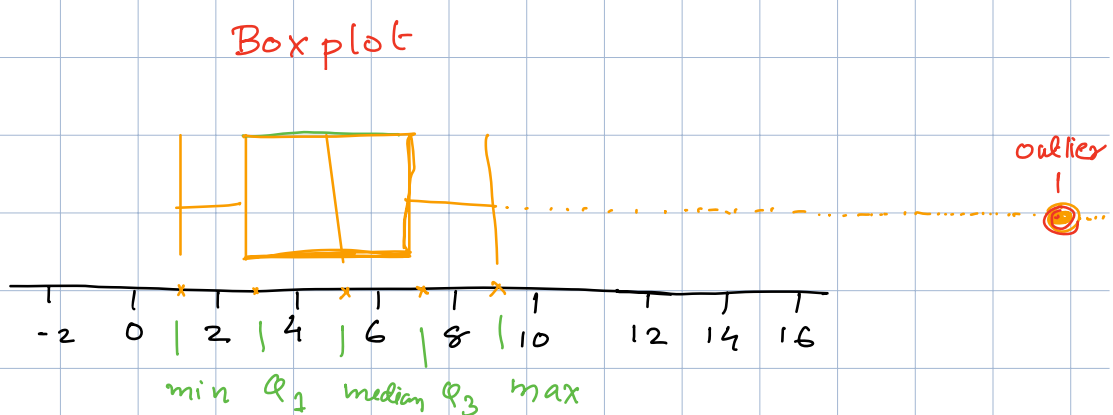
Q_1 = 3

Q_3 = 7

Median = (19 elements so 10^{th} element)

= 5

Maximum = 9 (excluding the outlier)



⇒ Outlier play a major role in data science project

Assignment

$$Y = \{-13, -12, -5, -6, 3, 4, 5, 6, 7, 7, 8, 10, 10, 11, 55\}$$

$n = 15$

$$Q_1 = \frac{25}{100} \times 16 = 4$$

$$Q_3 = \frac{75}{100} \times 16 = 12$$

$IQR = 12 - 4 = 8$

Lower Fence $= Q_1 - 1.5(8) = 4 - 1.5(8)$

$= 4 - 12 = -8$

Higher Fence $= 12 + 12 = 24$

$[-8, 24]$

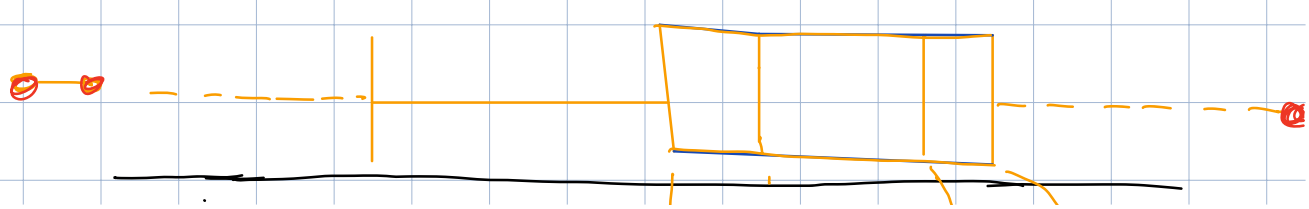
outliers $= [-13, -12, 55]$

Box plot

min $= -5$

max $= 11$

median $= 6$



-8 -6 -4 -2 0 2 | 4 6 8 10 | 12

min — Q1 — median — max — Q3

Histogram And Skewness

A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable and is used to visualize the shape, central tendency and variability of a dataset.

Age = { 11, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50 }

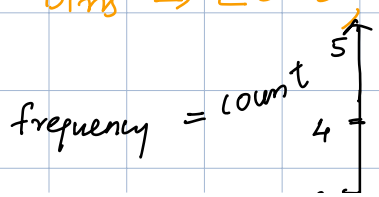
Range here = (11, 50)

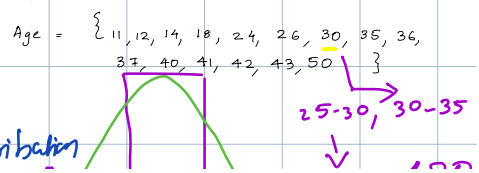
consider all the histograms between 0-50

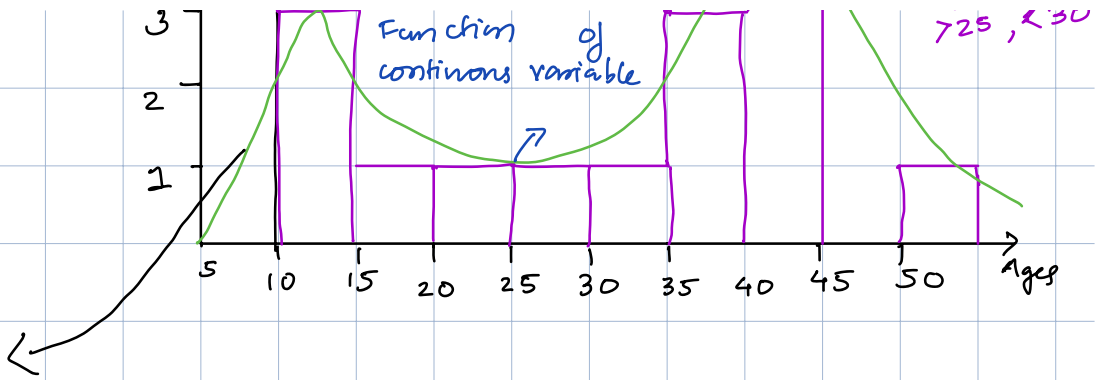
① No of bins = 10 $\frac{50}{10} \rightarrow 5$ = 5 bin size

ie 10 bins each have size will be 5.

Bins → [0-5, 5-10, 10-15, 15-20, ... 45-50]



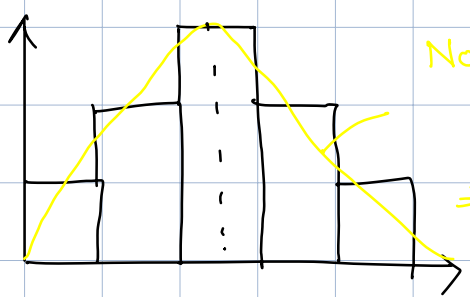




Kernel density Function.

Skewness

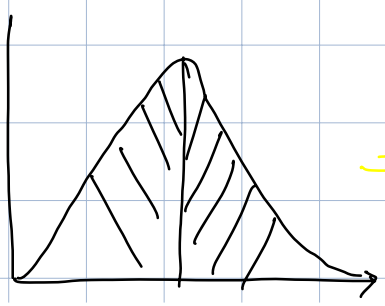
If we create a histogram



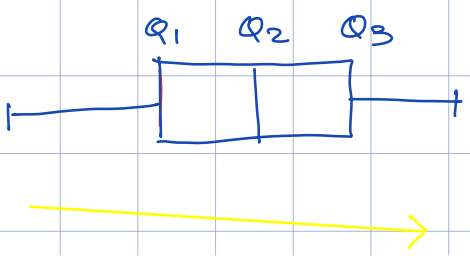
Normal / Gaussian Distribution
=> Symmetrical Distribution
Bell curve

No skewness in symmetrical distribution.

①



Boxplot

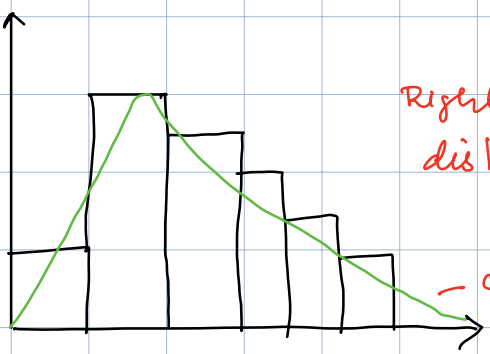


The mean, median and mode are all perfectly at the center in the symmetrical distribution

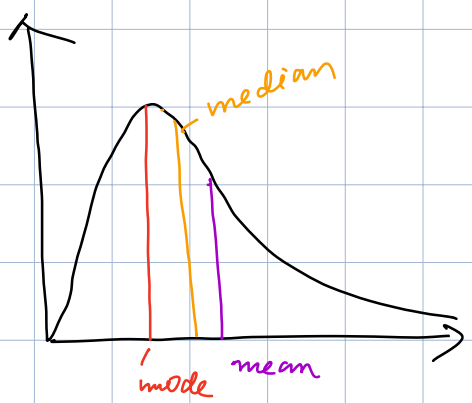
$Q_3 - Q_2 \approx Q_2 - Q_1$

Mean = Median = Mode

② Right Skewed



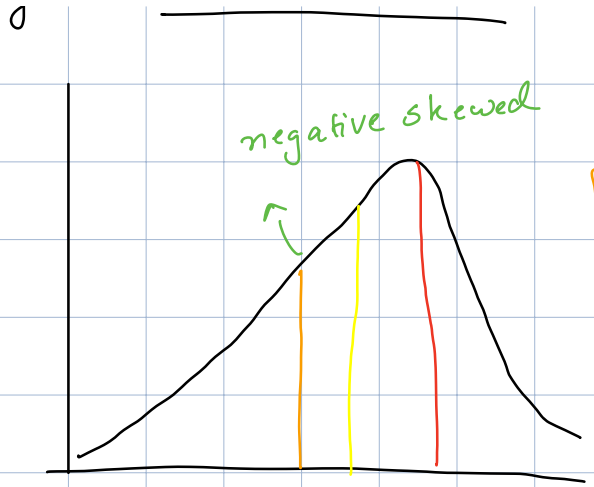
Right skewed / log Normal distribution.



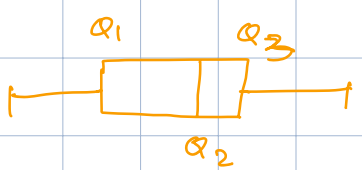
mean ≥ median ≥ mode

$Q_3 - Q_2 \geq Q_2 - Q_1$

③ Left Skewed Distribution



mean ≤ median ≤ mode



$Q_2 - Q_1 \geq Q_3 - Q_2$

Covariance and Correlation

Covariance and correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable are associated with changes in another variable.

Covariance

Covariance is a measure of how much two random variables change together. If the variables tend to increase and decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative.

to quantify the relationship between x and y

Eg :-

X	Y
2	3
4	5
6	7
8	9

X ↑ Y ↑

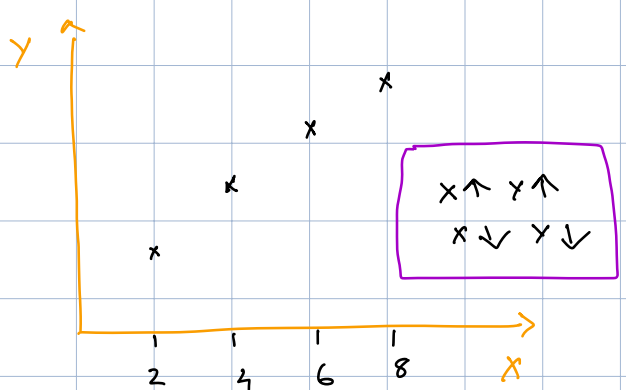
X ↓ Y ↑

X ↑ Y ↓

X ↓ Y ↓

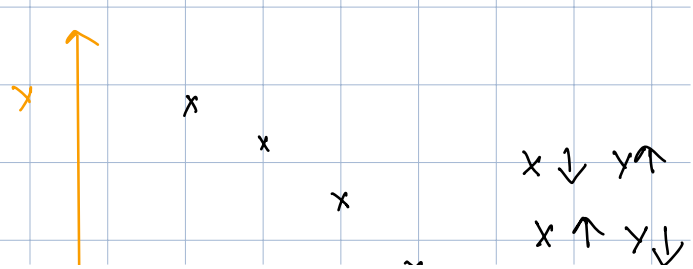
Data set

size of house	Price
1200	45000 Euro
1300	50000 Euros
1500	75000 Euro

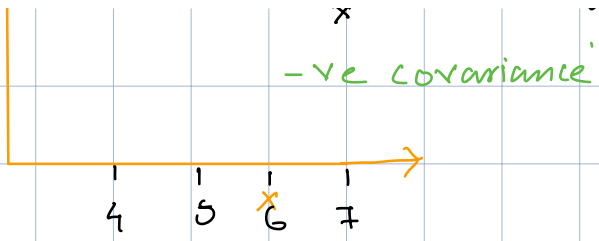


+ve covariance

X	Y
7	10



3	14
1	16



Covariance $\quad Cov(x, y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

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

⇓

Spread of data

x_i = Datapoint of random variable x

$\bar{x}$ = Sample mean of x

y_i = Datapoints of random variable y

$\bar{y}$ = Sample mean of y

Eg:

Students

Hours Studied	Exam Score (Y)
2	50
3	60
4	70

$X \uparrow \quad Y \uparrow$

X ↓ Y ↓

5	80
6	90

Calculate the Covariance

$$\bar{x} = \frac{2+3+4+5+6}{5} = \frac{20}{5} = 4$$

$$\bar{y} = \frac{50+60+70+80+90}{5} = 70$$

$$Cov(x,y) = \frac{-2(-20) + -1(-10) + 0(0) + 1(10) + 2(20)}{4}$$

$$= \frac{40+10+10+40}{4} = \frac{100}{4} = 25$$

Positive Co Variance

Indicates the number of hours studied increases the exam score.

Advantages

⇒ Quantify the relationship between X and Y

Disadvantage

① Covariance does not have a specific limit value

$(-\infty, \infty)$

limit value is not restricted

Correlation → Pearson correlation Coefficient
→ Spearman Rank Correlation.

Pearson correlation Coefficient ⇒ [-1 to 1]

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y}$$

① The more the value towards +1 the more positive correlated x and Y is.

② The more the value towards -1 the more -ve correlated it is.

For the the earlier problem

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{25}{\sigma_x \cdot \sigma_y}$$

$$\sigma_x = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{1^2+2^2+1^2+2^2}{4} = 5$$

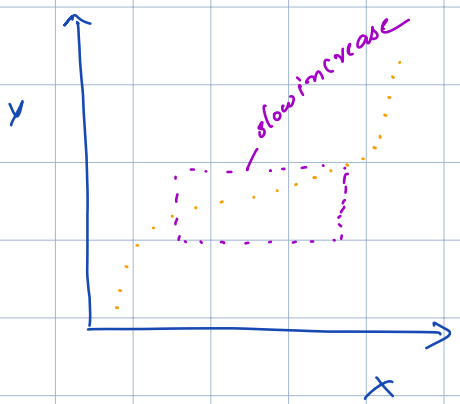
$$\sigma_y = \frac{20^2+10^2+20^2+10^2}{4} = \frac{1000}{4} = 250$$

$$\rho_{x,y} = \frac{25}{5(250)} = \frac{1}{10(5)} = \frac{1}{50} = 0.02$$

0 to 1
⇓
+ve

Spearman Rank Correlation

In pearson correlation = -1/1 in a linear
in between = scattered)



Spearman Correlation 1
pearson Correlation 0.88

Pearson correlation cannot capture relationship in nonlinear data.

Spearman Correlation $r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$ (Rank)

x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

Eg

Feature selection

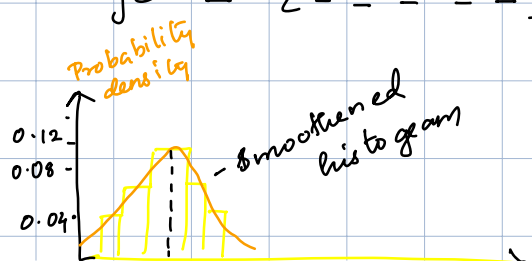
Size of House ↑	No of Room ↑	Location ↑	Haunted ↓	No. of people	Price
Important	Important	Important	Important	Not important	

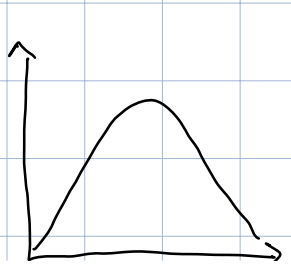
Probability Distribution Function and Types of Distribution.

Probability distribution function

Probability distribution functions describe how the probabilities are distributed over the values of a random variable.

Age = {- - - - - -}





55 Ages

PDF helps us to understand how the probabilities are distributed over random variables

2 Main type of probability distribution functions

① Probability Mass functions (PMF): used for discrete random variables

② Probability density function (PDF): used for continous random variables.

③ Cumulative density Function (CDF)

① Probability Mass Function [specifically for discrete random variable]

Eg: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$ = Fair Dice

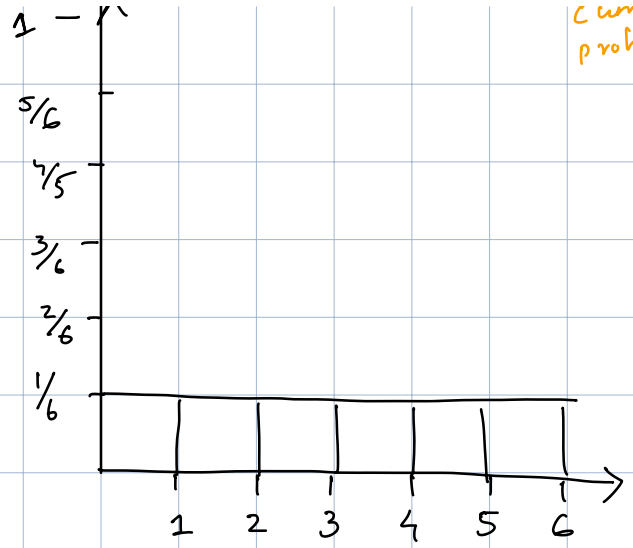
$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$$

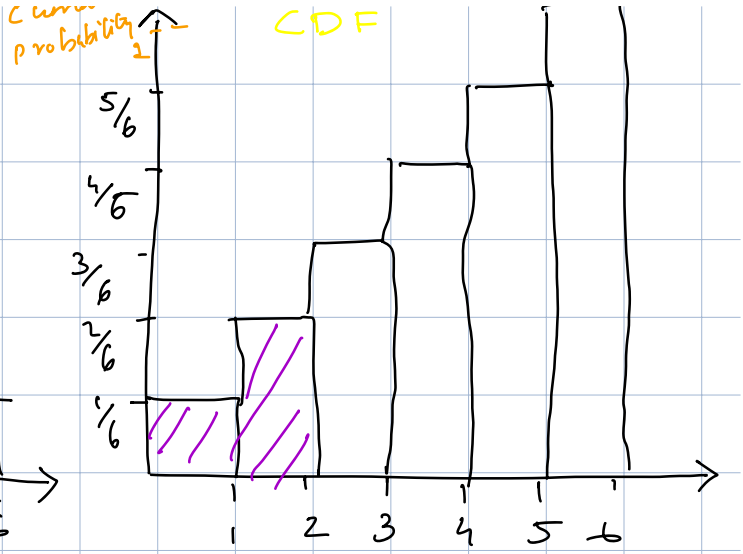
Fair dice

PMF

Probability

Cumulative density Function





combine probability as we go from 1 to 6

Why CDF?

$$P(X \le 2) = P(X=1) + P(X=2)$$

$$= P(X \le 2) = \frac{2}{6} = \frac{1}{3}$$

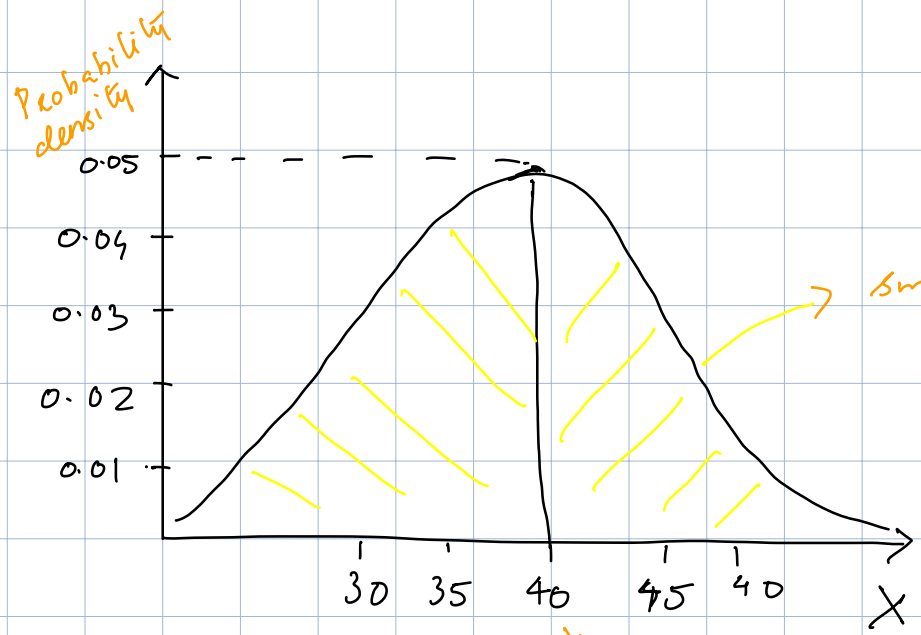
$$P(X \le 6) = P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6)$$

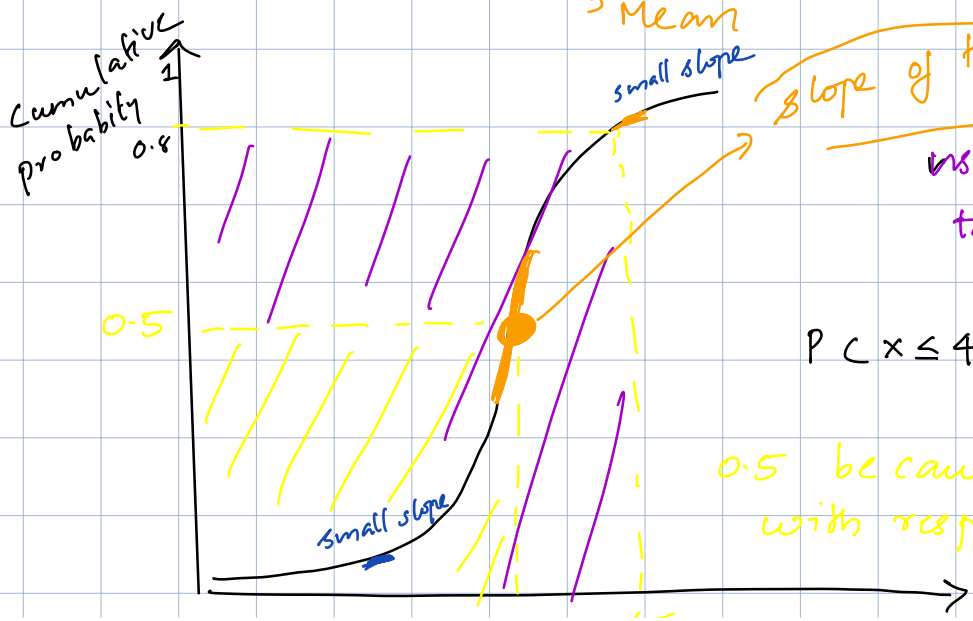
$$= 1$$

② Probability Density Function (PDF)

① Distribution of continuous random variable

X = Ages = { . . . . . }





40 45 X

① Area under the curve

② Probability density

So slope of CDF give PDF

$Pr(X \leq 45) = 0.8 = 80\%$

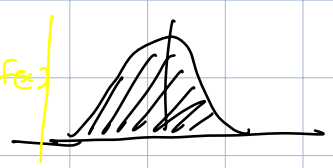
So Probability density is gradient of cumulative density function.

PDF properties ⇒

① Always non negative $f(x) \geq 0$ for all x

② The total area under the PDF curve is equal to 1

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$



With respect to different distribution fox) function is going to change.

Different distribution types [pdf, pmf, cdf]

Dataset often follows different types of distribution:

(1) Bernouli Distribution ⇒ Outcomes are binary
(pmf) = Discrete Random Variable

(2) Binomial Distribution → pmf

(3) Normal / Gaussian Distribution → pdf
- some sort of bell curve
- used in assumptions

(4) Poisson Distribution = pmf

(5) Log Normal Distribution = (pdf)

(6) Uniform Distribution = (pmf)

Eg: House Price Prediction Dataset

Size of house	No of Rooms	location	Floor	Sea side	Price
continous	Discrete		Discrete	Bernoulli 0 or 1	Continous
pdf			pmf	pmf	pdf

This step is important in explanatory data analysis and feature engineering.

(1) Bernoulli Distribution

Definition :- The Bernoulli Distribution is the simplest discrete probability distribution. It represents the probability distribution of a random variable that has exactly two

possible outcomes: success (with probability p) and failure (with probability 1-p). It is used to model binary outcomes, such as a coin flip or a yes/no question.

① Discrete Random Variable (prof)

② Outcomes are Binary

Eg: ① Tossing a coin {H, T}

$Pr(X = H) = 0.5 = p$ (H - head)

$Pr(X = T) = 1 - 0.5 = 0.5 = q$ (T - Tail)

$p, q \Rightarrow p = 1 - q \qquad q = 1 - p$

② Whether the person will Pass / Fail

$Pr(X = pass) = 0.4$

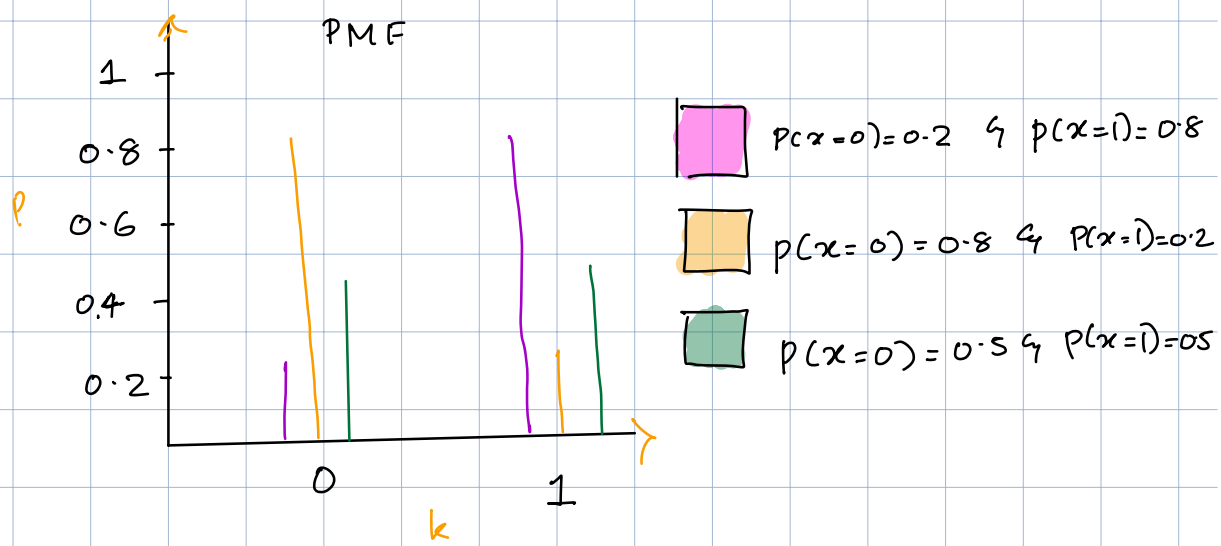
$Pr(X = Fail) = 1 - 0.4 = 0.6$

Parameters

$0 \leq p \leq 1 \qquad q = 1 - p$

$K = \{0, 1\} \Rightarrow$ 2 outcomes

Pr (Success) $\Rightarrow k=1$ Pr (Fail) $\Rightarrow k=0$



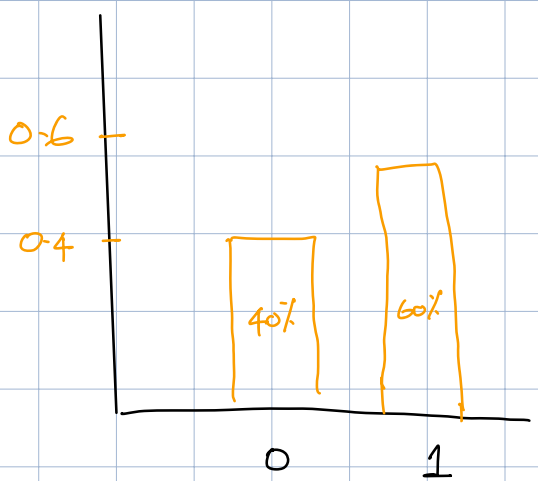
Eg

PMF :- company has launched a new smartphone 'A'

k

① use = 60%

② not use = 40%



$PMF = P^k * (1-P)^k$

if $k = 1$ $\quad p = 0.6$

ie $p(k=1) = p^1(1-p)^{1-1} = p$

$p(k=0) = p^0(1-p)^{1-0} = 1-p = q$

Simplified

$$Pmf \begin{cases} q = 1-p & \text{if } k=0 \\ p & \text{if } k=1 \end{cases}$$

Mean of Bernoulli Distribution

$k \in \{0, 1\}$.

$$E(x) = \sum_{i=0}^{k=1} k \cdot p(k)$$

$= 0 \cdot (0.4) + 1(0.6)$

$= 0 + 0.6 = 0.6 \Rightarrow p$

Median of Bernoulli Distribution

Median $\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$

Median $\begin{cases} 0 & \text{if } q > p \\ 0.5 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$

Mode

$p > q \Rightarrow$ p will be the mode

else q will be the mode.

Variance

$k = 0$ and 1 $\quad P(k=0) = 0.4$, $P(k=1) = 0.6$

$$\sigma^2 = 0.40 * (0 - 0.6)^2 + 0.6 (1 - 0.6)^2$$

$$= \sum_{i=1}^{n} p_i (x_i - \mu)^2 \quad \text{here } \mu = 0.6$$

$$= 0.4(0.36) + 0.6(0.16)$$

$$\sigma^2 = 0.24 \Rightarrow p(k=0) * p(k=1)$$

$$\sigma^2 = pqv$$

$$\sigma = \sqrt{p*q}$$

② Binomial Distribution

In probability theory and statistics the binomial distribution with parameters n and p is the discrete probability distribution of the numbers of success in a sequence of n independent experiments, each asking a yes - no question, and each with its own Boolean valued outcome: success (with probability p) or failure (with probability $q = (1-p)$. A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment and a sequence of outcomes is called a Bernoulli process; for a single trial, ie. $n=1$, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

③ Discrete Random Variable

(*) Every outcome of the experiment is binary

(*) These experiments are performed for n trials

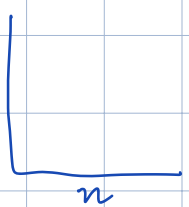
Eg: Tossing a coin 10 times, $n = 10$

$\Downarrow$

$\{H, T\}$

PMF function of Binomial function

Notation: $B(n, p)$



$n \in \{0, 1, 2, \ldots\} \Rightarrow$ no of trials or experiments.

$p \in [0, 1] \Rightarrow$ Success probability for each trial.

$q = 1 - p$

Support :- $k \in \{0, 1, 2, 3, \ldots, n\}$ = Number of success.

$$P(k, n, p) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$$

$${}^{n}C_{k} = \frac{n!}{k!(n-k)!} \Rightarrow \text{Binomial Coefficient}$$

mean = $n \cdot p$

variance = $n \cdot p \cdot q$

$\sigma = \sqrt{npq}$

Eg: Coin Flip

Number of trials (n) = 5)

Probability of success (p) = 0.5

No of success (k) = varies from 0 to 5

1) What is the probability of getting exactly 3 heads in 5 flips?

$n = 5$ $\quad k = 3$

$$Pr(X=3) = {}^5C_3 (0.5)^3 (1-0.5)^{5-3}$$

$$= 0.3125$$

Example: Quality control

Scenario: Inspecting 10 items in a factory where each item has a 10% chance of being defective.

* No of Trials (n) = 10
* Probability of success (p) = 0.1 (defective items)
* No of successes (k) = varies from 0 to 10

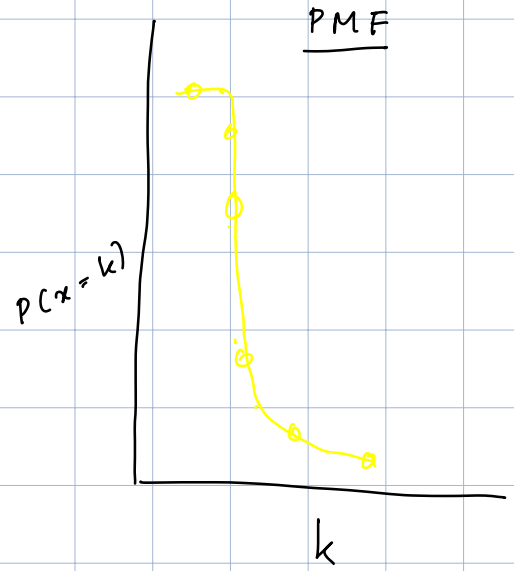
question : What is the probability of finding exactly 2 defective items in a sample of 10?

$$P(X=2) = {}^{10}C_2 (0.1)^2 (0.9)^{10-2}$$

$$= \frac{10!}{8!\,2!} (0.1)^2 (0.9)^8 = 0.1937 //$$

Poisson Distribution

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occuring in a fixed interval of time if these events occur with a known constant mean rate and independently of the time since the last event.

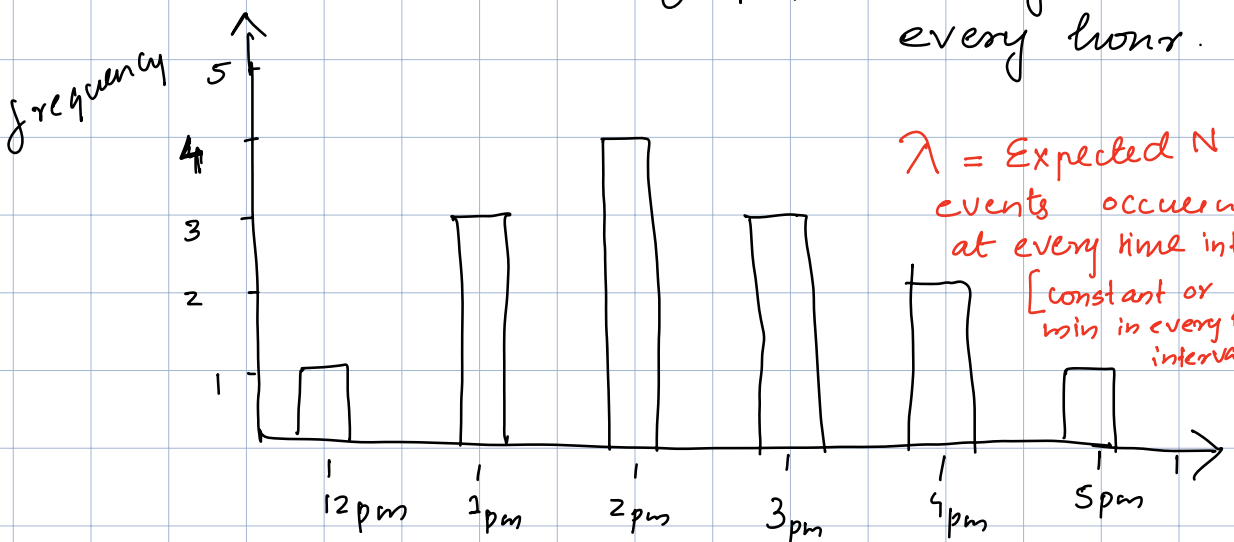


(1) Discrete random Variable (Pmf)

(2) Describe the numbers of events occuring in a fixed time intervals

Eg: No of people visiting hospital every hour

No of people visiting banks every hour.



λ = Expected No of events occuring at every time interval [constant or min in every time interval]

PMF

$$P(X=5) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad \lambda = 3$$

$$= \frac{e^{-3}\,3^{5}}{5!} = 0.101 = 10.1\%$$

Mean of Poisson Distribution

$$\text{Mean} = E(X) = \mu = \lambda * t$$

λ = expected No of events occur in every time interval

t = time interval

Normal / Gaussian Distribution

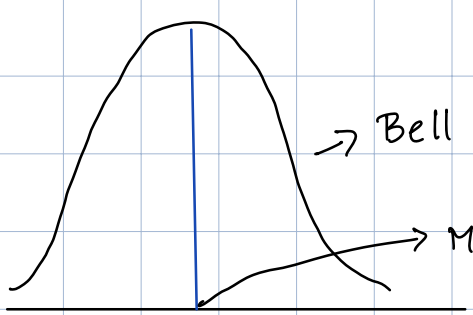
In probability theory and statistics, a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable

⇒ most data follows Normal distribution

⇒ For continuous R.V (Pdf)

$X = \{ \ldots \}$

⇒



→ Bell Curve
Symmetric Distribution.

→ Mean = median = mode

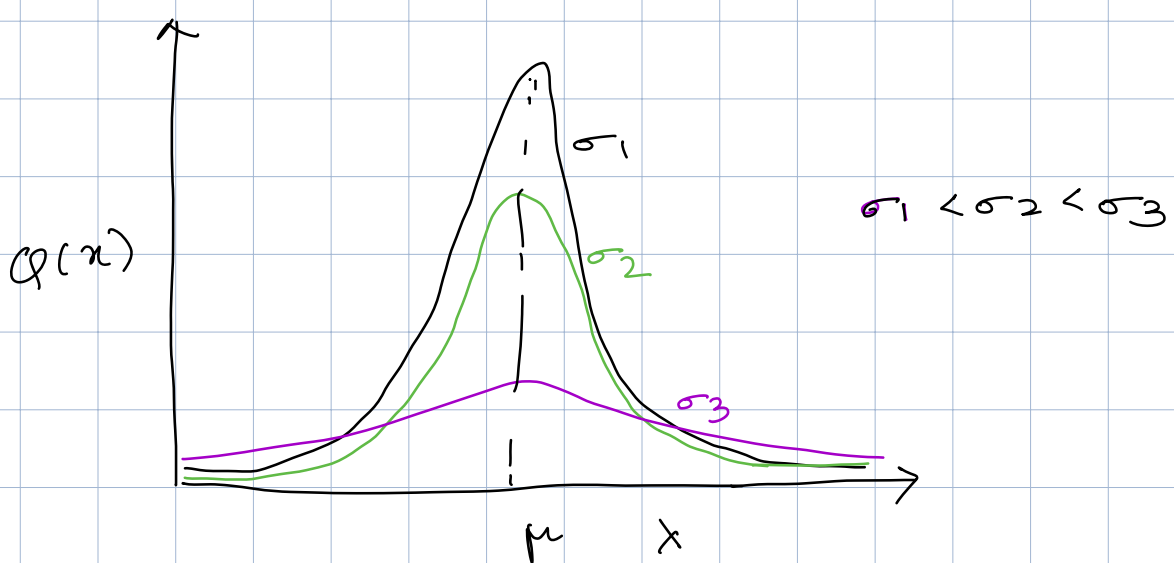
Notation

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

Parameters :- $\mu \in R$ = mean

$\sigma^2 \in R > 0$ = Variance

$x \in R$



Example : IRIS Dataset : Petal length, sepal length, Petal width sepal width

For Gaussian distribution : $PDF = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$

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

Mean

$$\mu = \sum_{i=1}^{n} \frac{x_i}{n}$$

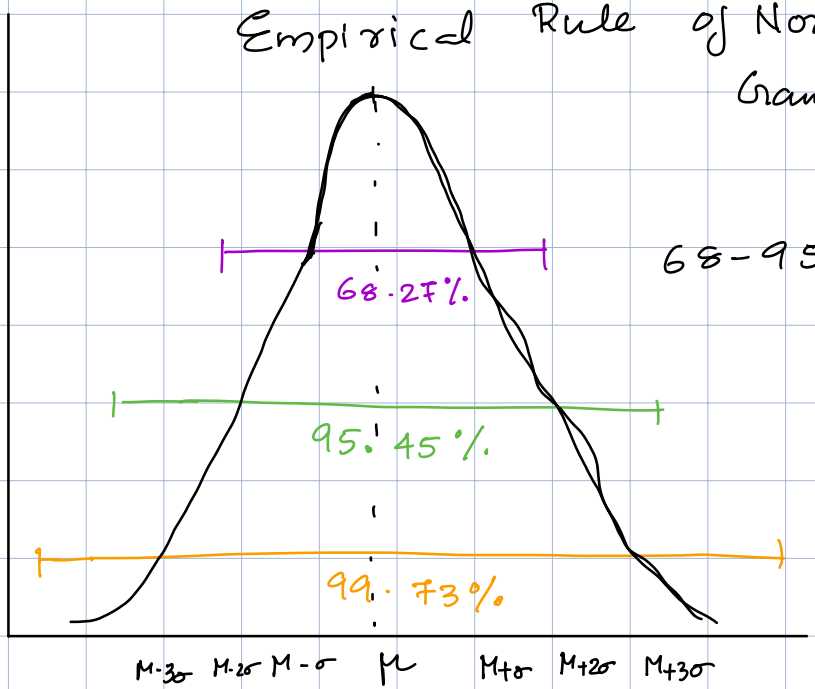
Variance

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

$$\sigma = \sqrt{\text{variance}}$$

Empirical Rule of Normal/ Gaussian Distribution

68-95-99.7 Rule



$X = \{ \ldots\ldots \} \Rightarrow$ Normal / Gaussian Distribution

Probability

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = 68\%$$

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 95.\%$$

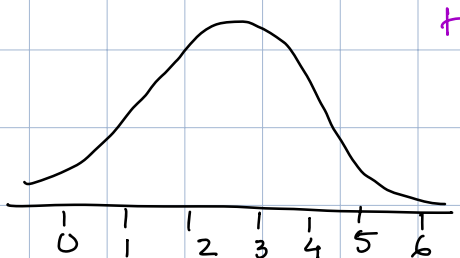
$$P(\mu + 3\sigma \leq x \leq \mu + 3\sigma) = 99.7\%$$

Standard Normal Distribution

$X = \{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma = 1.414 \approx 1$



trying to $\Longrightarrow$ $\mu = 0$, $\sigma = 1$

If $\mu = 0$ and $\sigma = 1$ for Gaussian distribution then we call it as Standard Normal distribution

Every Gaussian distribution can be converted to standard Normal,

⇓

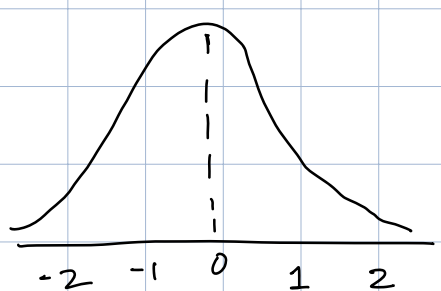
Transformation is done by using Zscore

$$Z\,score = \frac{x_i - \mu}{\sigma}$$

$x = \{1, 2, 3, 4, 5\}$ $\quad Z = \frac{x_i - \mu}{\sigma}$ $\quad \mu = 3$, $\sigma = 1$

$\Downarrow \; x_i - \mu / \sigma$

$y = \{-2, -1, 0, 1, 2\}$



$X \approx SND\ (\mu = 0, \sigma = 1)$

$\mu = 4$

$\sigma = 1$

How many standard deviation 4.25 is away from the mean.

$x_i = 4.25$

$$Z\text{-score} = \frac{4.25 - 4}{1} = 0.25$$

ie 4.25 is 0.25σ way from the mean.

Why Z-score / standard Normal Distribution?

Eg Dataset

Age (years)	Weight (kg)	Height (cms)	salary (INR)
24	70	175	40k
25	60	160	50K
26	55	180	80 k
27	40	130	30 k
30	30	175	20 k
31	35	180	70 k

To bring all data points into same unit to make better utilization by ML algorithms ⇒ Standardization

Standardization ⇒ uses Z score in every feature

$$Z\,score = \frac{x_i - \mu_{age}}{\sigma_{age}}, \quad \frac{x_i - \mu_{weight}}{\sigma_{weight}}$$

⇒ Better model efficiency & performance

Uniform Distribution.

(1) continuous Uniform Distribution (pdf)
(2) Discrete uniform Distribution (pmf)

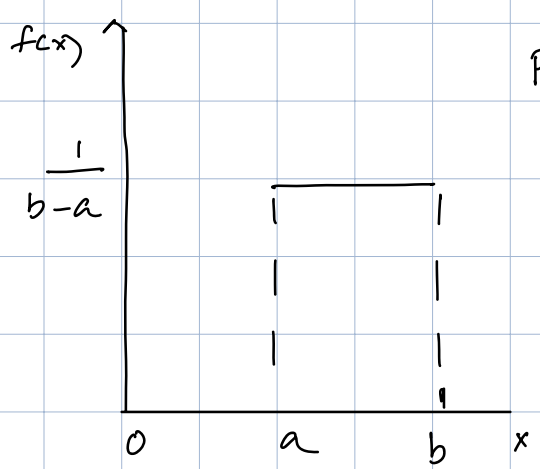
(1) Continuous Uniform Distribution

In probability theory and statistics, the continuous

uniform distribution or rectangular distribution are a family of symmetric probability distribution.

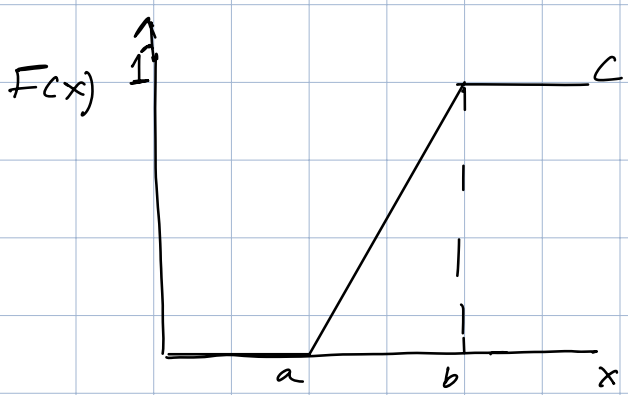
Such a distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds. The bounds are defined by the parameters, a and b. which are the minimum and maximum values

Notation : $U(a, b)$



Parameters: $-\infty < a < b < \infty$

$$Pdf = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$



$$cdf = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$$

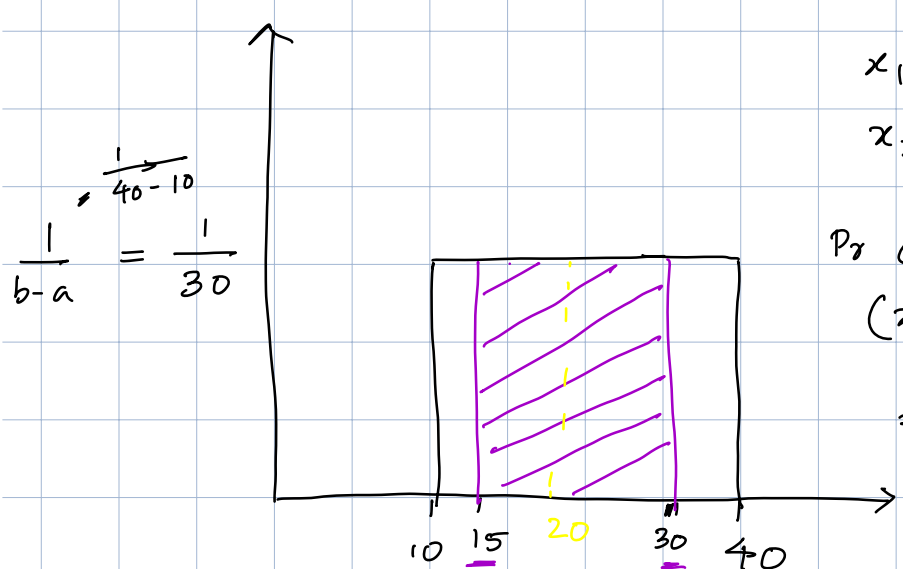
$$\text{Mean} = \frac{1}{2}(a+b)$$

$$\text{Median} = \frac{1}{2}(a+b)$$

$$\text{Variance} = \frac{1}{12}(b-a)^2$$

Eg :- The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 candies and minimum of 10 candies.

(i) Probability of daily sales to fall between 15 and 30?



$x_1 = 15$

$x_2 = 30$

$Pr(15 \le x \le 30) = (x_2 - x_1) * \frac{1}{b-a}$

$= (30-15) \times \frac{1}{30} = 0.5$

$$Pr(x \geq 20) = (40-20) \times \frac{1}{30} = 66\% = 0.66$$

Discrete Uniform Distribution

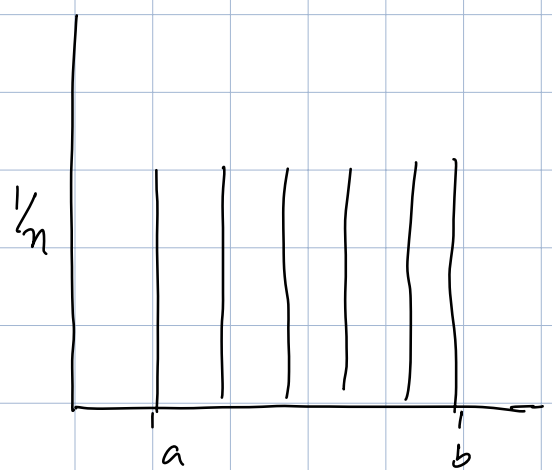
* Discrete Random Variable

* Pmf

$Pr(1) = 1/6$
$Pr(2) = 1/6$
$Pr(3) = 1/6$
⋮

Eg: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

(a) (b)



$1/n \Rightarrow n = b - a + 1$
$= 6 - 1 + 1 = 6$

Notation $U(a, b)$

Parameters a, b where $b > a$

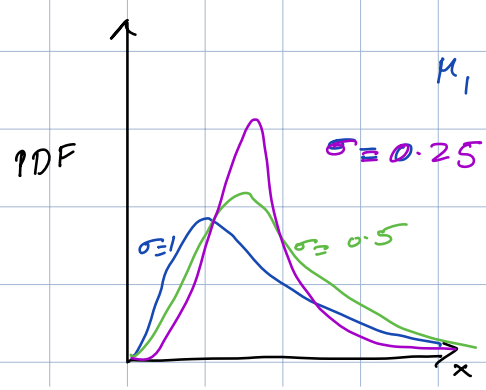
PMF $= 1/n$

In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution wherein a finite number of values

are equally likely to be observed; every one of n values has equal probability $1/n$. Another way of saying "discrete uniform distribution" would be "a known, finite number of outcomes equally likely to happen.

Log Normal Distribution

In probability theory, a log-normal (or lognormal) distribution is a continuous probabity distribution of R.V whose logarithm is normally distributed. Thus, if the random variable X is a log-normally distributed, then $Y = \ln(X)$ has a normal distribution. Equivalently, if Y has a normal distribution, then the exponential function of Y, $X = \exp(Y)$, has a log-normal distribution.



$\mu_1 = \mu_2 = \mu_3 = 0$

$\sigma = 0.25$

$\sigma = 1$

$\sigma = 0.5$

X = log normal Distribution (μ, σ)

Right-skewed

$Y \approx \ln(X)$ = Normal Distribution

↓

natural log

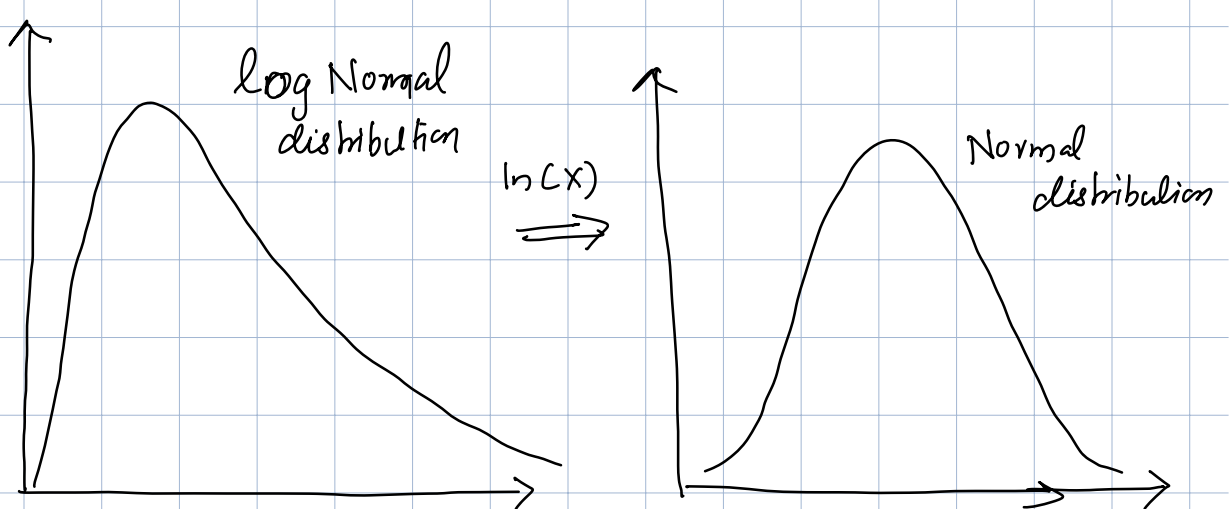
⇓

$\log_e$

If $Y \approx \ln(X)$

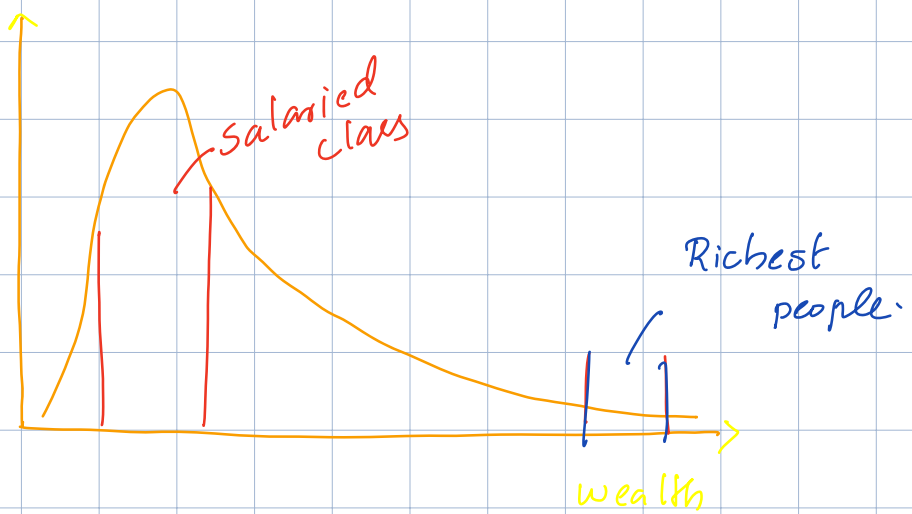
⇓

$X \approx \exp(Y)$ ⇒ log Normally Distributed



Q-Q plot used to check log normal distribution

Eg: ① Wealth distribution of the world



② Discussion Forum → length of the comments

long comments — few

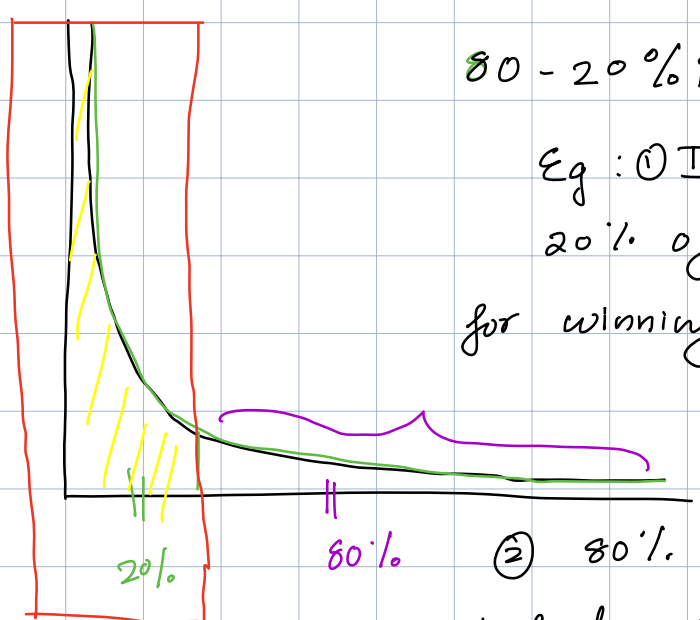
Medium sized comment — more people

small — large number

③ Time spend in Websites / Dwell time.

Power law Distribution

In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another



80-20% Rule

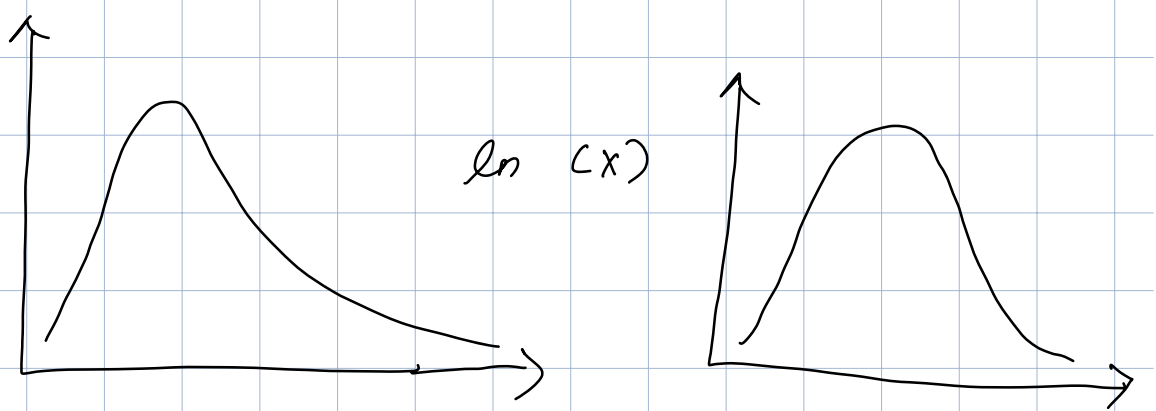
Eg: ① IPh
20% of team is responsible for winning 80% of match.

② 80% of wealth are distributed with 20% of the total population

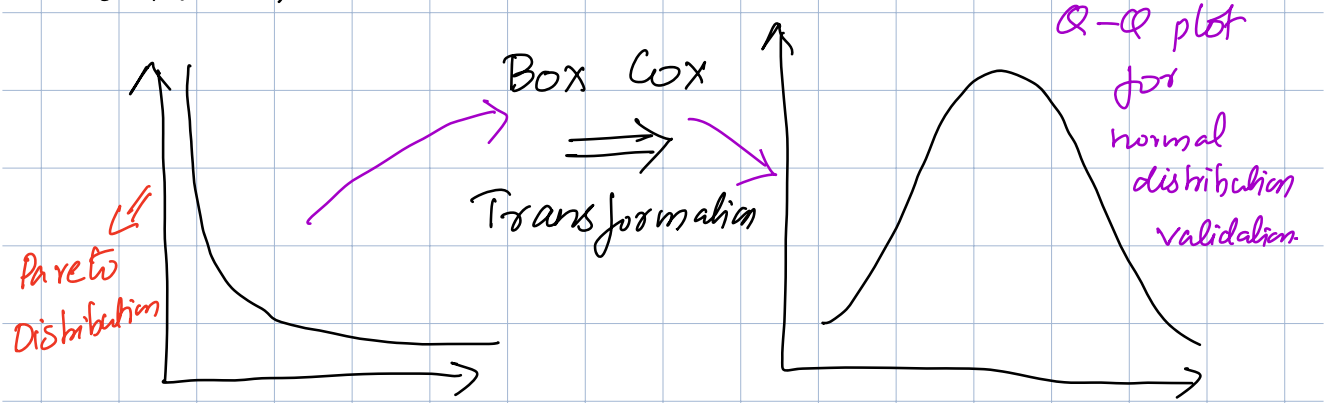
③ 80% of the total oil is with 20% of the nation.

④ Frequency of words in most languages

⑤ 20% of major defects fixes the 80%. of upcoming defects in software product



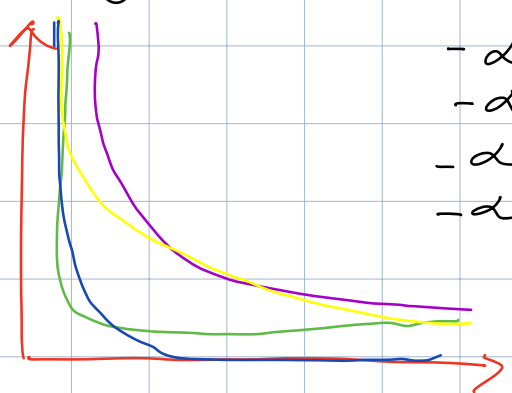
Normal distribution because many ML algorithms performs well will normal distribution.



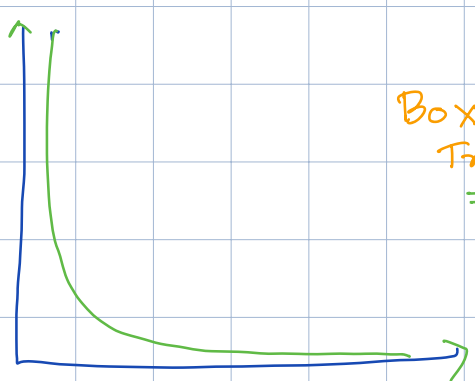
Pareto Distribution

Any distribution which follow power law (80 - 20 rule) is called Pareto Distribution.

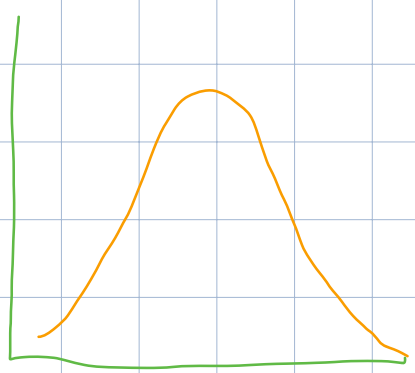
⇒ Non gaussian distribution



- $\alpha = \alpha$ ⇒ higher height
- $\alpha = 3$
- $\alpha = 2$
- $\alpha = 1$ ⇒ touch x axis



Box-Cox Transformation ⇒



Eg :- IT Industry ⇒ 80% of entire projects done by 20% of team.

② 80% of defects can be solved if we solve 20% of defects

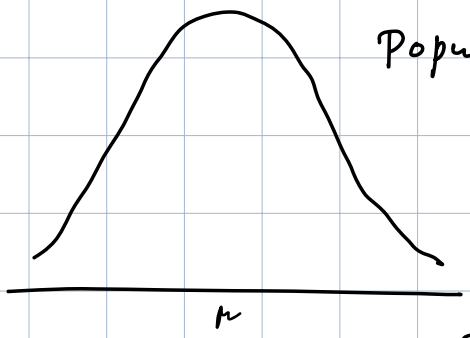
Central limit theorem

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial or any other distribution, the sampling distribution of the mean will be normal.

① $\bar{X} \approx N(\mu, \sigma)$

n = sample size = any value



Population DATA

$S_1 = \{x_1, x_2 \cdots x_n\} = \bar{x}_1$

$S_2 = \{x_2, x_3 \cdots x_n\} = \bar{x}_2$

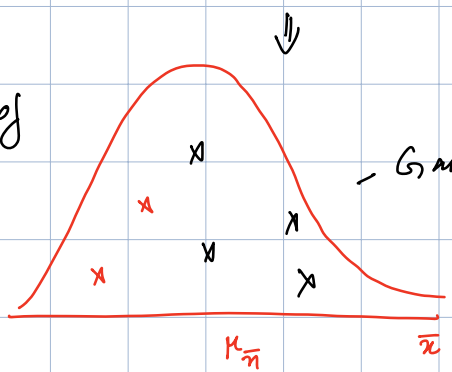
$S_3 =$

$\vdots$

$S_m =$ $\bar{x}_m$

x_i

Sampling Distribution of mean



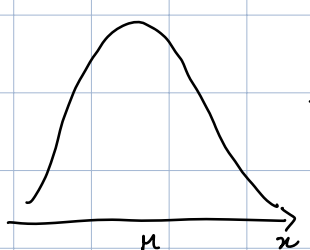
→ Gaussian Distribution

if R.V is not following a gaussian distribution then n-value should ≥ 30 ($n \geq 30$) sample size.

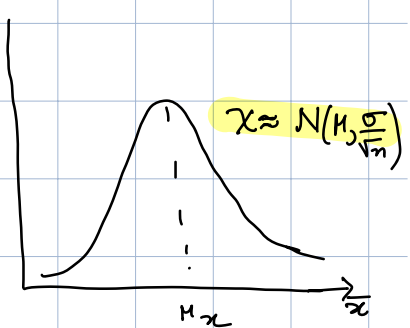
For Normal distribution, n can be any value.

① Normal

$X \approx N(\mu, \sigma)$



$\Rightarrow$ sampling distribution of mean



After sampling new mean = population mean

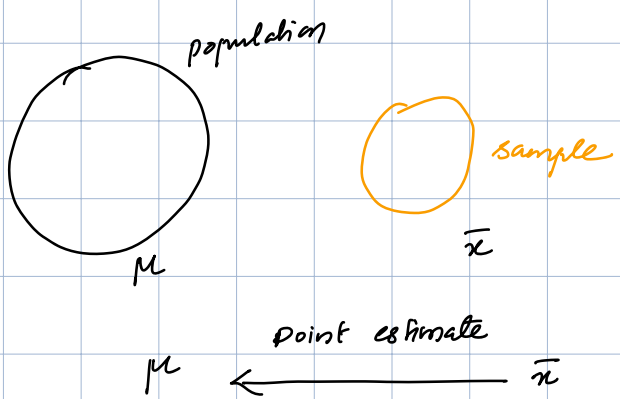
$$\text{new std} = \frac{\sigma}{\sqrt{n}}$$

Estimates

Estimate :- Is specified observed numerical value used to estimate an unknown population parameter.

(1) Point estimate : single numerical value used to estimate an unknown population parameter

Eg: sample mean $\bar{x}$ is a point estimate of population mean

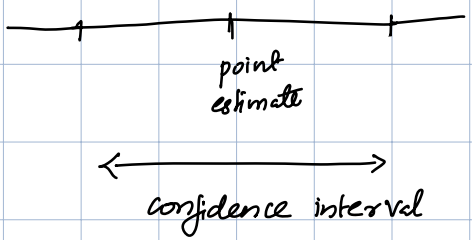


(2) Interval estimate

A range of values is used to estimate unknown

population parameter.

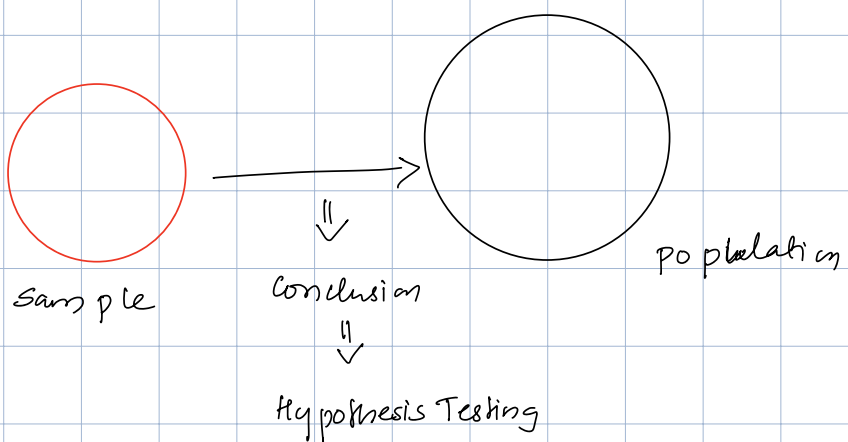
55 - 65



Inferential Statistics

↳ Conclusion or Inferences from sample data about population parameter.

A hypothesis and Hypothesis Testing Mechanism



Hypothesis Testing Mechanism

① Null hypothesis (H_0) — e.g. Person is not guilty

- The assumption you are beginning with.

② Alternate hypothesis (H_1) - The person is guilty

Opposite of null hypothesis

③ Experiments → Statistical Analysis
→ Collect proofs (DNA, finger Print)

④ Accept or reject the null hypothesis

Eg: Colleges at District A states its average passed percentage of students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass percentage of 90%. With a standard deviation of 4%

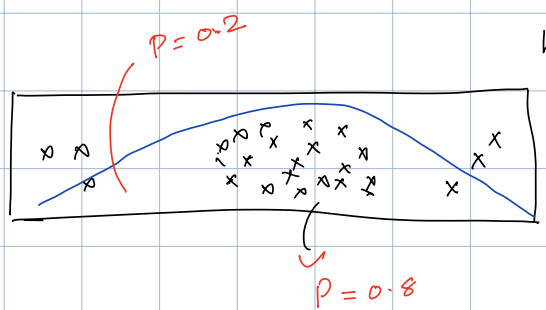
Does this college have a different passed percentage.

Ans). Null hypothesis (H_0) = $\mu = 85\%$

Alternate hypothesis (H_1) = $\mu \neq 85\%$

P value

The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.



keyboard space bar

maximum contact in the center

Hypothesis Testing

Eg: Coin is Fair or Not {100 Tosses} {100 times}

P(H) = 0.5 P(T) = 0.5

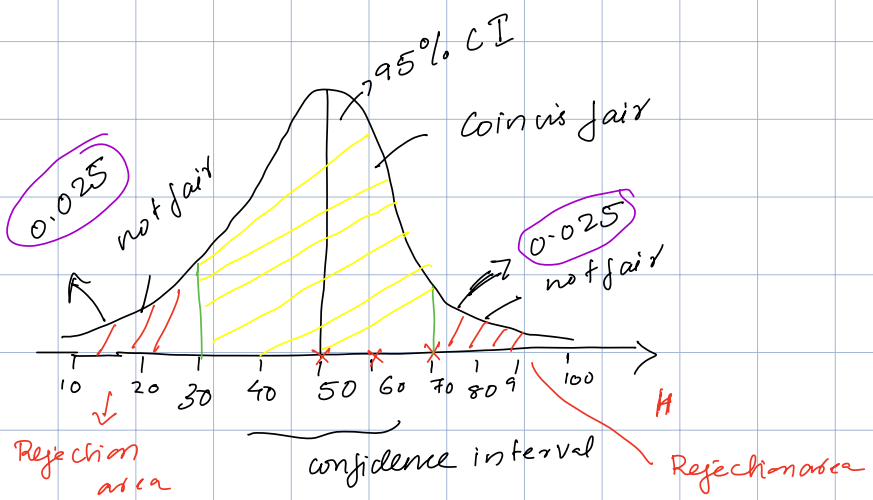
P(H) = 0.6 P(T) = 0.4

P(H) = 0.7 P(T) = 0.3

① Null hypothesis (H_0) = Coin is fair

② Alternate Hypothesis (H_1) = Coin is not fair

③ Experiment ⇒ toss for 100 times



④ Significance value : $\alpha = 0.05$

Confidence interval = $1 - 0.05 = 0.95$

⑤ Conclusion

If P < significance value
⇓
Reject the null Hypothesis > Rejection area

Else ⇒ Fail to reject the null hypothesis

Hypothesis Testing and Statistical Analysis

Z Test, t test } Data dealing with average data — Z table (Zscore, Pval), t table (tscore, Pval)

Chi square ⇒ Categorical data

Annova ⇒ For variance of data

Z test

Conditions to use z-test
= (i) population std known (ii) $n \geq 30$

The average height of all residents in a city is 168cm with $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5cm.

(a) State null and Alternate Hypothesis

(b) At a 95% confidence level, is there enough evidence to reject the null hypothesis.

Ans) $\mu = 168cm$ $\sigma = 3.9$

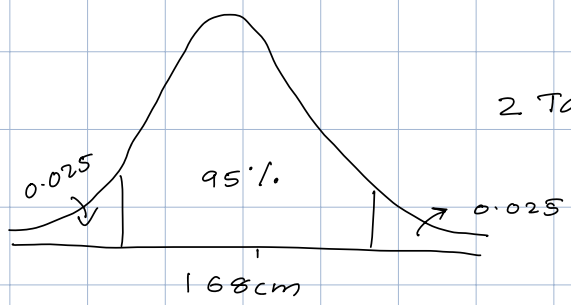
$n = 36$ $\bar{x} = 169.5cm$ C.I = 0.95

$\alpha = 1 - C.I = 0.05$

a) ① Null hypothesis $(H_0) \Rightarrow \mu = 168cm$

② Alternate hypothesis $(H_1) \Rightarrow \mu \neq 168cm$

③ Based on C.I we will draw decision boundary



$1 - 0.025 = 0.9750 \Rightarrow$ Z-Score

⇓

Z-table pdf

↳ check the value 0.9750

⇓

Area $\Rightarrow +1.96$

similarly in left $= -1.96$

if Z is less than -1.96 or greater than $+1.96$, Reject the Null hypothesis

Z will be calculated using statistical analysis

(*) Z - test

$$Z_d = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$Z\text{-score} = \frac{\bar{X} - \mu}{\sigma}$$

$\Downarrow$

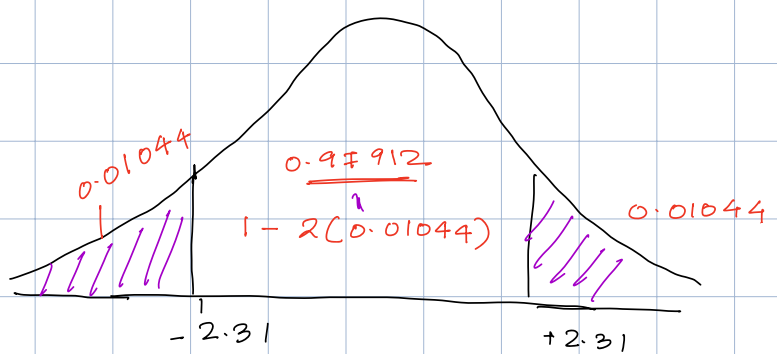
$\Downarrow$ sample s.td according to CLT

$$= \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.31$$

2.31 is greater than 1.96

$2.31 > 1.96 \Rightarrow$ Reject the null hypothesis

$P < 0.05$



again go to Z - table

2.31 correspond to area 0.98956

so shaded area is $1 - 0.98956 = 0.01044$

P value = 0.01044 + 0.01044

= 0.02088 $P < 0.05$

= 0.02088 < 0.05 = reject null hypothesis

⊛ Final conclusion the average height ≠ 168cm

The average height seems to be increasing based on sample data.

Question

A factory manufactures bulbs with a average warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will malfunction in less than 5 years. He test a sample of 40 bulbs and find the average time to be 4.8 years.

(a) State null and alternate hypothesis

(b) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

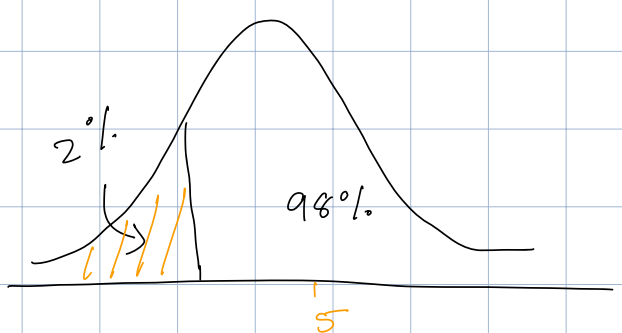
a) ① Null hypothesis $\mu = 5$

given $\sigma = 0.50$

② Alternate hypothesis is $\mu < 5$ { 1 Tail test }

given $n = 40$

b) decision boundary = one tail test



c) Z-test

$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.5/\sqrt{40}}$$

$$= -2.53$$

Area under the curve with Z score

$-2.53 = 0.0570$

P-value $= 0.0570$

Compare p-value with α

$0.0570 < 0.02 \Rightarrow$ False

P value is not falling in the rejection area

We accept the null hypothesis or failed to reject null hypothesis

Student test

In z stats when we perform any analysis using z-score. we require σ (population standard deviation) → is already known

How do we perform any analysis when we do not know the population standard deviation?

Here we use students t distribution.

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

population std

⇓

Z table used here

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

sample std

⇓

t table

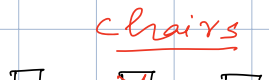
⇓

also need degree of freedom

Degree of freedom

$dof = n - 1$

let's say 3 people



1) 1st selection from 3 option

[ie when 3 people were there only two people were having options.]

2nd person selects from 2 option

⇓

So $dof = 3 - 1 = 2$

3rd person has no option

T table has

cum. prob	$t_{.50}$ ..... $t_{.9995}$
one tail	0.50 ..... 0.0005
two tail	1.00 ..... 0.001
degree of freedom	
1	
2	
.	
.	
.	

T-stats and T test → One sample t-test

① In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a position or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence.

⇒ either positive or negative effect on intelligence ⇒ Two tail test

Ans) $\mu = 100$

$n = 30$

$\bar{x} = 140$

$s_d = 20$

Initialize C.I = 95% $\alpha = 0.05$

① Null hypothesis H_0 : $\mu = 100$

Alternate hypothesis H_1 : $\mu \neq 100$

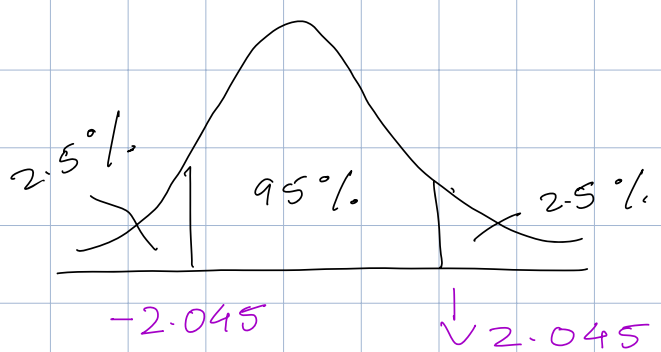
(two tailed test)

② $\alpha = 0.05$

③ Degree of freedom =

$n - 1 = 30 - 1 = 29$

④ Decision rule



Go to t - table

1) 2 tail with $df = 29$

$\alpha = 0.05$ with $d = 29$

$\Downarrow$

2.045

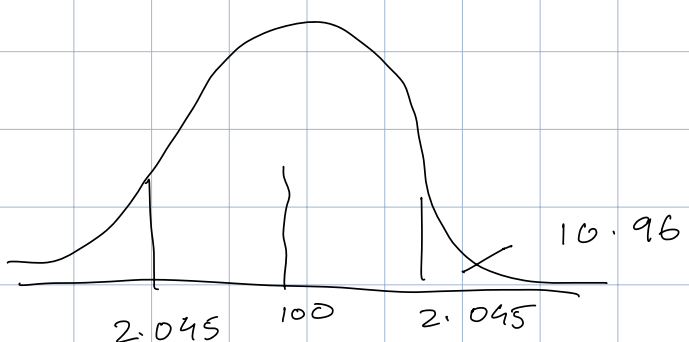
The assumption here is met- if my t-test is less than -2.045 and greater than 2.045, reject the null hypothesis

⑤ Calculating Test statistics

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = \frac{40}{3.65} = 10.96$$

$10.96 > 2.045$

so



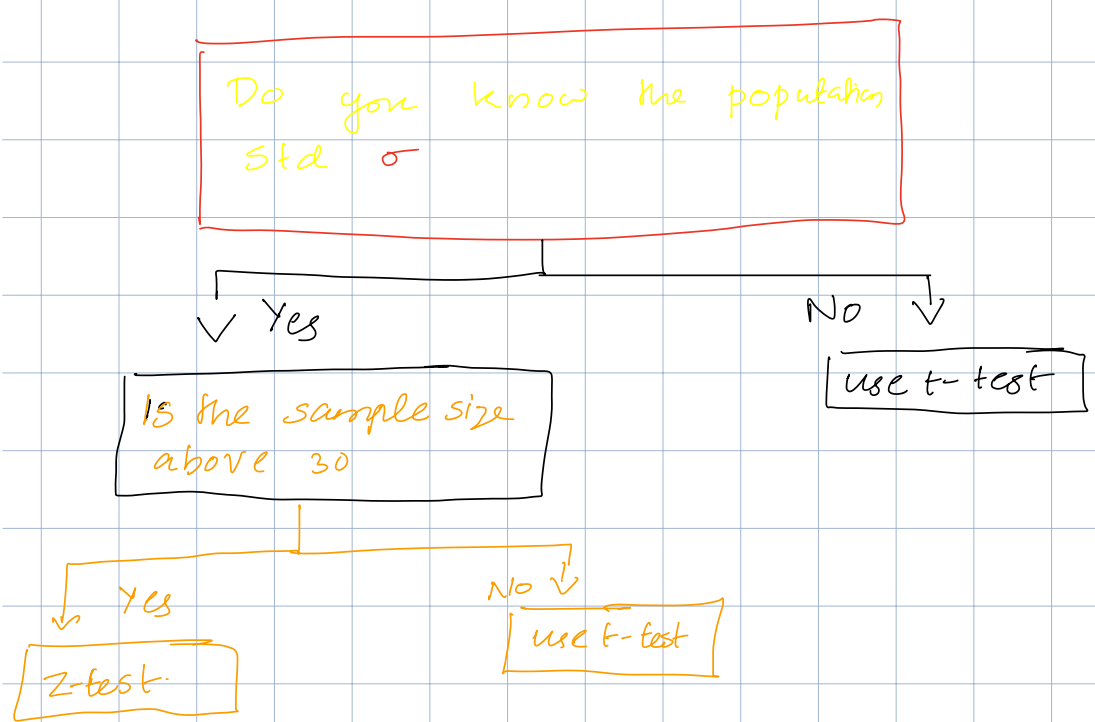
10.96 falls in the rejection area

Conclusion

Medication used has affected the intelligence

Specifically medication has increased the IQ.

When to use T-test Vs Z-test



Type I and Type 2 Errors

Reality : Null hypothesis is True or Null hypothesis is False

Decision : Null hypothesis is True or Null hypothesis is False

Outcome 1 :- We reject the null hypothesis when in reality it is false → Good

Outcome 2 : We reject the null hypothesis when in reality it is true ⇒ Error

⇓

Type I Error

Outcome 3 : We retain the Null hypothesis when in reality it is False

⇓

Type to Error

Outcome 4 :- We retain the Null hypothesis when in reality it is True ⇒ Good / correct scenario

Bayes theorem (Bayes Statistics)

Bayesian statistics is an approach to data analysis and parameter estimation based on Baye's theorem.

Baye's Theorem

Probability
- → Independent events
- → Dependent events

(1) Independent event

eg : Rolling a dice
{1, 2, 3, 4, 5, 6}

(2) Dependent event

Eg Tossing a coin

{ H, T }

P(red) = 2/5 = 1st event

P(yellow) = 3/4 = 2nd event

one after the other and there is dependency

So P(Red & Yellow) =

$P(R \& Y) = P(R) * P(Y/R)$

conditional probability

$= 2/5 \times 3/4 = 6/20$

In general:

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

$P(A) * P(B/A) = P(B) * P(A/B)$

$$P(B/A) = \frac{P(B) * P(A/B)}{P(A)}$$

Baye's theorem

$$P(A/B) = \frac{P(A) * P(B/A)}{P(B)}$$

A, B = events

$P(A/B)$ = Probability of A given B is true

$P(A), P(B)$ = Independent probabilities of A and B

DATASET

I/P ⇒ Independent feature

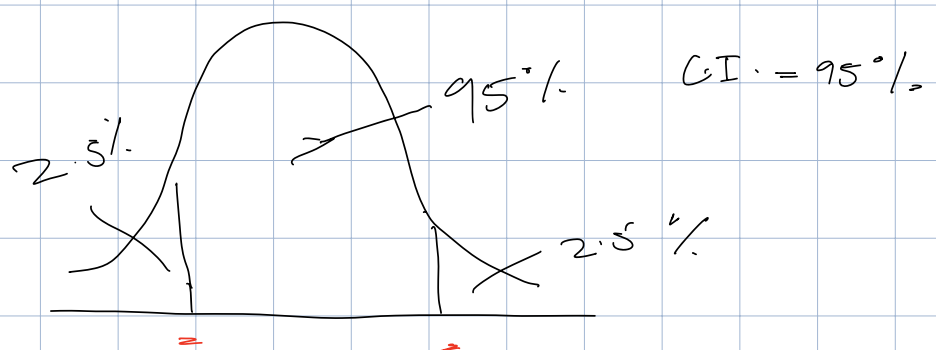
O/P dependent = feature

Size of house	Number of Rooms	Location	Price
x_1	x_2	x_3	y

Bayes theorem used here

$$P(Y/x_1, x_2, x_3) = \frac{P(Y) * P(x_1, x_2, x_3/Y)}{P(x_1, x_2, x_3)}$$

Confidence Intervals and Margin of Error



Point estimate

$\bar{x}$ → μ

$\bar{x} = 2.5$

$M = 3$

not always same

Confidence interval

Point estimate ± margin of Error

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

↑ for Z_{test}

$$\bar{x} \pm t_{\alpha/2} \sigma/\sqrt{n}$$

for t_{test}

Eg: On the verbal section of CAT exam, the standard deviation is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% CI about the mean.

$$\bar{x} \pm Z_{\alpha/2} \sigma/\sqrt{n}$$

$\alpha = 0.05$

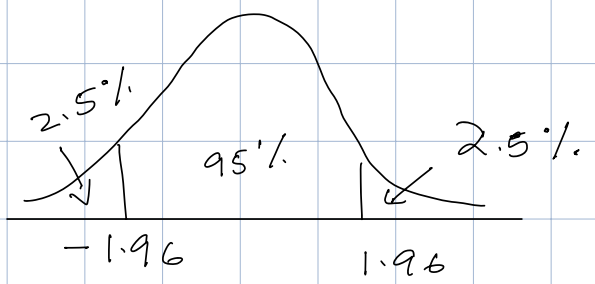
$1 - 0.025 = 0.9750$

$\downarrow\downarrow$

Z table

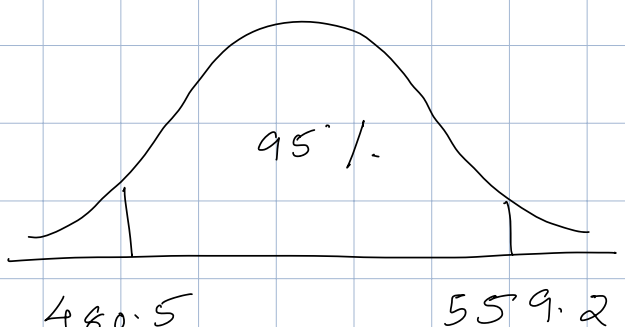
$\parallel$

± 1.96



$$\text{lower C.I} = 520 - (1.96)\frac{(100)}{\sqrt{25}} = 480.5$$

$$\text{Higher C.I} = 520 + (1.96)\frac{100}{\sqrt{25}} = 559.2$$



So CAT score will be 95% confidently between 480.5 and 559.2

CHI SQUARE TEST

⇒ for categorical variables

⇒ Chi-square test for Goodness of fit test

Claims about population proportion

It is a non parametric test that is performed on the categorical data. [Ordinal and nominal]

There is a population of male who likes different colour bikes.

		(Prob) known theory about population	Sample
Yellow	Bike	1/3	22
Red	Bike	1/3	17
Orange	Bike	1/3	59

We have to see if the sample information is a good fit for population

* Goodness of fit test

Goodness of fit test

In a science class of 75 students, 11 are left handed. Does this class fit the theory that 12% of people are left handed.

		Observed	Expected
Ans)	Left handed	11	$\frac{12}{100} \times 75 = 9$
	Right handed	$\frac{64}{75}$	$\frac{66}{75}$

Chi - square test we are comparing 2 categorical variables

CHI SQUARE FOR GOODNESS OF FIT

In 2010 census of the city, the weight of the individuals in a small city were found to be the following

≤50kg	50 - 75	> 75
20%	30%	50%

In 2020, weight of $n = 500$ individuals were sampled. Below are the results

≤50	50 - 75	>75
140	160	200

Using $\alpha = 0.05$, would you conclude the population difference of weights has changed in the last 10 years?

Ans) We use Chi-Square table here

1st take 2010 data

<50kg	50-75	>75
20%	30%	50%

Expected

take 2020 observed data $n = 500$

<50	50-75	>75
140	160	200

<50	50-75	>75
0.2×500 = 100	0.3×500 = 150	0.5×500 = 250

Expected info based 2010 census.

Apply Chi - Square

① Null hypothesis: H_0: The data meets the expectation

Alternate hypothesis H_1: The data does not meet the expectation

② $\alpha = 0.05$ C.I = 95%.

③ Degree of freedom

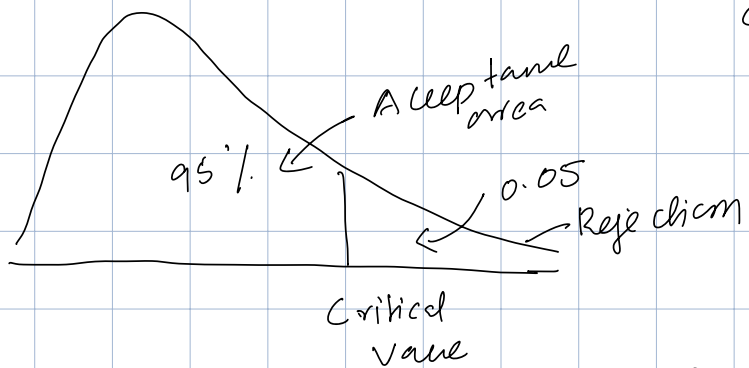
$df = k - 1 = 3 - 1 = 2$

k = number of categories

④ Decision boundary [In Z, t we used symmetrical distribution]

But we use a one tailed statistic

Chi square Probability



5.991 (by checking Chi-square table)

If χ^2 is greater than 5.99, Reject H_0

else

we fail to reject H_0

(5) Calculate Chi - Square Test Statistics

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$= 16 + 0.66 + 10 = 26.66$$

Decision rule was if χ^2 is greater than 5.99 Reject H_0

Here $26.66 > 5.99$ so Reject H_0

Answer

The weights of 2020 population are different than those expected in 2010 population.

What is ANOVA?

Analysis of Variance

Definition: ANOVA is a statistical method used to compare the means of 2 or more groups.

ANOVA

(1) Factors (variable)

(2) Levels

Eg: Medicine (Factor)

[Dosage] 5 mg 10 mg 15 mg [level]

Eg) Mode of Payment- [Factor]

GPAY PHONEPE PAYPAL [Levels]

Assumptions in ANOVA

(1) Normality of Sampling Distribution of Mean

The distribution of sample mean is normally

distributed

② Absence of Outliers

Outlier score need to be removed from dataset

③ Homogeneity of Variance

Population variance in different levels of each independent variable are equal

$$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$$

(4) Samples are independent and randomly selected

Types of ANOVA

3 types

1) One way ANOVA : One factor with atleast 2 levels and these levels are independent

Eg :- Doctor wants to test a new medication to decrease headache.
They split the participants in 3 conditions [10mg, 20 mg, 30mg]

Doctor ask the participants to rate the headache between [1 -10]

Medication → Factor

10mg	20mg	30mg
5	7	2
3	4	6
–	–	–
–	–	–

Type II) Repeated measures ANOVA :-

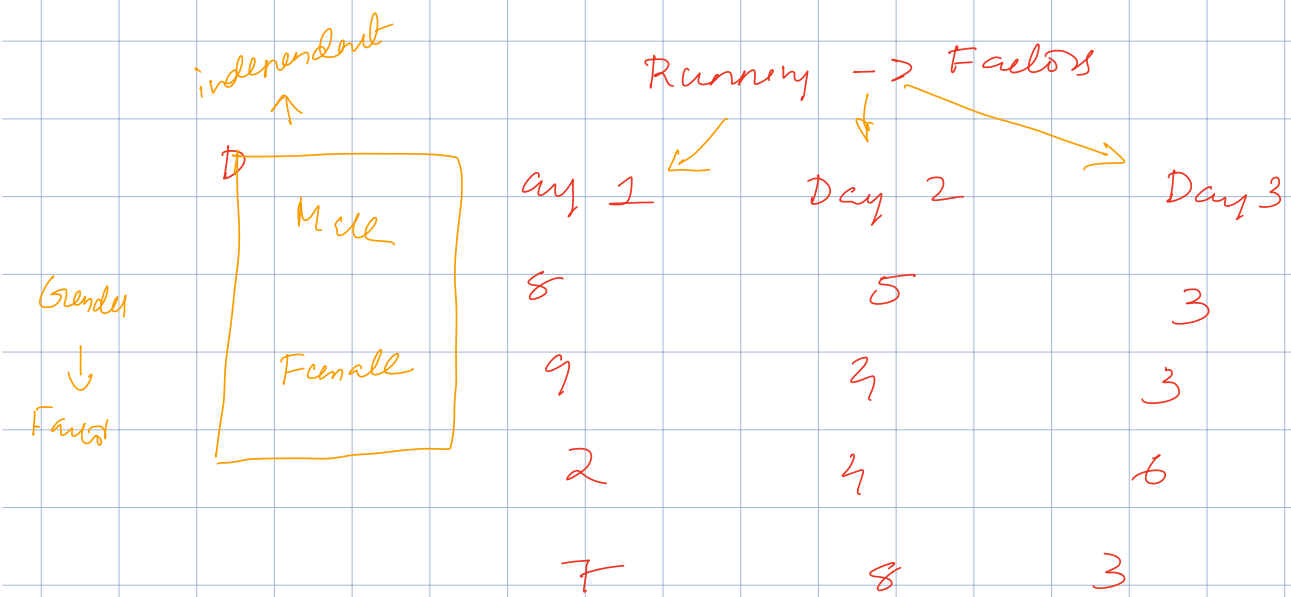
one factor with atleast 2 levels, levels are dependent.

Running → Factor

Day 1	Day 2	Day 3 ← levels
8	5	4
7	4	9
–	–	–

③ Factorial ANOVA:

Two or more factors each of which with atleast 2 levels, levels can be independent and dependent.



Hypothesis Testing In ANOVA

(Partitioning of Variance in the ANOVA)

1) Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 \cdots \mu_k$

Alternate hypothesis H_1: atleast one of sample mean is not equal

Test Statistics

$$F = \frac{\text{Variance between samples}}{\text{Variance within sample}}$$

Variance between samples

Samples = x_1	x_2	x_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	1

Variance within samples

$\bar{x}_1 = 3$ $\quad \bar{x}_2 = 19/5$ $\quad \bar{x}_3 = \frac{16}{4} = 4$

$H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3$

H_1 at least one sample mean is not equal

One Way ANOVA

One factor with atleast 2 levels, levels are independent

① Doctors want to test a new medication which reduces headache. They split the participants into 3 condition [15 mg, 30mg, 45mg). Later on the doctor ask the patient to rate the headache between [1–10]. Are there any differences between 3 conditions

alpha = 0.005?

	15mg	30mg	45mg
Any	9	7	4
	8	6	3
	7	6	2
	8	7	3
	8	8	4
	9	7	3
	8	6	2

① Define null and alternate hypothesis

H_0 $\quad \mu_{15} = \mu_{30} = \mu_{45}$

H_1 $\quad$ atleast one mean is not equal

② Significant value

$\alpha = 0.05$

C.I = 95%

③ Degree of freedom

$N = 21$ $\quad [7 \times 3]$

↓ rows ↓ columns

$a = 3$ ⇒ categories (levels)

$n = 7$ samples in each category

$df_{between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - a = 21 - 3 = 18$

$df_{total} = N - 1 = 20$

$df_{btwn} + df_{within} = df_{total}$

In F table df_1 df_2

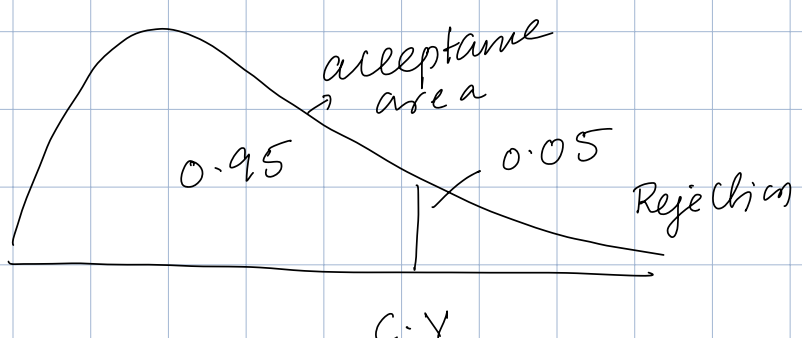
(21, 18)

$\alpha = 0.05$

⇓

Critical value

(4) Decision Boundary



(5) F table

check df_1 (between), df_2 (within)

$\alpha = 0.05$ (2, 18)

$\alpha = 0.05$

Decision Rule

If F is greater than 3.5546, reject the Null Hypothesis

(5) Calculate the F test statistics

$$F = \frac{\text{Variance between sample}}{\text{Variance within sample}}$$

	Sum of squares	df	M S	F
Between				
Within				
Total				

15mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

$$\text{S.S between} = \sum \frac{(\sum a_i)^2}{n} - T^2/N$$

$\sum a_i$

15mg $\Rightarrow 9+8+7+8+8+9+8 = 57$

30mg $\Rightarrow$ $7+6+6+7+8+7+6 = 47$

45mg $\Rightarrow$ $4+3+2+3+4+3+2 = 21$

$$S.S_{btwn} = \sum \frac{(\sum a_i)^2}{n} - \frac{T^2}{N}$$

$$= \frac{57^2+47^2+21^2}{7} - \frac{[57^2+47^2+21^2]}{21}$$

$$= 98.67$$

(2) $$S.S_{within} = \sum y^2 - \sum \frac{(\sum a_i)^2}{n}$$

$$= \sum y^2 = 9^2+8^2+7^2+8^2+8^2+\ldots$$

$$= 853 - \frac{57^2+47^2+21^2}{7}$$

$$= 10.29$$

Variance between / variance within → F

	Sum of squares	df	mean squared M.S	F
Between	98.67	2	49.34	
Within	10.29	18	0.54	
Total	108.96	20		

$$F \text{ test} = \frac{49.34}{0.54} = 86.56$$

If F is greater than 3.5546, we Reject H_0

Here $86.56 > 3.5546$ so reject H_0