Date / /
Page L&W

* Data Engineering

- Creating and maintaining infrastructure for Data Analysts and Data Scientists, Big Data Analytics, Data visualization

* Data Analytics

- extracts insights from data and presents findings.

4 type of Data Analytics

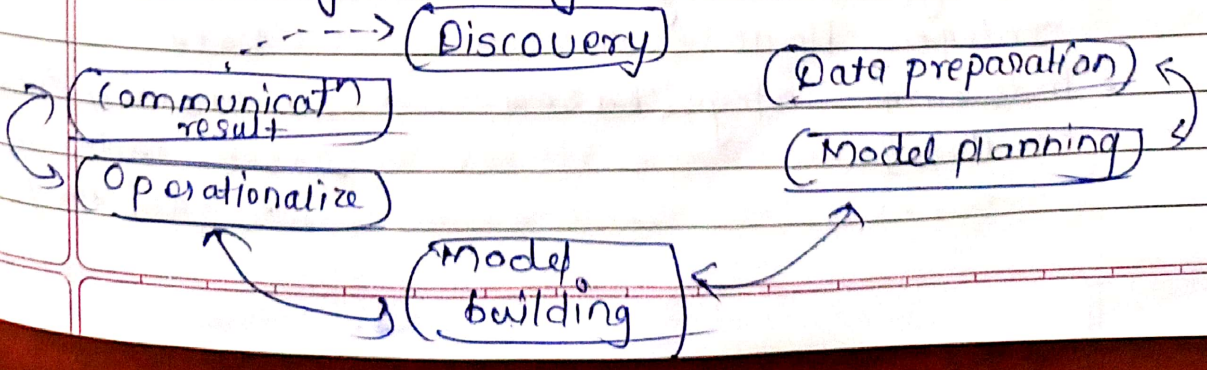
1] Descriptive Analytics :- Summarize past data to understand changes and patterns.

2] Diagnostic Analytics :- ex. Analyzing a drop in sales. def :- Identify reasons or causes behind events.

3] Predictive Analytics :- Forecast future trends using historical data.

4] Prescriptive Analytics :- Recommend actions based on predictions.

* Data analytics life cycle



* Numpy
- More efficient than lists: Homogeneous unlike lists (All elements are of same type) and are stored in contiguous memory location
- More efficient Vectorization: Apply a function on entire array without needing a for loop
- Broadcasting: Align arrays of different shapes, e.g result = arr_2d + arr_1d
- can handle missing value

* Pandas
- Pandas is a powerful python library used for data manipulation, analysis, and checking mising values, data cleaning. It provides easy-to-use data structures and functions to work with structure data, espedially tabular data

* Diff betn Numpy and Pandas

Numpy	Pandas
- Numerical computation using n-dimensional arrays	• Data processing using series and dataFrames
- Data type:- Mainly Integer, float	• Numeric, Text, Date
- Indexing: Integer based [0,1]	• Additionaly support label-indexing eg. df['age']

- BuiltinOperation Numerical and linear-algebra related
- Data analysis tools such as merging, sorting, joining, handling missing data etc.

* Pandas series : A data structure that holds an array of information along with a named index
- Dataframe : Table of columns and rows that can be easily restructured / filtered.
- crosstab() : function in pandas is used to compute a cross-tabulation (contingency table) of two or more factors (categories). It's helpful for summarizing the ~~real~~ relationship between categorical variables.

* Matplotlib

- One of the most popular plotting libraries in python.
- Seaborn / pandas built in visualization are built on top of matplotlib.
- Main types of plots
  - Line plot : Great for showing functional relationships and continuous data
  - Scatter plot : Useful for plotting raw data points and understanding the correlation bet^n two variables.
  - Bar plot : Useful for categorical data to show comparisons bet^n diff groups.

- Histogram: Good for showing the distribution of a single variable
- Pie charts: Used for showing proportion or percentage of categories

* Seaborn

- statistical plotting library
- Built on top of matplotlib, but uses a simpler one-line syntax

Type of plot

- scatter plot: - Relationship between two continuous variables (Trends, correlation)
- Distribution plots: How a single variable is distributed (patterns, skew, outliers) (Histogram, KDE plot)
- categorical plot: Categorical variables and their relationships with continuous data (Box plot, bar plot, count plot)
- Comparison plots: Compare two or more variables (pair plot)
- Matrix plots: Complex relationship in a matrix form (heatmap)

Interview

Difference between matplotlib and Seaborn

- Matplotlib: is a basic, low-level plotting library in python that provides full control over plot elements like axes,

labels, and styles. It's highly customizable and widely used for creating static, interactive, and animated visualizations.

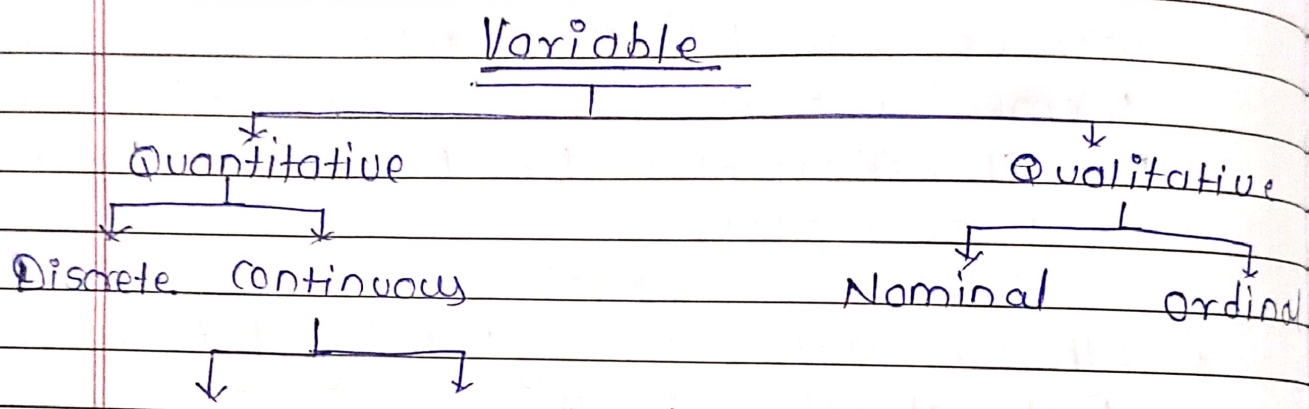
- Seaborn: is a high-level visualization library built on top of Matplotlib. It simplifies complex visualizations and comes with beautiful default styles and themes. Seaborn is particularly powerful for statistical plots like heatmaps, boxplots, and pair plots

* What is data cleaning? Why it is important
- Data cleaning is the process of identifying and correcting errors or inconsistencies in a dataset to improve its accuracy, completeness, and reliability before analysis or modeling.

- Why important?
→ • Ensures data quality
• Removes inaccuracies, inconsistencies, and redundancies
• Make analysis more efficient and meaningful
• Increases trust in data-driven decisions.

* Basic of Statistics

- Qualitative variables: Categorical, Non-numeric.
- Quantitative variables: Numeric. Can be measured



- Discrete :- Countable and take specific values
  Rank : 1, 2, 3 but not 1.8
- Continuous :- Measurable, but not a specific
  Height : 5 feet 5 inches 5cm 2mm
- Nominal :- categorical data with no inherent order or ranking
  ex. gender, color.
- Ordinal :- Categorical data with a meaningful order or ranking, but intervals betn ranks are inconsistent or unknown
- Interval :- Numeric data with meaningful and consistent intervals betn them without a ~~zero~~ true zero point
  ex. Temp. IQ, calender data

Midrange :- This is simply the avg of the largest of the largest and smallest values in the dataset

- Ratio: that has all properties of an interval variable, plus a true zero input point. This means you can meaningfully say that one value is "twice as much" as another

* Descriptive statistics

- Measures of central Tendency: A single value that represents the "centering" of a set of data, e.g. average.

1] Mean :- Average $= \frac{\sum_{i=1}^{n} (add)}{n}$

- Better if the data is normally distributed and there are no outliers.. used for interval and ratio data

2] Median : Better when the data is skewed (has extreme values). use for ordinal, interval, and ratio data.

3] Mode : useful for identifying the most common value or values in a dataset used in all the four scales. Best for categorical data.

- Measure of Dispersion

1] Range: Diff between the maximum value and the minimum value in the data set.

(i) Percentile (Relative) : ≠ Percentage (Absolute)
Percentile : A value below which certain percentage of observations lie
- k^{th} percentile = k% data is below it, and rest is above it
- Divides data into 100 equal parts

(ii) Quartile : Divides data into 4 parts
first quartile (Q1) = 25th percentile
Median = second quartile (Q2) = 50th percentile = medi
Third quartile (Q3) = 75th percentile

* Inter quartile range
= Q3 - Q1 = middle 50% data
- Handles outliers better than range, since the extreme values at both the ends are ignored in IQR
- since it uses percentiles rather than actual values, it is less affected by skewed dat
~~(see~~
- Outliers : Data points that are significantly outside of the typical range of values.
lower bound : Q1 - (1.5 * IQR)
upper bound : Q3 + (1.5 * IQR)

2] Variance :- How far data points from their mean ?
- population variance (whole data)

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

• sample variance (selected data)

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

n = no. of observation (sample size)

3] standard deviation :- square root of variance

- The scale of variance is not to the scale of the original data
- Scale of S.D is the same as scale of the original data

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

* coefficient of variation (cv)

$$cv = \frac{SD}{mean} \times 100\%$$

- lower cv → less variability relative to the mean
- Higher cv → more variability relative to the mean.
- cv is especially useful when comparing the relative dispersion of data sets with diff units or means.

* Problem solving with Analytics phases.

1. Recognizing a problem: Identify that a gap exists bet^n the current state and desired outcome. It involves observing symptoms or inefficiencies that need resolution.

Date / /
Page

2] Defining the problem: clearly articulate the problem's scope and objectives. This helps in focusing efforts and avoiding vague or misdirected analysis.

3] Structuring the Problem: Break the problem into smaller, manageable components. Establish variables, relationships and assumptions for analysis

4] Analyzing the Problem: Use data analytics tools and techniques to explore, model, and test hypotheses. This phase reveals patterns, causes and potential solutions.

5] Interpreting Results and Making Decision: Draw insights from the analysis and evaluate alternatives. Choose the best course of action based on evidence and buisness goals.

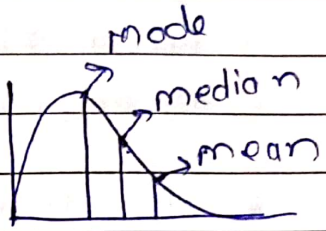
6] Implementing the Solution: Apply the chosen solution in the real-world setting. Monitor outcomes and adjust if needed to ensure effectiveness.

* Skewness - Measures and Interpretation
- Skewness is key statistical measure that shows how data is spread out in a dataset.

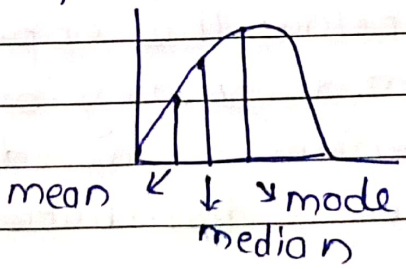
- It is imp because it helps us to understand the shape of the data distribution which is imp for accurate data analysis and helps in identifying outliers and finding the best statistics method to use for analysis.

Types of Skewness

1] Right skewness
- positive skewness
- right tail longer.
- Most data points are on left with few large values pulling the distribution to the right.
relationship: Mean > Median > Mode



2] Negative skewness (LEFT skewness)
- left tail longer
- most data points are on right with few large values puling the ~~distan~~ distribution to the left.
Relationship: Mean < Median < Mode.



3) Zero skewness (Symmetric Distribution)
- Zero skewness shows a perfectly symmetrical distribution where mean, median and mode are equal. Mean = Median = Mode

Tests of skewness

- **Skewness coefficient (Pearson's first coefficient of skewness)**

→ ~~This a skewness~~ This is a numerical measure of skewness based on the relationship between the mean and mode. It helps us to find if the data is skewed when the mean and mode are not equal.

- Positive skew: If the mean is greater than the mode.
- Negative skew: If the mean is smaller than the mode.
- Zero skew: If the mean is equal to the mode, ~~the~~

* Cumulative Relative Frequency Distribution

- The cumulative relative frequency represents the proportion of the total no. of observations that fall at or below the upper limit of each group. A tabular summary of cumulative relative frequencies is called a cumulative relative frequency distribution.

* Cross Tabulation

- used to summarize categorical data and examine the relationship bet^n two categorical variables is cross tabulation.

- is tabular method that displays the no. of observations in a data set for diff subcategories of two categorical variables
- also called contingency table.
- subcategories of the variables must be mutually exclusive and exhaustive, meaning that each observation can be classified into only on subcategory and taken together over all subcategories, they must constitute the complete data set

ex. Data

Person	Gender	Preferred product
1	Male	A
2	female	B
3	female	A
4	Male	C
5	male	A
6	female	B

Cross-Table.

	A	B	C
Male	2	0	1
female	1	2	0

* A standardized values (z-scores)
- A standardized values, commonly called z-score, provides a relative measure of the distance an observation is from the mean, which is

independent of the units of measurement. The z-score for the ith observation in a data is calculated as follows:

$$z_i = \frac{x_i - \bar{x}}{s}$$

* Descriptive statistics for categorical Data:

- The proportion
- statistics such as means and variances are not appropriate for categorical data.
- proportion $= \frac{x}{n}$

where x = no. of observat^n having a certain characteristic
n = sample size

Note: - proportions are analogous to relative frequencies for categorical data

* Covariance
- is a measure of the linear association bet^n two variables, x and y.
- diff formulae for popul^{at} and samples
- Quantifies how changes in one variable relate to changes in another.

$$cov(X,Y) = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(Y_i - \mu_y)}{N}$$

For sample $= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

ex.

Height	Weight
65	68
67	69
68	70
66	69
64	65

mean Height($\bar{x}$) = 66

mean Weight($\bar{y}$) = 68.2

- For Height $(x_i - \bar{x})$ = 65 − 66 = −1
  67 − 66 = 1
  68 − 66 = 2
  66 − 66 = 0
  64 − 66 = −2

- for weight $(y - \bar{y})$ = 68 − 68.2 = −0.2
  69 − 68.2 = 0.8
  70 − 68.2 = 1.8
  69 − 68.2 = 0.8
  65 − 68.2 = −3.2

$$cov = \frac{(-1 \times 0.2) + (1 \times 0.8) + (2 \times 1.8) + (0 \times 0.8) + (-2 \times 3.2)}{4}$$

$$= \frac{11}{4} = 2.75$$

Note :- The value of covariance depends on the scale of variables. It is not standardized like correlation, so you can't directly say whether a value is 'high' or 'low' unless you interpret it in context.

* covariance measure direction of the linear relationship between two vari

- (+) ~~variable~~ covariance :- as one variab increases, the other trends to increase
- (-) covariance :- as on increases, the other tends to decrease

* Correlation
- measure of the linear relationship of two variables, which does not depend on the units of measurement
- measured by correlation coefficient, also known as the pearson product ~~mean~~ moment correlation coefficient

$$\text{correlation}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

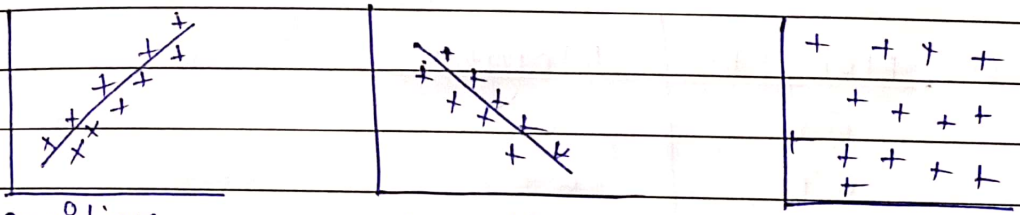
- ~~correlation~~ coefficient (r)

(a) Pearson correlation coefficient : Measures linear relationships between continuous variables

(b) Spearman Rank correlation coefficient: Measures relationships, even if they are not strictly linear

(a) Pearson correlation coefficient (r)

- measures strength and direction of the linear relationship between two continuous variables
- It ranges from -1 to 1
- +r means as one variable increases, the other ~~bands~~ tends to increase
- -r means as one variable increases, the other tends to decrease.
- Pearson correlation assumes
  - variables are continuous and normally distributed
  - No significant outliers



Positive correlation | Negative correlation | No correlation

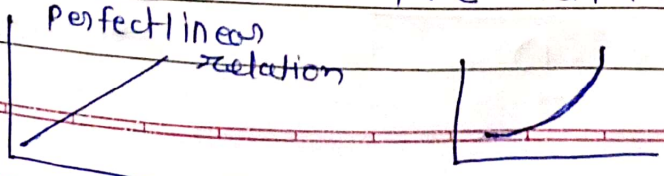
ex (covariance example)

$$corr(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$= \frac{2.75}{1.58 \times 1.92}$$

$$= 2.090$$

- Works well when data is linear, but not well when the data is not linear



(b) Spearman Rank correlation
- measures the strength and direction of a monotonic relationship betn two ranked (ordinal or continuous) variable
- It is non-parametric (distribution free)

$$r_s = 1 - \frac{6 \sum d^2}{(n^2 - 1)}$$

d^2 = square of differences in the ranks of the two co-ordinates for each point (x, y)
n = no. of data points

eg

Height	Weight
65	68
67	69
68	70
66	~~70~~ 69
64	65

x_i	R_x	y_i	R_y	$d_i = R_x - y_y$	d_i^2
65	2	68	2	0	0
67	4	69	3.5	0.5	0.25
68	5	70	5	0	0
66	3	69	3.5	-0.5	0.25
64	1	65	1	0	0

$$r_s = 1 - \frac{6 \times 0.50}{(5^2 - 1)} = 1 - \frac{3}{24} = \frac{21}{24} = 0.87$$

* Sample vs Population

- Population :- The entire group you're interested in studying (e.g. all students in India)
- sample :- A subset of the population, used to draw conclusions about the population.

* Univariate and Bivariate sampling

(a) Univariate sampling
- Data collected involves one variable
- To analyze the distribution or central tendency of that single variable.
- example :- survey of height of students

(b) Bivariate sampling
- Data collected involves two related variables.
- To analyze the relationship or correlation betⁿ them.
- ex :- survey of height and weight of students

* Re-Sampling
- is a statistical method that involves drawing repeated samples from the original data.
- Often using computers - to assess variability or improve accuracy.
- useful when the theoretical distribution is unknown.
- Doesn't rely on strict statistical assumptions
- Helpful in model validation (e.g. cross-val in ML)

* Techniques:

(a) Bootstraping :- Sampling with replacement from the original sample to estimate confidence intervals or standard errors

(b) Jackknife :- Systematically leaving out one observation at a time to estimate variability

(c) Permutation Test : Randomly shuffle labels or values to test a hypothesis

* Oversampling :- Increase the size of the minor class by duplicating samples or generating synthetic data

ex: Smote (Synthetic Minority Oversampling Technique): Handles imbalanced datasets, where the minority class has significantly lower samples than the others e.

* Sampling Techniques

1) Probability sampling :- Random Techniques
- Every member of the population has a chance of being selected.
- Most appropriate
- Mainly used in quantitative search

• Types of Probability sampling:

(i) Simple Random sampling :-
- Every individual has an equal chance of selection (like lottery)

(ii) Systematic sampling :-
- Select every k^{th} item (e.g. every 10^{th} person)

(iii) Stratified Sampling :-
- Population is divided into strata (groups), and samples are taken from each

(iv) cluster sampling
- Population is divided into cluster (e.g. cities) and a few clusters are fully surveyed.

b] Non-probability sampling :- Non-Random selection
- Not every individual has a chance of being included
- Easier and cheaper to access, has a higher risk of sampling bias
- Used in qualitative research

• Types of Non-probability sampling:

(i) Convenience sampling
- Sample is taken from what is easily available

(ii) Judgmental sampling
- Researcher selects what they think is a representative sample

(iii) Snowball sampling
- Existing participants recruit future subjects (used for hidden populations)

(iv) Quota sampling
- Sample has quotas to match population characteristics (e.g. 50% male / female)

* Sample Space and Events

1. Sample space (S)
- The sample space is the set of all possible outcomes of a random experiment.

ex. Tossing a coin
$S = \{Heads, Tails\}$

» Rolling a 6-side die:
$S = \{1, 2, 3, 4, 5, 6\}$

2) Event (E)
- An event is a subset of the sample ~~data~~ space. It includes one or more outcomes of interest.

ex.
Event: getting an even no. when rolling a die
$A = \{2, 4, 6\}$

• Types of Events:
(a) Simple Event : An event with one outcome
e.g. rolling a 4 : $\{4\}$

b) Compound Event :- An event with multiple outcomes
e.g. rolling even no : {2, 4, 6}

c) Certain Event :- The event that always occur
e.g. (getting a no. betn 1-6 when die roll)

d) Impossible Event : The event that can never happen
e.g (rolling 7 on a 6-sided die)

e) mutually Exclusive : Events that cannot happen together
e.g. getting heads and tails in one coin toss)

f) Independent Event : Events where the outcome of one does not affect the other

* Probability
probability may be defined from one of three perspectives.

1) classical (Theoretical) Approach
- probability is the ratio of favorable outcomes to the total no. of equally likely outcomes

$$P(E) = \frac{\text{No. of favorable outcomes}}{\text{total no. of possible outcomes}}$$

ex. Rolling a fair - 6sided die :
- Event: getting a 4

$$P = \frac{1}{6}$$

2) Empirical (Experimental) Approach
- based on actual experiments or observation
- It is relative frequency of an event occurring in repeated trials.

$$P(E) = \frac{\text{No. of times event E occurs}}{\text{total no. of trials}}$$

ex. toss a coin 100 times.
- Head appeared 53 times

$$P(\text{Heads}) = \frac{53}{100}$$

3) Axiomatic (Modern Approach)
- this approach defines probability based on a set of axioms (rules) without assuming anything about the nature of outcomes.

key axioms

1. $P(E) \geq 0$ for any event E (probabilities are non-negative)
2. $P(S) = 1$ (the probability of the sample space is 1)
3. If events A and B are mutually exclusive, then:

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

* Marginal probability
- is probability of single event, regardless of the outcome of other variables.

P(A) or P(B)

ex	Science	Arts
Male	30	20
Female	40	10

$$P(\text{Female}) = \frac{50}{100} = 0.5$$

$$P(\text{Science}) = \frac{70}{100} = 0.7$$

* Joint Probability:
- Event A and B are happening together, whether they are independent or dependent: $P(A, B)$

• Independent Event: If A and B are independent, $P(A, B) = P(A) \times P(B)$

• Dependent events: If A and B are dependent, $P(A, B) = P(A) \times P(B|A)$, where $P(B|A)$ is the conditional probability of B given A

ex. A: student is Female
B: Student studies science

$$P(A \cap B) = \frac{40}{100} = 0.4$$

* Conditional probability
- is the probability of occurrence of one event A, given that another event B is known to be true or has already occurred.

conditional probability of an event A given that event B

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$P(A \text{ and } B) = P(A|B)\,P(B) = P(B|A)\,P(A)$ is called the multiplication law of probability.

Bayes theorem

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$

- $P(A|B)$: posterior probability
- likelihood: $P(B|A)$
- Prior Probability of A: $P(A)$
- total probability of B: $P(B)$

Q. Bag A = 2 red balls and 3 blue balls
Bag B = 4 red balls and 4 blue balls.
You know that you picked the ball from Bag B, what is probability that you picked the ball from Bag B, given that the ball is red?

→ given

$P(A) = 0.5$

$P(B) = 0.5$

$P(\text{Red}|A) = \frac{2}{5}$

$P(\text{Red}|B) = \frac{4}{5}$

→ $P(\text{Red}) = P(\text{Red}|A) \cdot P(A) + P(\text{Red}|B) \cdot P(B)$

$= \frac{2}{5} \times \frac{1}{2} + \frac{4}{5} \times \frac{1}{2}$

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

$P(B|\text{Red}) = \frac{P(\text{Red}|B) \cdot P(B)}{P(\text{Red})} = \frac{4/5 \cdot 1/2}{3/5}$

$= \frac{2}{5} \times \frac{5}{3} = 2/3$

Date / /
Page L&W

* Random variables

- A random variable is a numerical outcome of a random experiment
- It assigns a no. of each possible outcome of a probabilistic event.

Two type:

(a) Discrete :- Takes countable values (usually whole numbers)

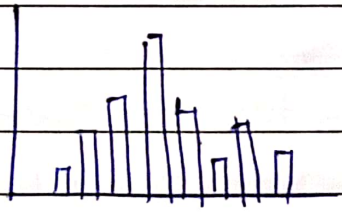
(b) continuous :- Takes uncountable / infinite values within a range

* Probability Distribution

- Distribution :- Describes all the probable outcomes of a variable.
- tells you how likely each value is

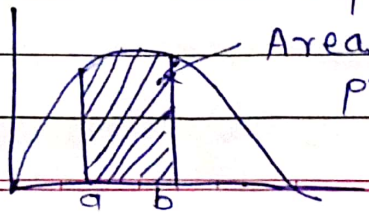
(a) Discrete distribution :

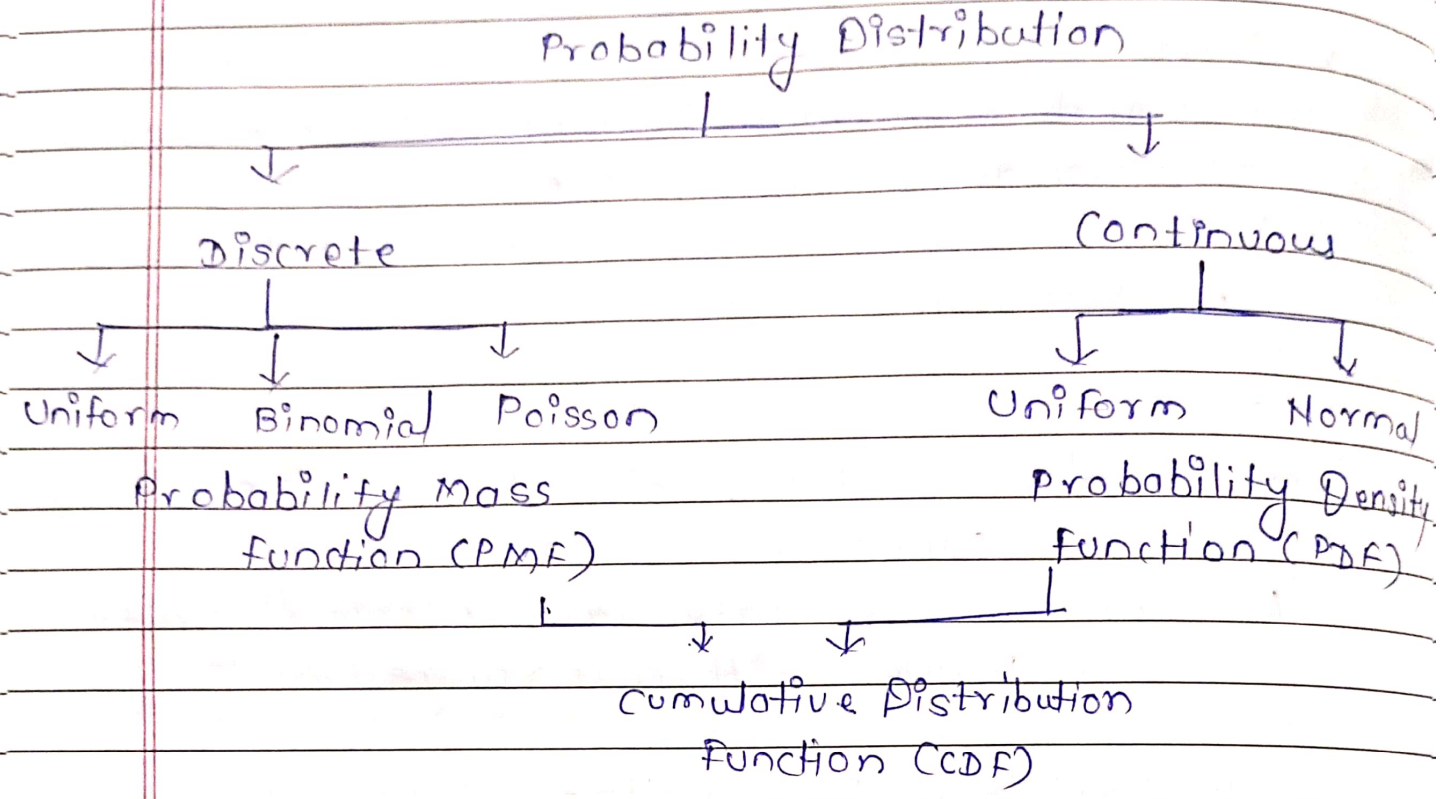
- Sum of all individual probability = 1
- the probability distribution of the discrete outcomes is called a probability mass function



(b) continuous distribution :

- Total area under the probability curve = 1





1] Discrete.

(a) Uniform

- A discrete uniform distribution is a probability distribution where each outcome in a finite set is equally likely to occur.

Formal def:- If a random variable X can take on n distinct values $x_1, x_2, \ldots x_n$ and each value has the same probability then:

~~$P(x = x)$~~

ex:- Rolling a fair 6-sided die:

Possible values: $X = \{1, 2, 3, 4, 5, 6\}$

$P(x = x) = \frac{1}{6}$ for $x = 1$ to 6

Binomial Distribution

b] Bernoulli Trial: An experiment with only two possible outcomes. success or failure.
many such bernoulli trials = Binomial distribution.

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

n = total no. of trails
k = no. of successes
p = probability of success

ex. suppose you flip a fair coin 5 times. What is the probability of getting exactly 3 heads.
n=5 k=3 p=0.5

$$P(X=3) = {}^{5}C_{3} \times (0.5)^{3} \times (0.5)^{2}$$
$$= 10 \times (0.5)^{5}$$
$$= 0.3125$$

properties.
- mean (el) = $n \cdot p$
- variance (σ^2) = $\sigma^2 = n \cdot p \cdot (1-p)$

c) Poisson Distribution
- The poisson distribution is a discrete probability distribution that models the no. of times an event occurs in a fixed interval of time or space, given a known average rate and that the events happen independently.

$$P(x=k) = \frac{\lambda^{k} \cdot e^{-\lambda}}{k!}$$

X: no. of events (k = 0, 1, 2...)
λ = avg rate (mean no. of occurences in the interval)
$e \approx 2.71\epsilon$ (Euler's no.)

ex. If a store receives an avg of 3 customers per hour ($\lambda = 3$), what is the probability that exactly 5 customers arrive in the next hour?

$$P(X=5) = \frac{3^5 \cdot e^{-3}}{5!} = \frac{243 \cdot e^{-3}}{120} \approx 0.1008$$

- property

Mean (μ) = λ

Variance (σ^2) = λ

d] Geometric distribution

- Used to modal number of trials till we get the first success

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

ex. probability of getting first head on the third trial = $PMF(x=3) = (0.5)^{3-1}(0.5$

$= 0.25 \times 0.5$

$= 0.125$

2] continuous probability distribution.

- curve that characterizes outcomes of a continuous random variable is called a probability density function

a] Uniform Distribution

- The uniforom distribution characterizes a continuous random variable for which all out comes betn some minimum and max value are equally likely.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

and cumulative distribution function is

$$f(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \le x \le b \\ 1, & \text{if } b < x \end{cases}$$

b] Exponential Distribution

- used to model the time between events in a poisson process (i.e, events that occur randomly and independently at a constant avg rate.)

$$f(x) = \lambda e^{-\lambda x}, \text{ for } x \ge 0$$

where:

x : time or distance betn events

λ : rate parameter (avg no. of events per unit time)

e : Euler's no. (~ 2.718)

ex. suppose a call center receives 5 calls per hr. What is the probability that the next call arrives within 10 minutes?

→ 10 min = $\frac{1}{6}$ hr.

$\lambda = 5$

$$P(x \le 1/6) = 1 - e^{-5(\frac{1}{6})} = 1 - e^{-5/6} \approx 0.565$$

* properties

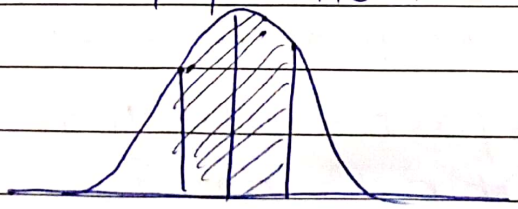
mean(μ) = $1/\lambda$

variance(σ^2) = $1/\lambda^2$

c) Normal distribution
- also known as Gaussian distribution.
- is symmetrical and bell-shaped curve

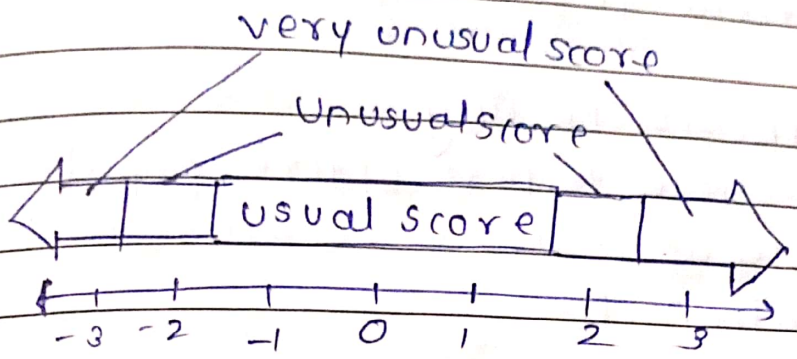
properties:

1. The distribution is symmetric hence skewness = 0
2. The mean, median, and mode are all equal. Thus half the area falls above the mean and half falls below it.
3. The range of x is unbounded, meaning that the tails of the distribution extend to -ve and +ve infinity.
4. Empirical rule:
   - 68% of the data falls within mean $\pm$ SD
   - 95% of the data falls within mean $\pm$ 2 SD
   - 99.97% of the data falls within mean $\pm$ 3SD
5. Z score (standard score): Useful in finding relative position of an observation with respect to the overall population



* Standard Normal Distribution :- is a special case of the normal distribution with:
  - Mean (μ) = 0
  - Standard Deviation (σ) = 1

* z-score



- tells you how many standard deviations a data point is from the mean of a distribution.
- It standardizes values so they can be compared on the same scale, even if the original data had different units or scales.

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

- $z = 0$ # value is exactly avg
  $z > 0$ # value is above the mean
  $z < 0$ # value is below the mean

* Estimation :- is the process of using sample data to infer or estimate population parameters (like mean, variance, proportion).

Type :-

a) Point Estimation : Gives a single best estimate of a parameter (e.g. sample mean for population mean)

b) Interval Estimation : Gives a range where the parameter is likely to lie (e.g. confidence interval)

* Statistical Inference

* Hypothesis Testing
- Not possible to test entire population, so test sample and draw conclusions about the population.

- Hypothesis-Testing Procedure
Conducting a hypothesis test involves several steps:
1. Identifying the population parameter of interest ~~of~~ and formulating the hypothesis to test
2. Selecting a level of significance, which defines the risk of drawing an incorrect conclusion when the assumed hypothesis is actually true
3. Determining a decision rule on which to base a conclusion.
4. Collecting data and ~~cleaning~~ calculating a test statistic
5. Applying the decision rule to the test statistic and drawing a conclusion.

• Significance level: Alpha (α)
- is used to decide whether to reject the null hypothesis in a hypothesis test.

ex. You test whether a new medicine works better than the old one.
• H_0: New medicine is not better.
• H_1: New medicine is better.
You choose $\alpha = 0.05$

If the p-value < 0.05, you reject H_0 and conclude that the medicine is better with 95% confidence.

If p-val
$< \alpha$ ⟶ Reject the null hypothesis
$> \alpha$ ⟶ Do not reject the null hypothesis

Note
confidence level ↑ ∝ ↓ critical z-value ↑ Range of acceptance ↑

- Two types of errors can occur in this decision-making:

1. Type I Error (False (+)ve)
- Rejecting the null hypothesis when it is actually true.
- It's like raising a false alarm.
- also known a Alpha (α)
ex. You test a new drug and conclude its effective, but it's actually false.
- It's +

2. Type II (False Negative)
- Failing to reject the null hypothesis when it is actually false.
- It's like missing a real effect.
- known = β
ex. You test a new drug and conclude it's not effective, but is actually is.

key relationships

- Lowering α reduces Type I error but may increase Type II error.
- Increasing sample size can reduce both errors.
- Good test design aims to balance both errors.

* What are Tailed sets?
- tails refer to the extreme ends of the probability distribution where we check for statistical significance.

1] One-tailed test.
- A test where the critical region (rejection zone) is in only one tail of the distribution - either left or right.

ex. claim: New drug increases recovery rate.
Hypotheses:
H_0: recovery rate $\leq$ 70%.
H_1: recovery rate $>$ 70%. → Right-tail

2] Two-tailed test
- A test where the critical regions are in both tails of the distribution.

ex. claim: New medicine produces exactly 500ml bottles.
Hypotheses:
H_0 mean = 500ml
H_1 mean $\neq$ 500ml → Two-tailed

Date / /
Page L&W

* t-Test

- A t-test is a statistical hypothesis Test used to determine whether there is a significant diff betn
  - The means of two groups, or
  - A sample mean and a known / population mean

Types

a] A one-sample t-Test.

- A one-sample t-Test is used to determine whether the mean of a sample is significantly different from a known or hypothesized population standard deviation is unknown

where Ho: sample mean ($\bar{x}$) = population mean(μ)

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

$\bar{x}$ = sample mean
μ = population mean
s = sample sd
n = Sample size.

- This is the same as z-test formula, but it is expected that here, the sample size should be ≤ 30

2] Two Independent samples T-Test

- Compare the mean of two independent groups (i.e Samples from two diff populations)
- Ho: There is no significant diff between the means of the two groups.

- formula: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- $\bar{x}_1$ and $\bar{x}_2$: sample means
  s_1 and s_2 : Sample S.D.
  n_1 and n_2 : sample sizes

3] Paired t-test
- Data from the same sample is tested before and after some event
- Different from two independent samples t-test (where data comes from two completely diff groups)
- Ho:- There is no significant diff betn the means of the two groups
- $t = \frac{\bar{x}\ diff}{\left(\frac{s\ diff}{\sqrt{n}}\right)} = \frac{\text{sample mean of the difference}}{\left(\frac{\text{sample standard deviations of the differences}}{\sqrt{\text{sample size}}}\right)}$

4] ANOVA:- Analysis of variance.
- It's a statistical test used to compare the means of three or more groups to determine if there's a statistically significant difference betn them.
- Looks at variance:- It achieves this by examining the variation within each group and comparing it to the variation betn the groups.

- Within-group variance: How much the data points within a single group differ from that group's mean.

- Betn-group variance: How much the means of different groups differ from the overall mean of all data.
- H_0 = All group means are equal.

* Chi-square (χ^2) test:
- is a non-parametric statistical test used to analyze categorical data. Unlike tests like ANOVA or t-tests that compare means of continuous data, the Chi-square test focuses on frequencies and proportions within different categories.

- The core idea of the Chi-square test is to compare observed frequencies (what you actually see in your data) with expected frequencies (what you would expect to see if there were no relationship or diff betn the variables being tested, based on the null hypothesis).

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where
c = Degrees of freedom
O = observed value
E = Expected value

Degree of freedom = (Row-1) * (col-1)

* Three types:

1] Chi-square test of Independence (Are two categories independent? : Are gender and pet preference related?)
- It checks if the occurence of one variable is independent of the occurence of the other.

2] Chi-Square of Homogeneity: (Is data within a category proportionate? : Do diff servers of a company fail roughly at the same rate?)
- is a statistical test used to determine if the distribution of a single categorical variable is the same across two or more diff populations or groups.

3] Chi-Square test of Goodness of Fit: (Does data fit a particular distribution? : Is tossing a fair coin following binomial distribution?)
- it checks if your observed frequencies for different categories are significantly diff from what you hypothesized or expected.
- Low chi-square value - High correlation bet^n Observed and expected value

~~Observe~~

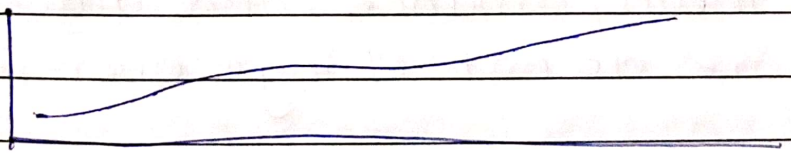
Date / /
Page

Time series.

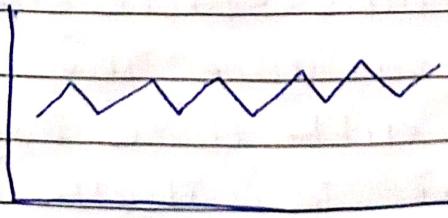
- sequence of data points: It means a collection of individual measurements or observation
- Indexed or listed: Each ~~steps~~ data point has a corresponding time stamp. This time stamp is crucial because it defines the order of the data points.
- In time order: The observations are arranged chronologically. This is the defining characteristic of a timeseries, as the order of the data is not arbitrary but dictated by when the data was collected.

Characteristics of Timeseries.

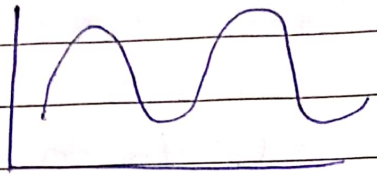
- Trend: A long-term increase or decrease in the data. Think of the general upward trend in global temp over decades.



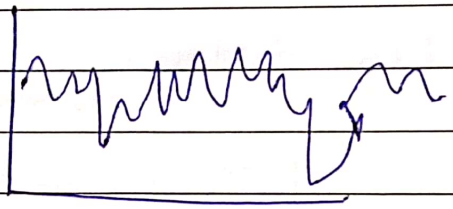
- Seasonality: Regular, predictable patterns that repeat over a fixed period (e.g. daily, weekly, monthly, yearly). Retail sales often show seasonality, peaking during holidays.



- Cyclicality: Fluctuations that are not of a fixed period, but rather associated with economic or other cycles. These are typically longer than seasonal patterns.



- Irregularity/Noise: Random, unpredictable variations that can't be explained by the other components.



* Stationary time series
- is a time series whose statistical properties remain constant over time. This means that its mean, variance, and autocorrelation structure do not change over time.

- key characteristics
1. constant mean: The avg value of the series stays the same throughout time.
2. Constant Variance: The spread (volatility) of the series remains stable over time
3. constant Autocovariance: The relationship betn values at diff time lags depends only on the lag itself, not on time

* ARIMA (AutoRegressive Integrated moving Average)

- ARIMA is powerful statistical model used for time series forecasting. It combines three key components:

1. AR (AutoRegressive)
- Uses past values to predict the current value.
- parameter: P (number of lag observations).

2. I (Integrated)
- Represents differencing to make the time series stationary.
- parameter: d (no. of differencing needed)

3. MA (Moving Average)
- uses past forecast errors in a regression-like model.
- Parameter q (no. of lagged forecast errors) (use ACF to calculate q)

* ACF (Auto correlation function (ACF) is a regression model that tells us about the correlation of y with its own lags, i.e
- Between y and lag 1 y
- Between y and lag 2 y
- Between y and lag 3 y

* PACF:- Partially Auto correlation Function (PACF) is similar to ACF. Conveys the relationship of

y with lags, but after removing the effects of the intermediate lags

* SARIMAX : Seasonal Autoregressive Integrated Moving Average with exogenous variables

- SARIMAX is an extension of ARIMA that supports:
  - Seasonality
  - Exogenous variables (external predictors)

- SARIMAX (p,d,q)(P,D,Q,s)

- (p,d,q) : Non-seasonal ARIMA parameters
  - p : # of diff of autoregressive terms
  - d : # diff of non-seasonal differences
  - q : diff of moving avg terms

- (P,D,Q,s) : Seasonal components
  - P : diff of seasonal autoregressive terms
  - D : diff of seasonal difference
  - Q : diff of seasonal moving avg terms
  - s : seasonality period (e.g., 12 for monthly data with yearly seasonality)

- Exogeneous variables : optional variable that influence the series (e.g. temp. holidays, marketing campaigns)

* Predictive analysis (PPT)