DAV - Last Min Notes

Owner	N nithu
Tags	

Unit 1 - Data Science Process

Facets of data: Structured,unstructured(email), Natural Language(linguistical knowledge), Multimodal, Machine generated, Graph based and streaming

Data Sci.Lifecycle: Business Understanding, Data Acquisition(Data source, Pipeline, Environment, Wrangling and Cleaning) and Understanding, Modelling (Feature engn. ,Model Training and Model evaluation)and Deployment.

Data Science Process:

1. Setting the research goal(understand the goals and context of your research and prep of project charter)
2. Retrieving data: Internal Data(data retrieval and ownership), External data

   ETL → Extract, Transform and Load

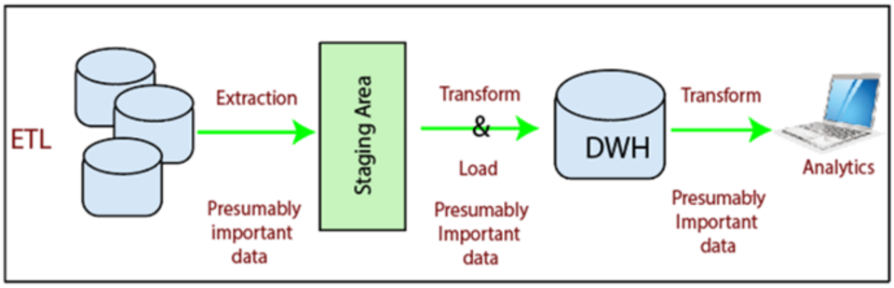
   Adv: Improved data qty.

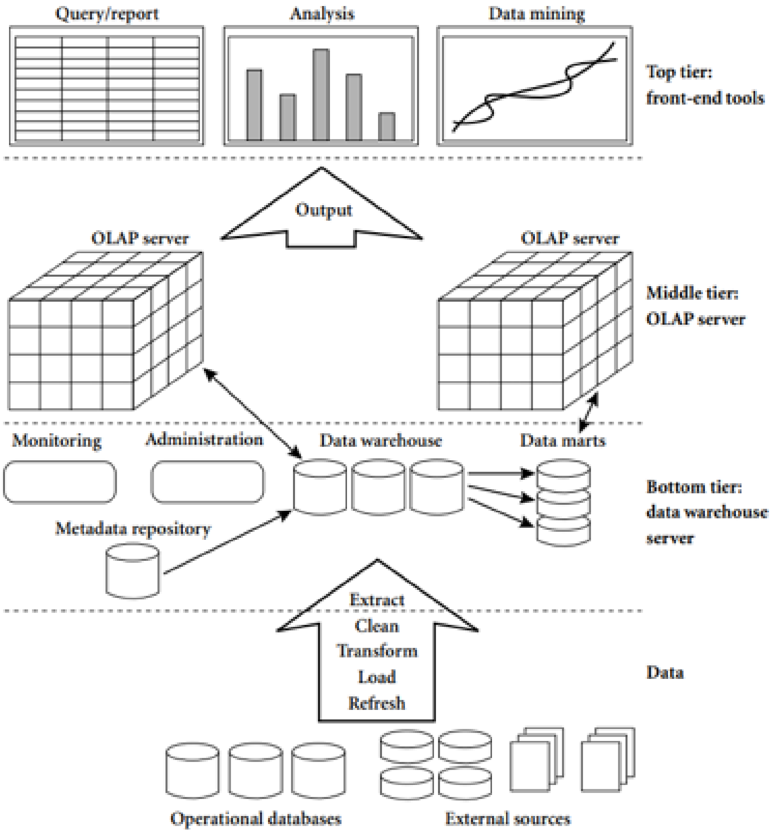
   Better data integration

   Data secu.

   Improved scalability

   Increased automation


ETL
Extraction
Presumably important data
Staging Area
Transform & Load
Presumably Important data
DWH
Transform
Presumably Important data
Analytics



Exploratory Analysis: Learning from data, stats like distributions, correlations and outliers.

3. Data prep(enhance qty of data)

   Cleansing,Transformation, Combining data

   An outlier is an observation that seems to be distant from other observations or, more
   specifically, one observation that follows a different logic or generative process than
   the other observations

   Noise: Random error or variance in a measured variable.

4. Data exploration(EDA)
5. Data modelling
6. Presentation and automation

Distributed file systems
A distributed file system is similar to a normal file system, except that it runs on multiple servers at once.

Eg: Hadoop File System

Once you have a distributed file system in place, you need to add data. You need to move data from one source to another, and this is where the data integration frame works such as Apache Sqoop and Apache Flume excel.

The agile project model is an alternative to a sequential process with iterations.

A Pareto diagram is a combination of the values and a cumulative distribution.

Model fit—For this the R-squared or adjusted R-squared is used. This measure is an indication of the amount of variation in the data that gets captured by the model.

Unit 2 - Freq, Normal Distributions and Regression

Frequency Distributions:

Frequency distribution of various outcomes(how many times they occur)

Identify patterns(how is the distribution , above or below a parameter)

- Grouped→ then sort and group the observations
- Ungrouped → take the individual observations and find the freq
- Relative - based on freq of the event/total number of events(here sum of the relative freq should equal 1)
- Cumulative(Less than and more than type)

Variability:

Spread or dispersion

→ Range

→ IQR(measure of spread-out from the middle set)
→ variance(rough idea on spread-out)→ small means tight packing

→ SD → tight clustering around the mean(small SD means taller gaussian curve , else spread apart)

Variability: Number of inconsistencies in the data

Low variability: Better prediction

else:
Values are less consistent

C.V = S.D/Mean *100

When to use: Say in one dataset , height is measured in cm and in other it is measured in metres, means comparing is hard.

Normal Distribution: Cont.prob.dist that is symmetrical around mean. Unimodal. No room for skewed distributions

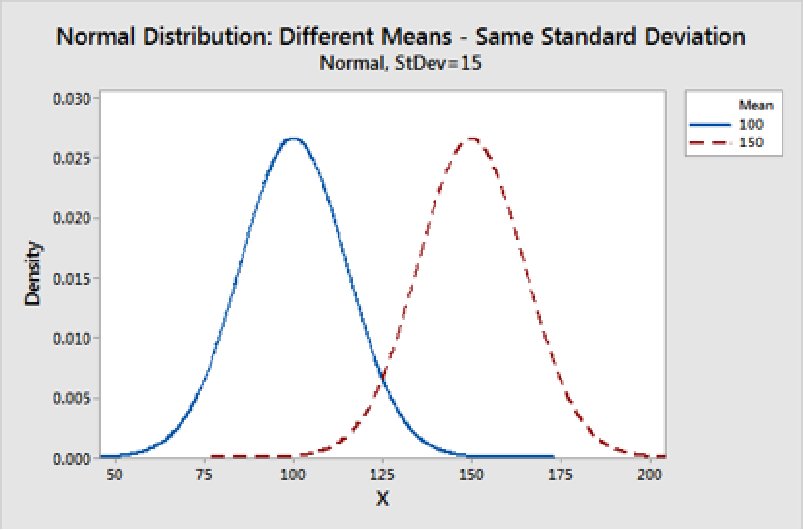
Mean=Median=Mode

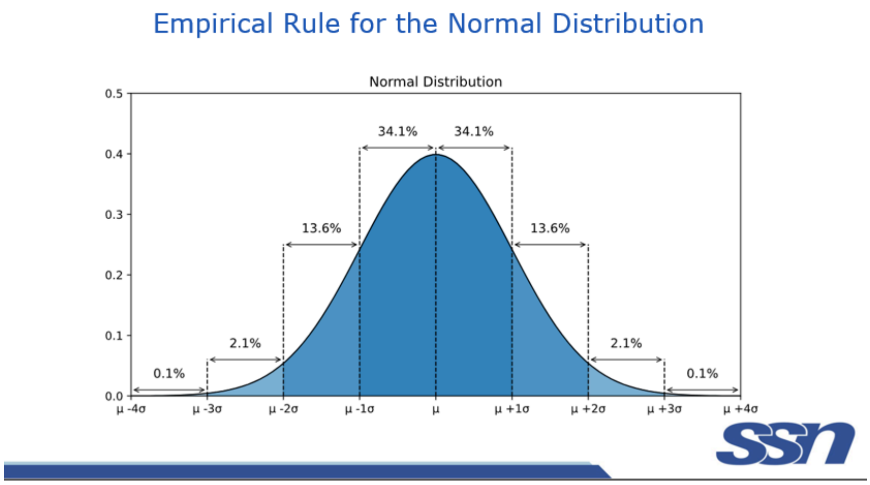
- If continuous data are skewed use Weibull, lognormal, exponential, or gamma distribution.

Clustering around the central peak

Basically , the least ends and the most ends are equally infrequent and the avg> and avg< mean is same.

- Parameters: Mean and S.D/Variance
- Taller G.C → smaller variance/S.D
- Shorter G.C → larger variance/S.D





Empirical Rule for the Normal Distribution

68 percent within mean-1*S.D and mean+1.S.D

68+26=94 percent within -2*S.D and +2*S.D

the rest are at the ends

When mean =0, S.D =1 ⟶Z-dist/Std.Normal Dist

- A value on the standard normal distribution is known as a standard score or a Z-score. A standard score represents the number of standard deviations above or below the mean that a specific observation falls.
- The mean has a Z-score of 0.
- Z-score ⟶ tells us where a specific obs. falls relative to the entire norm.dist
- A positive Z-score(more than the avg) else less than the avg

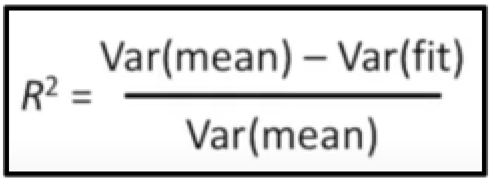
$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

Regression:

Linear rel. from dependent and independent variables to predict the cont. dependent variables from the independent

How? Best fitting line. Fit a line.. improve upon it.. and repeat

For the formulas, use Prob and Stats. That is better



$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

Var(mean)=10; Var(fit)=0; R^2 value is 100%(at max)

Mean X can completely explain Y

Multi-regression: Several indep.var and one dependent var

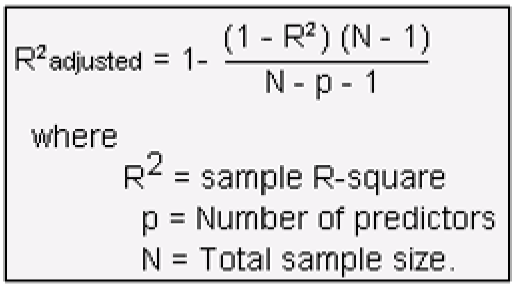
Prob with R^2 → no effect on an additional cause, that can really alter/ increasing R^2 in case of independent variables bringing no change → **Adjusted R^2**

Correct the overestimation.

Adjusted R^2 ≤ R^2

if Adj.R^2=1, perfect prediction

if val ==0, nothing predictive from the model



$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

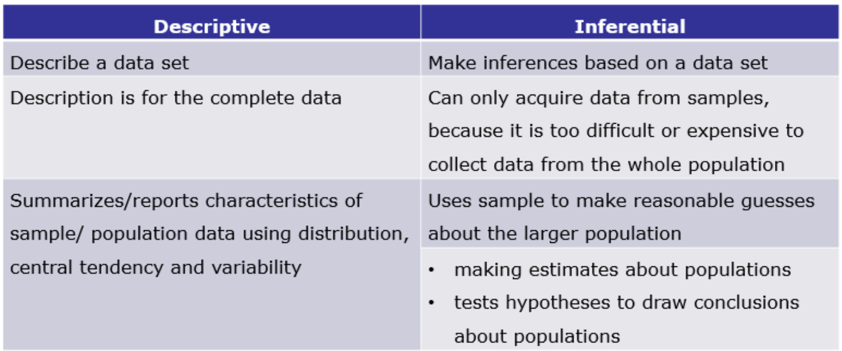
N = Total sample size.

S.E → amt of discrepancy that can be expected.

Widely scattered data→ higher score

0→ perfect match

Unit-3 - Hypothesis, Inferential and Descriptive Stats, z-test, one and two tailed tests



Descriptive	Inferential
Describe a data set	Make inferences based on a data set
Description is for the complete data	Can only acquire data from samples, because it is too difficult or expensive to collect data from the whole population
Summarizes/reports characteristics of sample/ population data using distribution, central tendency and variability	Uses sample to make reasonable guesses about the larger population
	• making estimates about populations • tests hypotheses to draw conclusions about populations

Population: The broader data

Sample: Specifics out of the population

- **Simple random Sampling: Random ass(fair chance)**
- **Stratified Random: Split into groups then random ass**
- **Cluster: Group random and get every member from some groups**
- **Systematic random sample: Random order - every nth member in the sample**
- **Convenience: Non-random (biased)**
- **Voluntary: Put out and let it on the ppl to decide whether they are interested**

Hypothesis:

A testable statement that proposes something(relationship b variables)

Prediction:

Measurable outcome that is expected

- **Theory**: A well-supported explanation of a phenomenon, often based on multiple pieces of evidence.

TYPES:
Simple- straight forward if then case

Complex: More than two dep and inde.variables

Null: Proposes No relationship

Alternative: Opp to null

Logical: Logical reasoning instead of facts

Empirical: Tests, trials⟶ so conditions can change(opp to logical)

Statistical: Tested and statistically verified

Type I error(alpha) → true negatives

Accept null hypothesis = 1-alpha

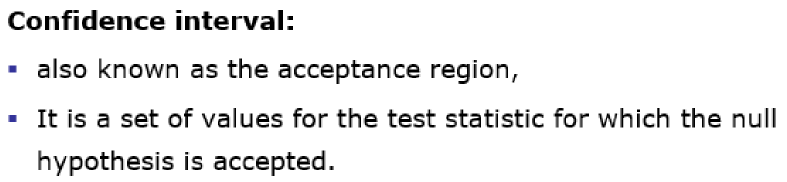
Type II error(beta) → false positives

Reject null hypothesis=1-beta

Level of Significance(LOS) = Prob of Type I error(alpha)

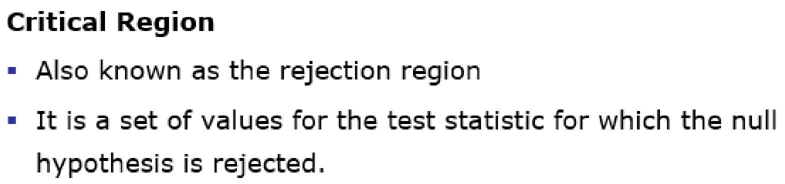
Power: 1-beta and 1-alpha should be high

Left-tailed: alternative hypo. says that true val less than the null hypothesis else Right-tailed tests



Confidence interval:

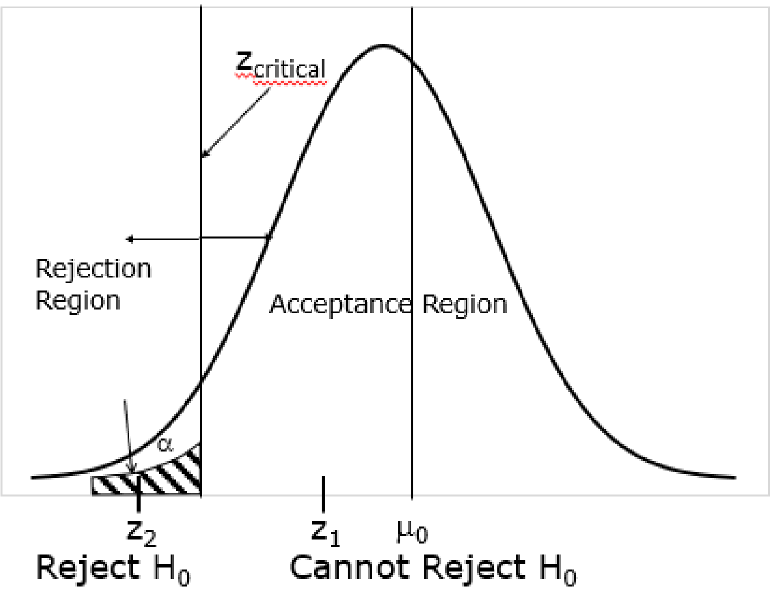
- also known as the acceptance region,
- It is a set of values for the test statistic for which the null hypothesis is accepted.



Critical Region

- Also known as the rejection region
- It is a set of values for the test statistic for which the null hypothesis is rejected.

Keywords



One -Sample z test:

One sample grp(sample mean with hypothesized val.)

→Whether a sample comes from a known population

$$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

Two-sample (Paired) z-test: If two grps are equal or different

$$Z = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}}$$

- P-value is a probability value which is set as threshold that should be satisfied by the sample taken so that it supports the null hypothesis giving sufficient evidence (sample) to accept the null hypothesis.

Unit-4 - t-test, chi-square and ANOVA(f-test)

T-test:

Significant difference bw the means of two grps and their relation. Determines if two samples come from the same pop.

NULL : Two means are same

ALT: There exist diff.

Large t-score → a lot of diff. else small diff(means grps are similar)

PAIRED t-test

DOF=n-1

$$\boxed{T = \frac{mean1 - mean2}{\frac{s(\text{diff})}{\sqrt{(n)}}}}$$

POOLED T-TEST

DOF=n1+n2-2

$$t = \frac{(\overline{x_1} - \overline{x_2})}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$ where s is sqrt($\frac{(n1-1)\times var1^2 + (n2-1)\times var2^2}{n1+n2-2}$)

Welch’s T-test(if variance is diff)

$$\text{T-value} = \frac{mean1 - mean2}{\sqrt{\left(\frac{var1}{n1} + \frac{var2}{n2}\right)}}$$

$$\text{Degrees of Freedom} = \frac{\left(\frac{var1^2}{n1} + \frac{var2^2}{n2}\right)^2}{\frac{\left(\frac{var1^2}{n1}\right)^2}{n1-1} + \frac{\left(\frac{var2^2}{n2}\right)^2}{n2-1}}$$

In cases of one sample

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

Using differences

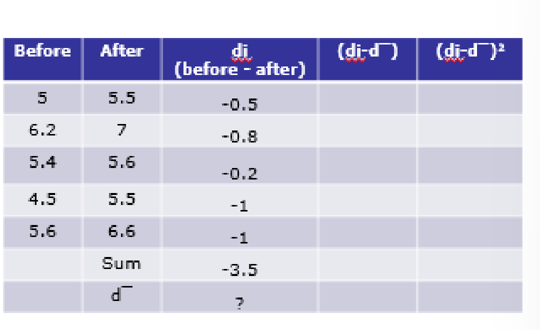
$$t = \frac{\bar{d} - \mu}{s_d / \sqrt{n}} \text{ with } \quad (n-1) \quad d.f$$

$$di = x_i(before) - x_i(after)$$

$$\bar{d} = \frac{\sum d_i}{n}$$

$$\mu = \mu_1 - \mu_2$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$



Before	After	di (before - after)	(di-d¯)	(di-d¯)²
5	5.5	-0.5		
6.2	7	-0.8		
5.4	5.6	-0.2		
4.5	5.5	-1		
5.6	6.6	-1		
	Sum	-3.5		
	d¯	?		

Imp Observation: More spread out distribution and has heavier tails(means greatness or least) to the std.normal dist.

Chi-square: Non-parametric test

Always ≥0

Positively right-skewed curve

Diff for each DOF

used for categorical ind. and dep.variables

TYPES: Single S.D/Variance , Independence, Goodness of fit(relevant from regression p.o.v)

Single S.D/Variance= S^2/S.D^2 at DOF
DOF=n-1

$$S^2 = [\Sigma(Xi- \overline{X})^2 /(n-1)]$$

Independence of variables:

Test statistics:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

F-test

Ratio b/w variances = (Var1)^2/(Var2)^2

- Variances measure the dispersal of the data points around the mean.
- Higher variances occur when the individual data points tend to fall further from the mean.
- F-Test is utilized to determine whether two populations' variances are equal.

Two-tailed F-test: Variances of two samples are equal or not

One-tailed F-test: One pop.variance is greater or less than other.

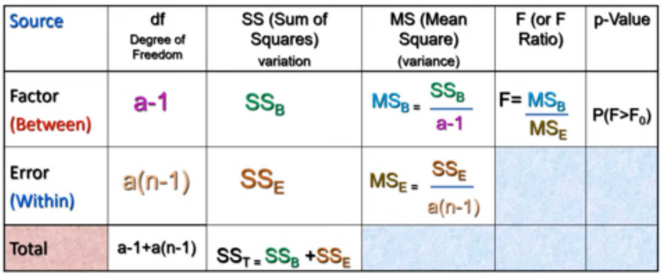
If test statistic>critical value , null hypo. is rejected

- **Normality**: the populations must have a normal distribution.
- **Independent and random selection of sample items**: the selection of the samples' components should be independent event and random.

How is it useful in Feature Engn?

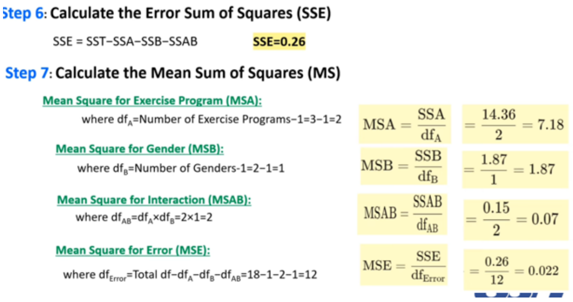
F-ratio is high, means it is a significant feature

One-way ANOVA



Source	df Degree of Freedom	SS (Sum of Squares) variation	MS (Mean Square) (variance)	F (or F Ratio)	p-Value
Factor (Between)	a-1	SS_B	$MS_B = \frac{SS_B}{a-1}$	$F = \frac{MS_B}{MS_E}$	$P(F>F_0)$
Error (Within)	a(n-1)	SS_E	$MS_E = \frac{SS_E}{a(n-1)}$		
Total	a-1+a(n-1)	$SS_T = SS_B + SS_E$			

Two-way ANOVA



Step 6: Calculate the Error Sum of Squares (SSE)

SSE = SST−SSA−SSB−SSAB SSE=0.26

Step 7: Calculate the Mean Sum of Squares (MS)

Mean Square for Exercise Program (MSA):

where df_A=Number of Exercise Programs−1=3−1=2

$$MSA = \frac{SSA}{df_A} = \frac{14.36}{2} = 7.18$$

Mean Square for Gender (MSB):

where df_B=Number of Genders-1=2−1=1

$$MSB = \frac{SSB}{df_B} = \frac{1.87}{1} = 1.87$$

Mean Square for Interaction (MSAB):

where $df_{AB}=df_A \times df_B=2\times1=2$

$$MSAB = \frac{SSAB}{df_{AB}} = \frac{0.15}{2} = 0.07$$

Mean Square for Error (MSE):

where df_{Error}=Total df−df_A−df_B−df_{AB}=18−1−2−1=12

$$MSE = \frac{SSE}{df_{Error}} = \frac{0.26}{12} = 0.022$$

So, we have two factors, Total and Their interaction in play here

Unit -5 Time series - Regression and Survival Analysis

Time-series analysis:

Consider time as the inde.variable

Historical datasets ku this is useful

Compare current trends with prev. trends

Mainly used for forecasting
Components:

- **Trend**: In which there is no fixed interval and any divergence within the given dataset is a continuous timeline. The trend would be Negative or Positive or Null Trend
- **Seasonality**: In which regular or fixed interval shifts within the dataset in a continuous timeline. Would be bell curve or saw tooth
- **Cyclical**: In which there is no fixed interval, uncertainty in movement and its pattern

- **Irregularity (Noise)**: Unexpected situations/events/scenarios and spikes in a short time span.
- **Time series data** is data that is recorded over consistent intervals of time.
- **Cross-sectional data** consists of several variables recorded at the same time.
- **Pooled data** is a combination of both time series data and cross-sectional data.

Survival analysis is a branch of statistics that focuses on analyzing and interpreting the time until an event of interest occurs. This event can vary depending on the context, such as the time until a patient dies, a machine fails, or a customer churns(or a radioactive decay, just got reminded by JEE physics lol). It is widely used in fields like medicine, engineering, and business.

Key Concepts in Survival Analysis:

1. **Survival Time**:
   - The duration from a defined starting point (e.g., diagnosis of a disease) to the occurrence of the event (e.g., death or recovery).
2. **Censoring**:
   - Not all observations experience the event during the study period. These incomplete observations are called *censored data*. For example, a patient might withdraw from the study, or the study ends before the event occurs.
3. **Survival Function (S(t))**:
   - Represents the probability that the event has not occurred by time . Mathematically:

     $S(t)=P(T>t)$

     T is the survival time
4. **Hazard Function (h(t))**:
   - The instantaneous rate at which the event occurs at time , given survival until t
5. **Kaplan-Meier Estimator**:

   - A non-parametric method to estimate the survival function from censored data.

6. **Cox Proportional Hazards Model**:
   - A semi-parametric model used to examine the effect of covariates (e.g., age, treatment type) on the hazard rate.
7. **Exponential, Weibull, and Log-normal Models (this is the one used for skewed distributions)**
   - Parametric models that assume specific distributions for survival times.

Applications of Survival Analysis:

1. **Medical Research**: To study patient survival rates and the efficacy of treatments.
2. **Engineering**: For reliability analysis and predicting the failure time of machines or systems.
3. **Business and Marketing**: To analyze customer retention, churn, and the lifetime of products or subscriptions.
4. **Social Sciences**: To study events like marriage duration or employment tenure.

Example:

If you're studying the survival times of cancer patients undergoing different treatments, survival analysis helps:

- Estimate the median survival time.
- Compare survival curves between treatment groups.
- Identify factors that significantly influence survival.

Survival analysis is powerful because it can handle censored data and provides insights into both the probability and the risk of events over time.