What is Statistics

09 March 2023 14:56

Statistics is a branch of mathematics that involves collecting, analysing, interpreting, and presenting data. It provides tools and methods to understand and make sense of large amounts of data and to draw conclusions and make decisions based on the data.

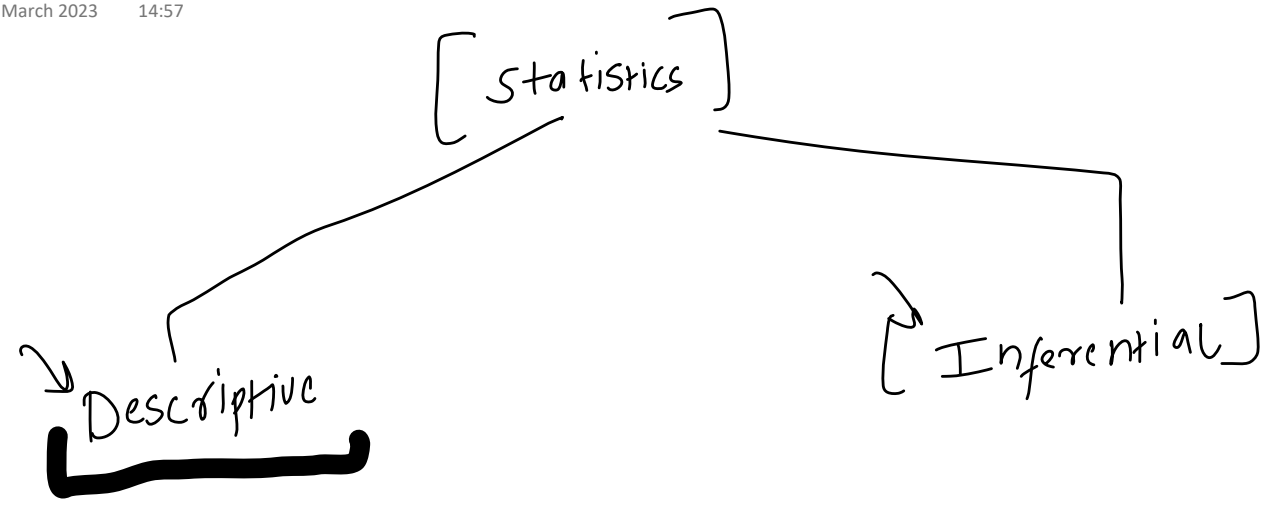
In practice, statistics is used in a wide range of fields, such as business, economics, social sciences, medicine, and engineering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions, and make forecasts and predictions.

Examples:

1. Business - Data Analysis(Identifying customer behavior) and Demand Forecasting
2. Medical - Identify efficacy of new medicines(Clinical trials), Identifying risk factor for diseases(Epidemiology)
3. Government & Politics - Conducting surveys, Polling
4. Environmental Science - Climate research

Types of Statistics

09 March 2023 14:57



Descriptive statistics deals with the collection, organization, analysis, interpretation, and presentation of data. It focuses on summarizing and describing the main features of a set of data, without making inferences or predictions about the larger population.

Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables

Population Vs Sample

09 March 2023 14:57

Population refers to the entire group of individuals or objects that we are interested in studying. It is the complete set of observations that we want to make inferences about. For example, the population might be all the students in a particular school or all the cars in a particular city.

A **sample**, on the other hand, is a subset of the population. It is a smaller group of individuals or objects that we select from the population to study. Samples are used to estimate characteristics of the population, such as the mean or the proportion with a certain attribute. For example, we might randomly select 100 students.

Examples

1. All cricket fans vs fans who were present in the stadium
2. All students vs who visit college for lectures

Things to be careful about which creating samples

1. Sample Size
2. Random
3. Representative

Parameter Vs Statistics

A parameter is a characteristic of a population, while a statistic is a characteristic of a sample. Parameters are generally unknown and are estimated using statistics. The goal of statistical inference is to use the information obtained from the sample to make inferences about the population parameters.

Inferential Statistics

09 March 2023 14:57

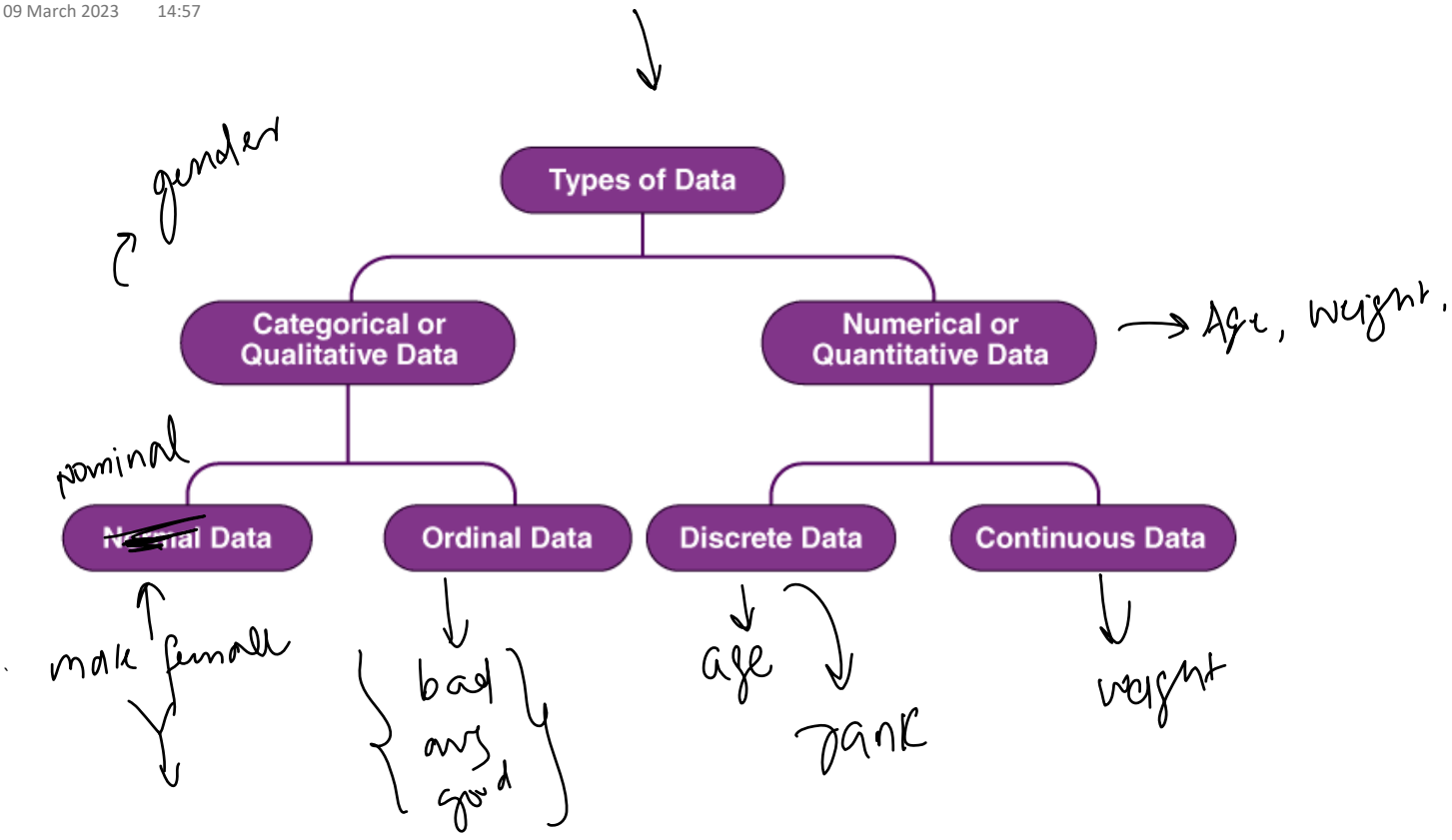
Inferential statistics is a branch of statistics that deals with making inferences or predictions about a larger population based on a sample of data. It involves using statistical techniques to test hypotheses and draw conclusions from data. Some of the topics that come under inferential statistics are:

1. **Hypothesis testing**: This involves testing a hypothesis about a population parameter based on a sample of data. For example, testing whether the mean height of a population is different from a given value.
2. **Confidence intervals**: This involves estimating the range of values that a population parameter could take based on a sample of data. For example, estimating the population mean height within a given confidence level.
3. **Analysis of variance (ANOVA)**: This involves comparing means across multiple groups to determine if there are any significant differences. For example, comparing the mean height of individuals from different regions.
4. **Regression analysis**: This involves modelling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a product based on advertising expenditure.
5. **Chi-square tests**: This involves testing the independence or association between two categorical variables. For example, testing whether gender and occupation are independent variables.
6. **Sampling techniques**: This involves ensuring that the sample of data is representative of the population. For example, using random sampling to select individuals from a population.
7. **Bayesian statistics**: This is an alternative approach to statistical inference that involves updating beliefs about the probability of an event based on new evidence. For example, updating the probability of a disease given a positive test result.

Why ML is closely associated with statistics?

Types of Data

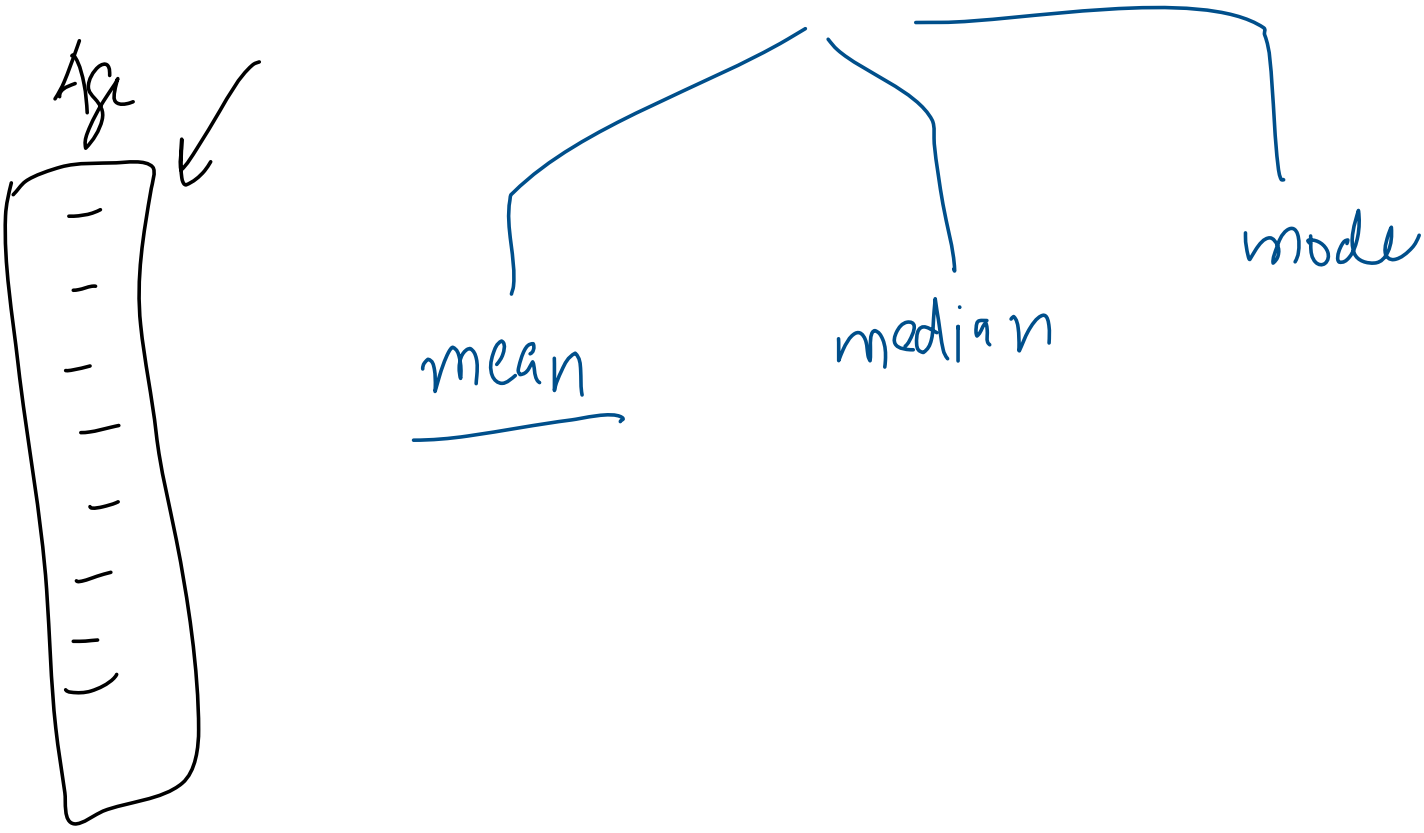
09 March 2023 14:57



Measure of Central Tendency

09 March 2023 14:58

A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.



1. Mean

09 March 2023 16:35

Mean: The mean is the sum of all values in the dataset divided by the number of values.

3
4
1
2
5

$= \frac{15}{5} = 3$

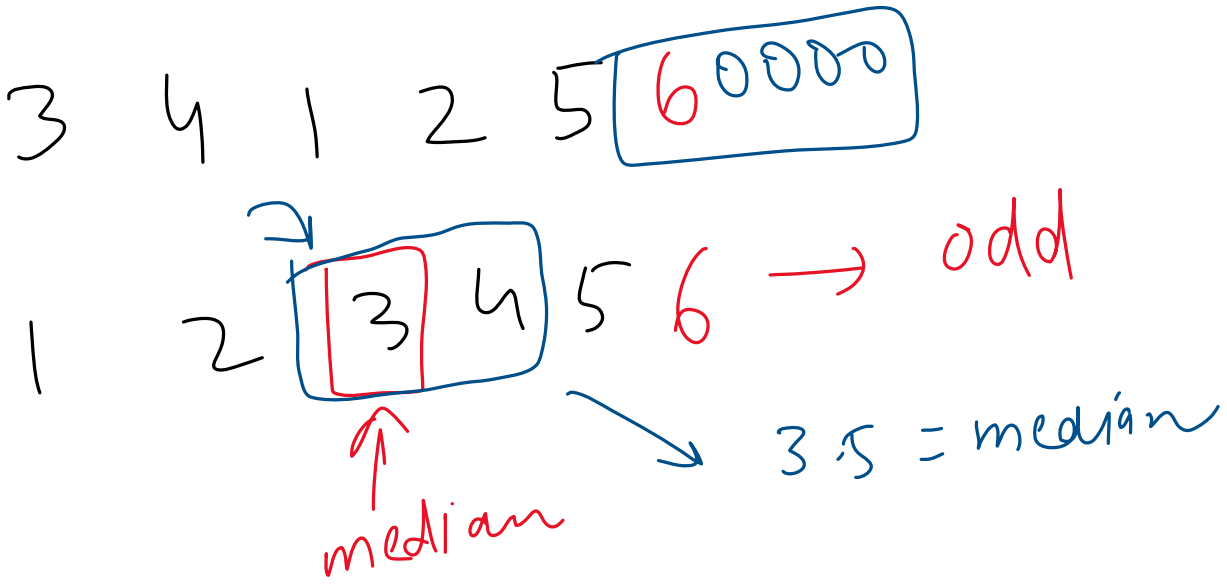
$\sum_{i=1}^{N} x_i$ / N $\qquad$ $\sum_{i=1}^{n} x_i$ / n

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

2. Median

09 March 2023 16:36

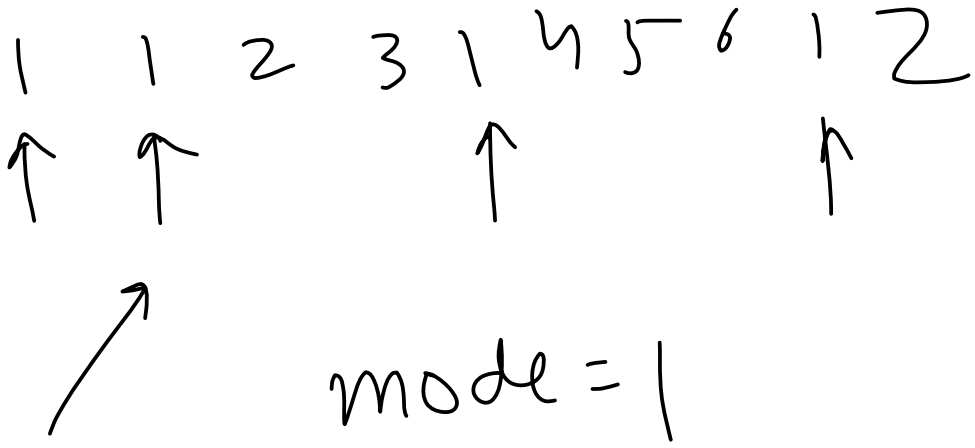
Median: The median is the middle value in the dataset when the data is arranged in order.



3. Mode

09 March 2023 16:36

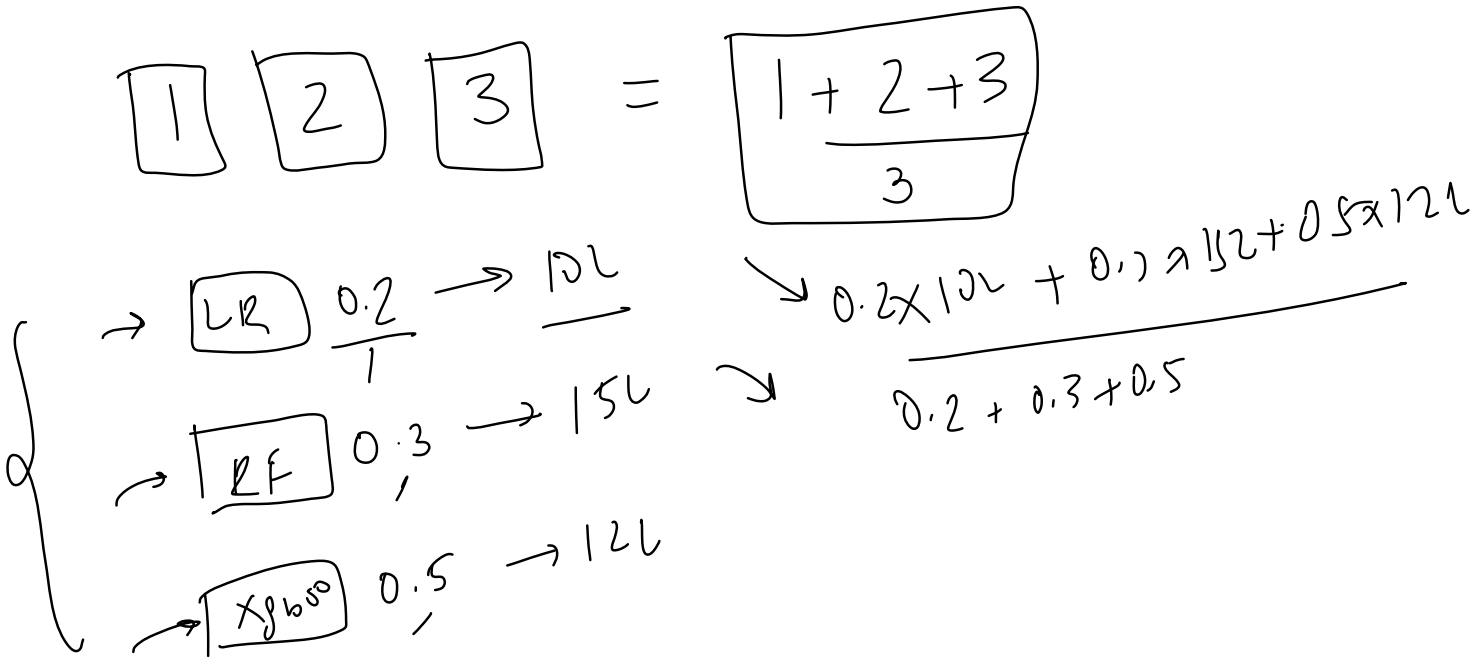
Mode: The mode is the value that appears most frequently in the dataset.



4. Weighted Mean

09 March 2023 16:39

Weighted Mean: The weighted mean is the sum of the products of each value and its weight, divided by the sum of the weights. It is used to calculate a mean when the values in the dataset have different importance or frequency.



5. Trimmed Mean

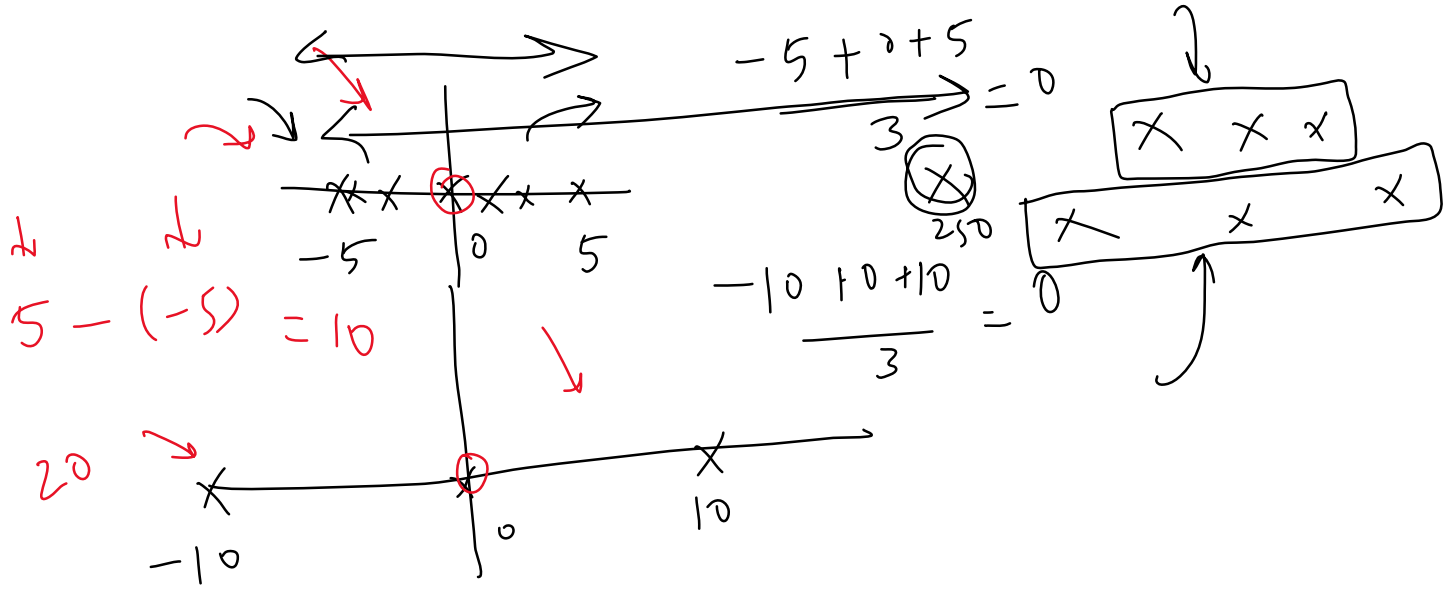
10 March 2023 09:37

A trimmed mean is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called the trimming percentage.

Measure of Dispersion

09 March 2023 14:58

A measure of dispersion is a statistical measure that describes the spread or variability of a dataset. It provides information about how the data is distributed around the central tendency (mean, median or mode) of the dataset.



1. Range

09 March 2023 16:36

Range: The range is the difference between the maximum and minimum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.

2. Variance

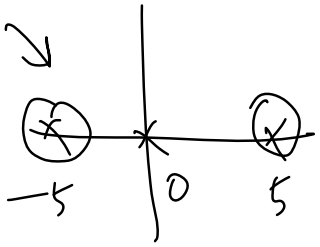
09 March 2023 16:36

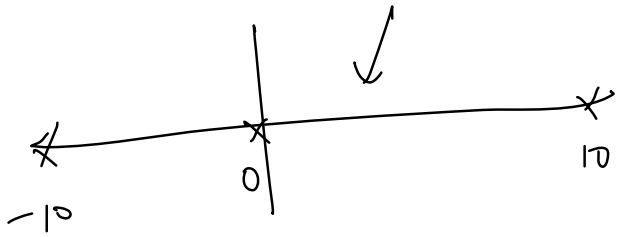
Variance: The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

	X-mean	(X-mean)^2
3	3-3	0
2	2-3	1
1	1-3	4
5	5-3	4
4	4-3	1

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$ Population Variance

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$ Sample Variance





Mean Absolute Deviation

$$MAD = \frac{\sum |x_i - \bar{x}|}{n}$$

informal

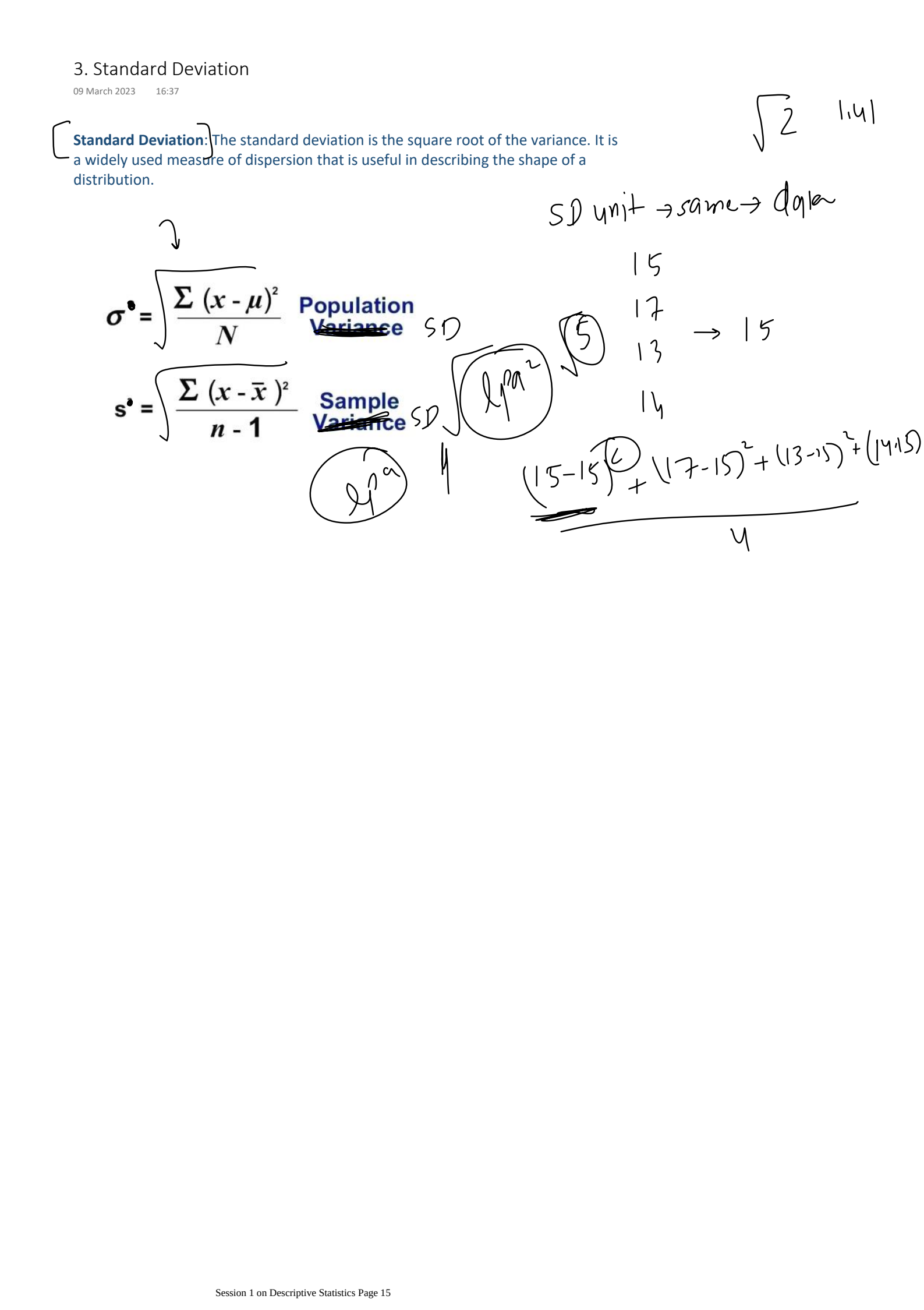
3. Standard Deviation

09 March 2023 16:37

Standard Deviation: The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$ Population Variance

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$ Sample Variance



4. Coefficient of Variation

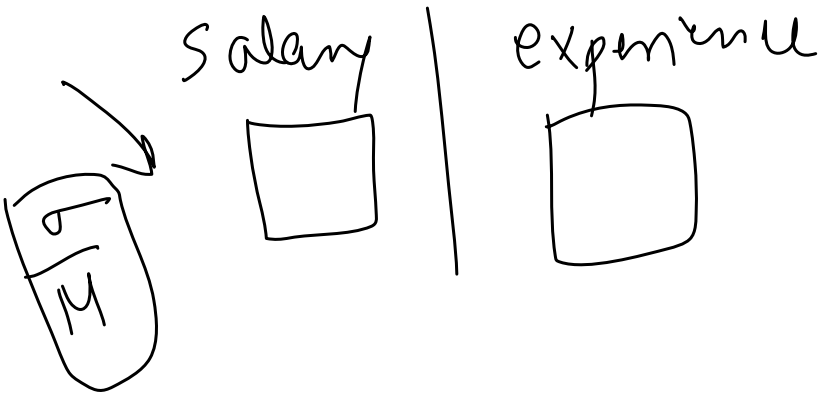
09 March 2023 16:37

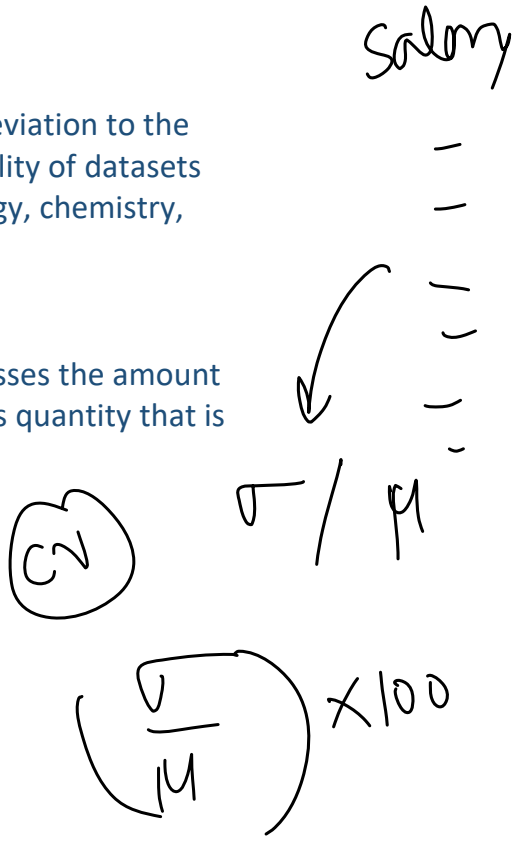
Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in fields such as biology, chemistry, and engineering.

The coefficient of variation (CV) is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is a dimensionless quantity that is expressed as a percentage.

The formula for calculating the coefficient of variation is:

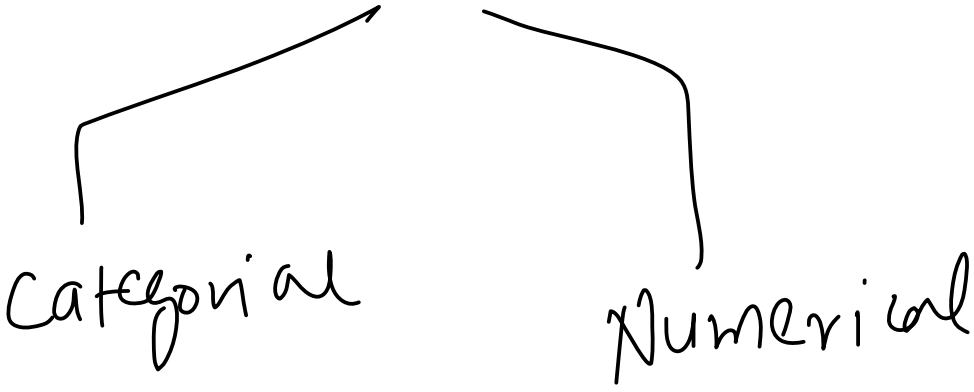
CV = (standard deviation / mean) x 100%





Graphs for Univariate Analysis

09 March 2023 14:58



1. Categorical - Frequency Distribution Table & Cumulative Frequency

09 March 2023 16:50

A **frequency distribution table** is a table that summarizes the number of times (or frequency) that each value occurs in a dataset.

Let's say we have a survey of 200 people and we ask them about their favourite type of vacation, which could be one of six categories: Beach, City, Adventure, Nature, Cruise, or Other

Type of Vacation	Frequency
Beach	60
City	40
Adventure	30
Nature	35
Cruise	20
Other	15

Relative frequency is the proportion or percentage of a category in a dataset or sample. It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample.

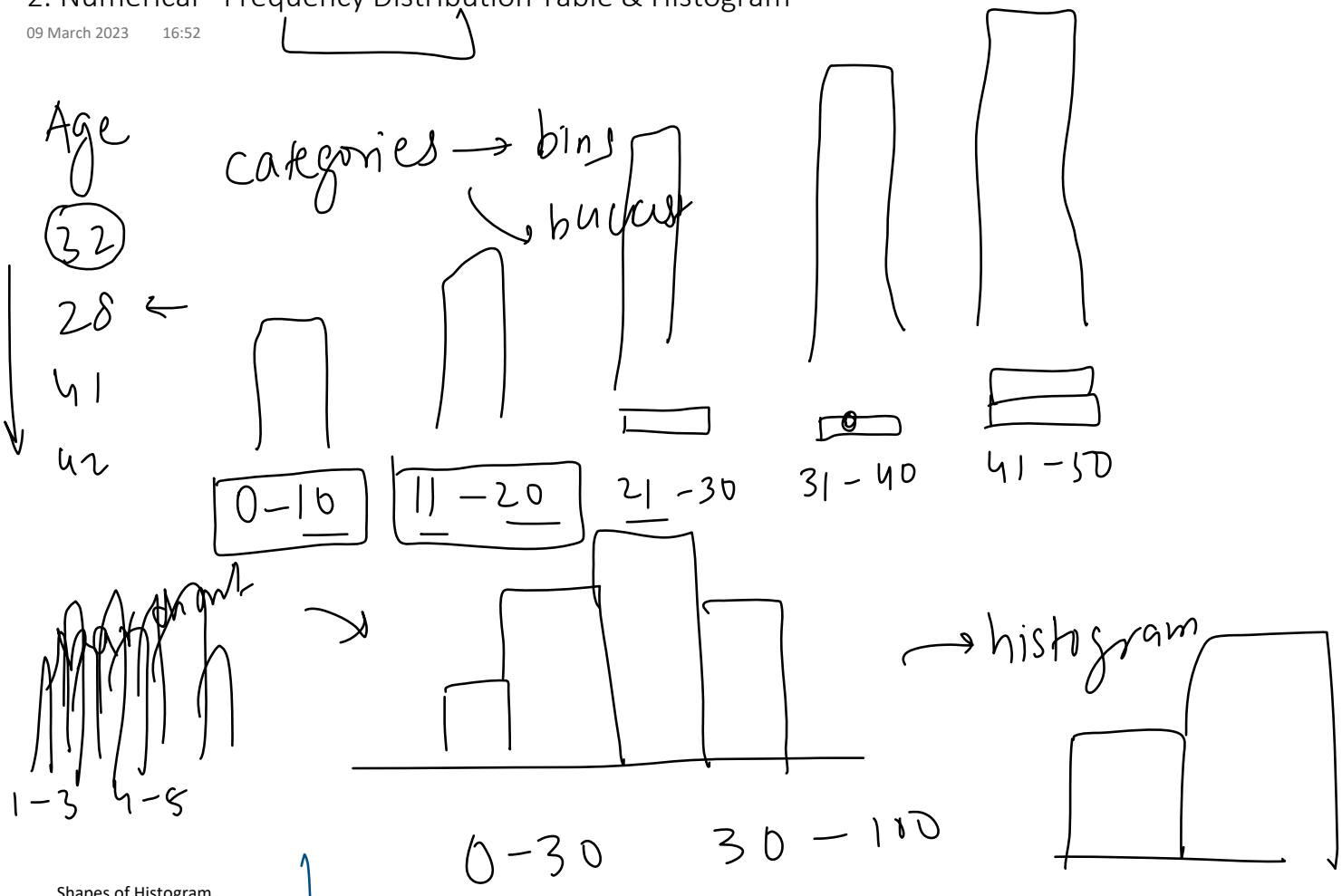
Type of Vacation	Frequency	Relative Frequency
Beach	60	0.3
City	40	0.2
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
Other	15	0.075

Cumulative frequency is the running total of frequencies of a variable or category in a dataset or sample. It is calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample.

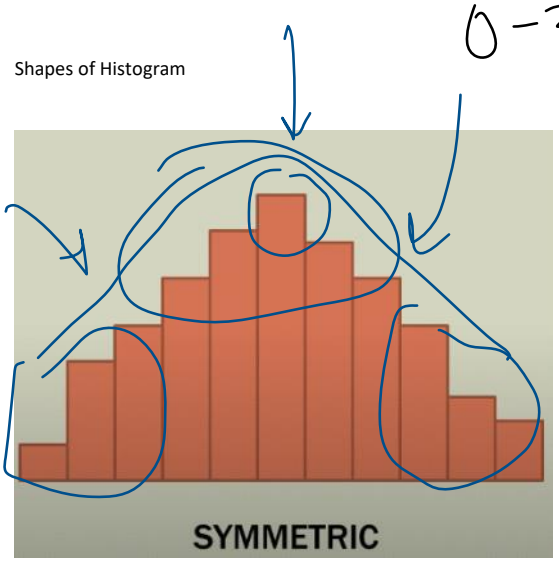
Type of Vacation	Frequency	Relative Frequency	Cumulative Frequency
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200

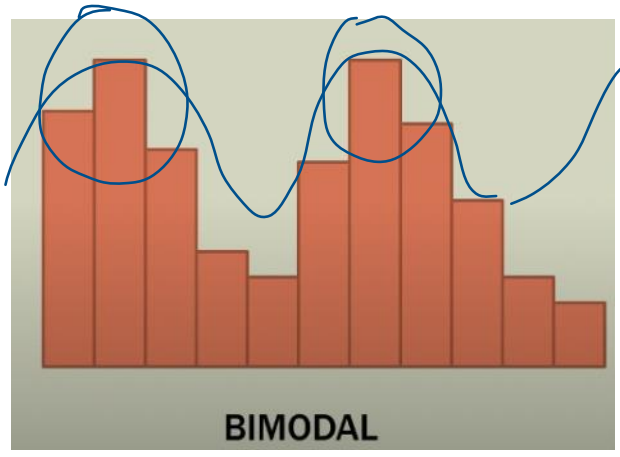
2. Numerical - Frequency Distribution Table & Histogram

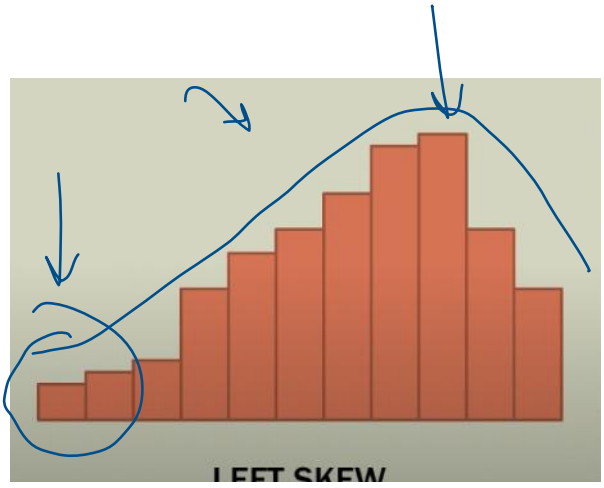
09 March 2023 16:52

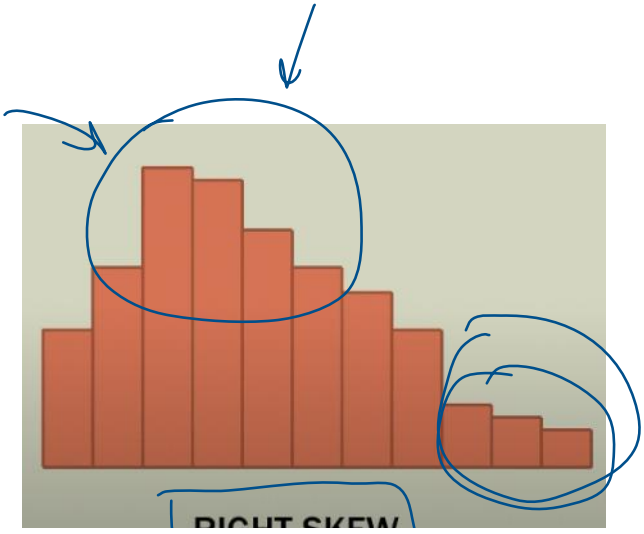


Shapes of Histogram



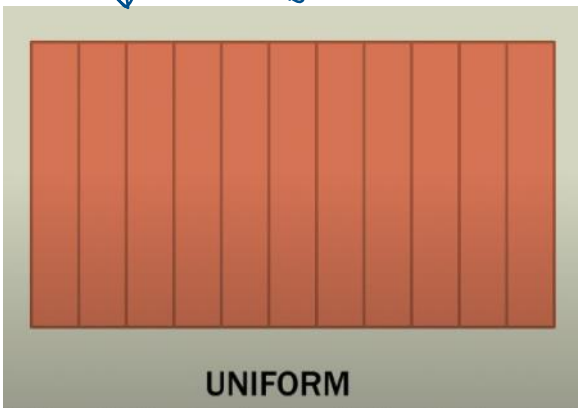


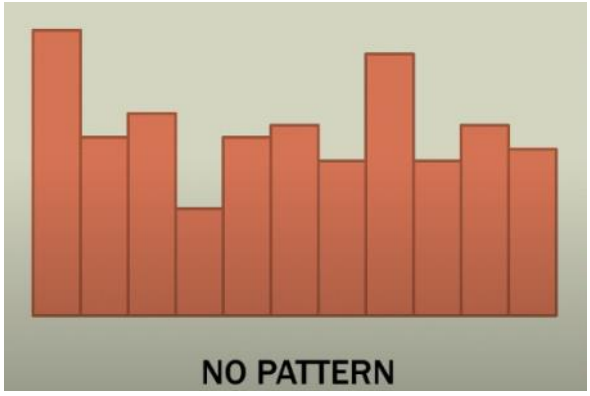





LEFT SKEW

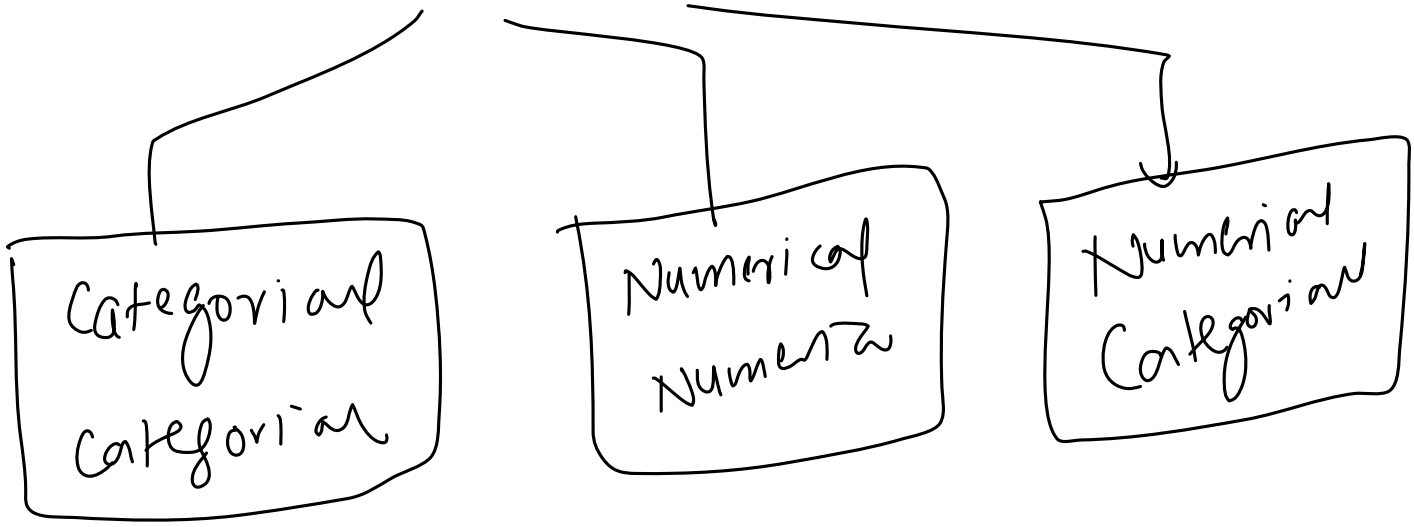

RIGHT SKEW


UNIFORM


NO PATTERN

Graphs for Bivariate Analysis

09 March 2023 14:59

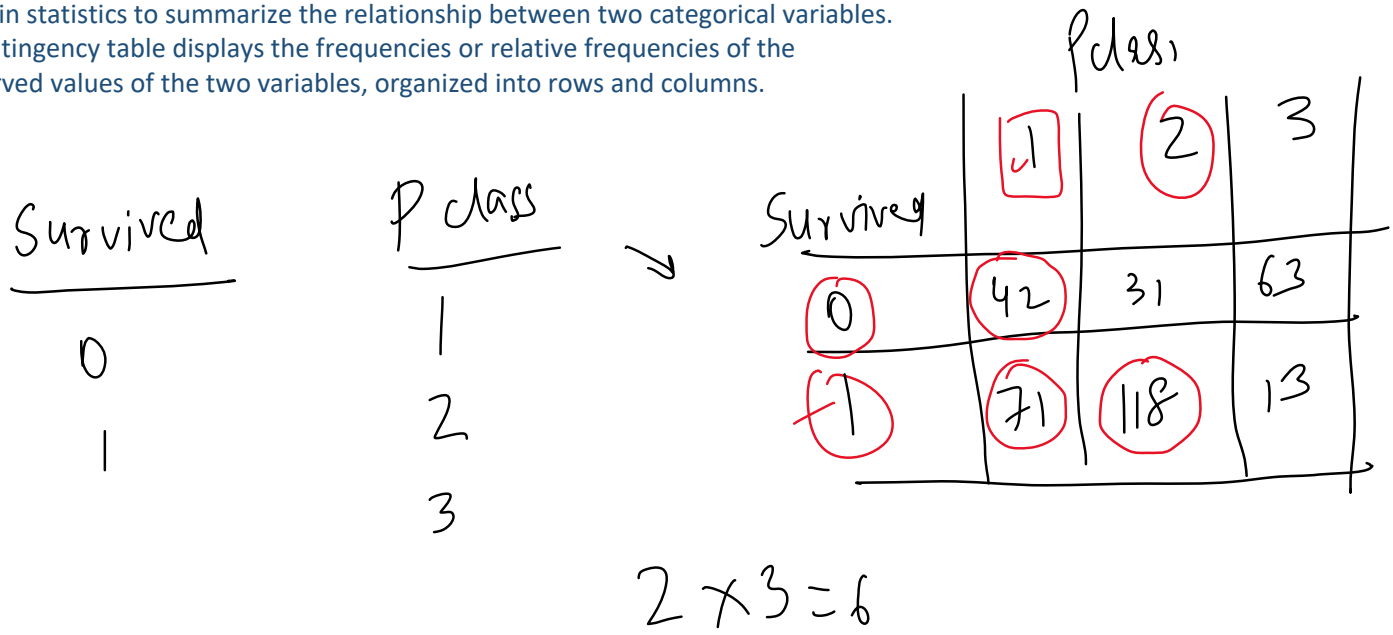


1. Categorical - Categorical

09 March 2023 16:58

Contingency Table/Crosstab

A contingency table, also known as a cross-tabulation or crosstab, is a type of table used in statistics to summarize the relationship between two categorical variables. A contingency table displays the frequencies or relative frequencies of the observed values of the two variables, organized into rows and columns.

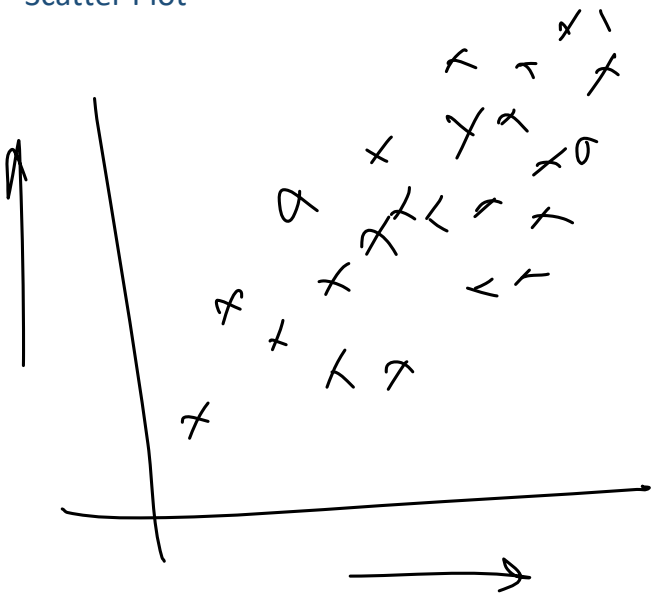


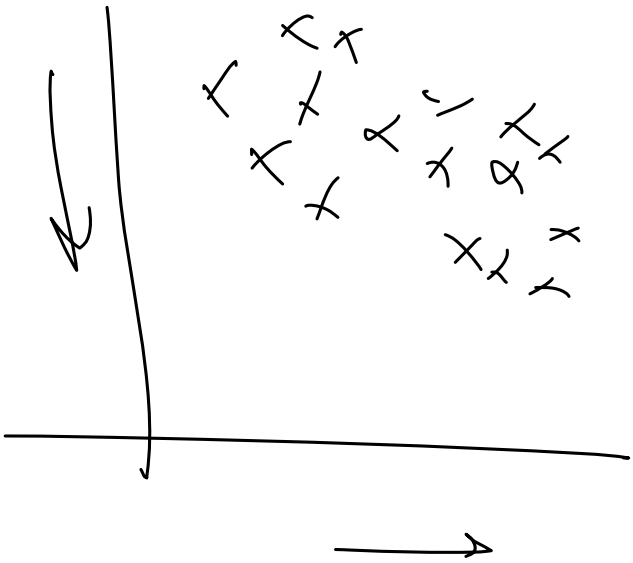
$2 \times 3 = 6$

2. Numerical - Numerical

09 March 2023 16:58

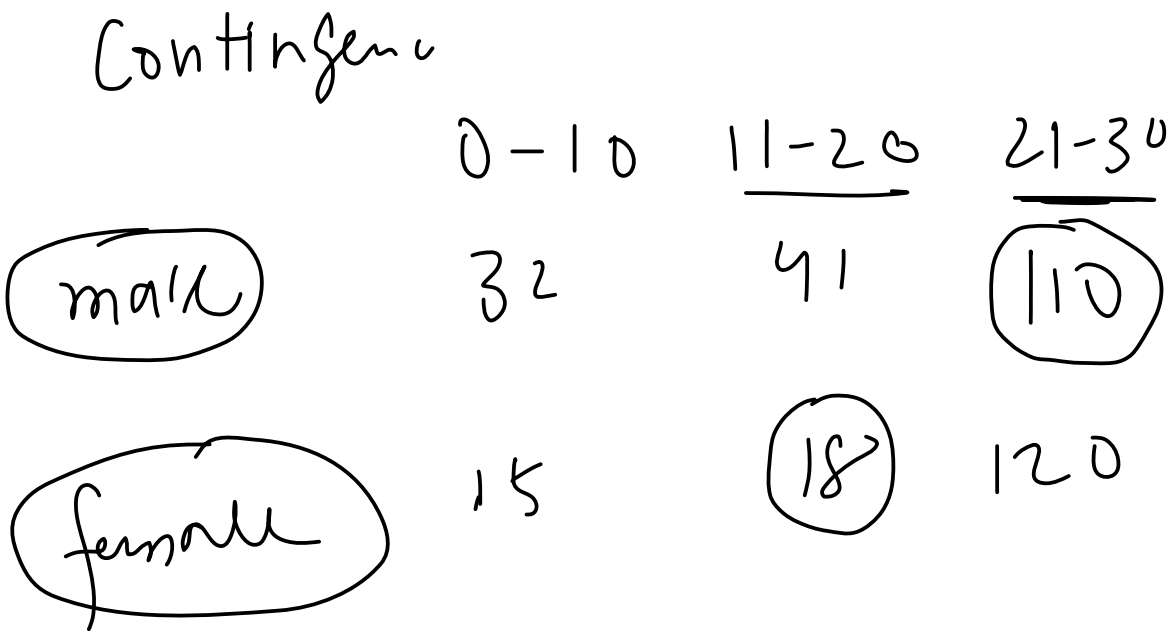
Scatter Plot





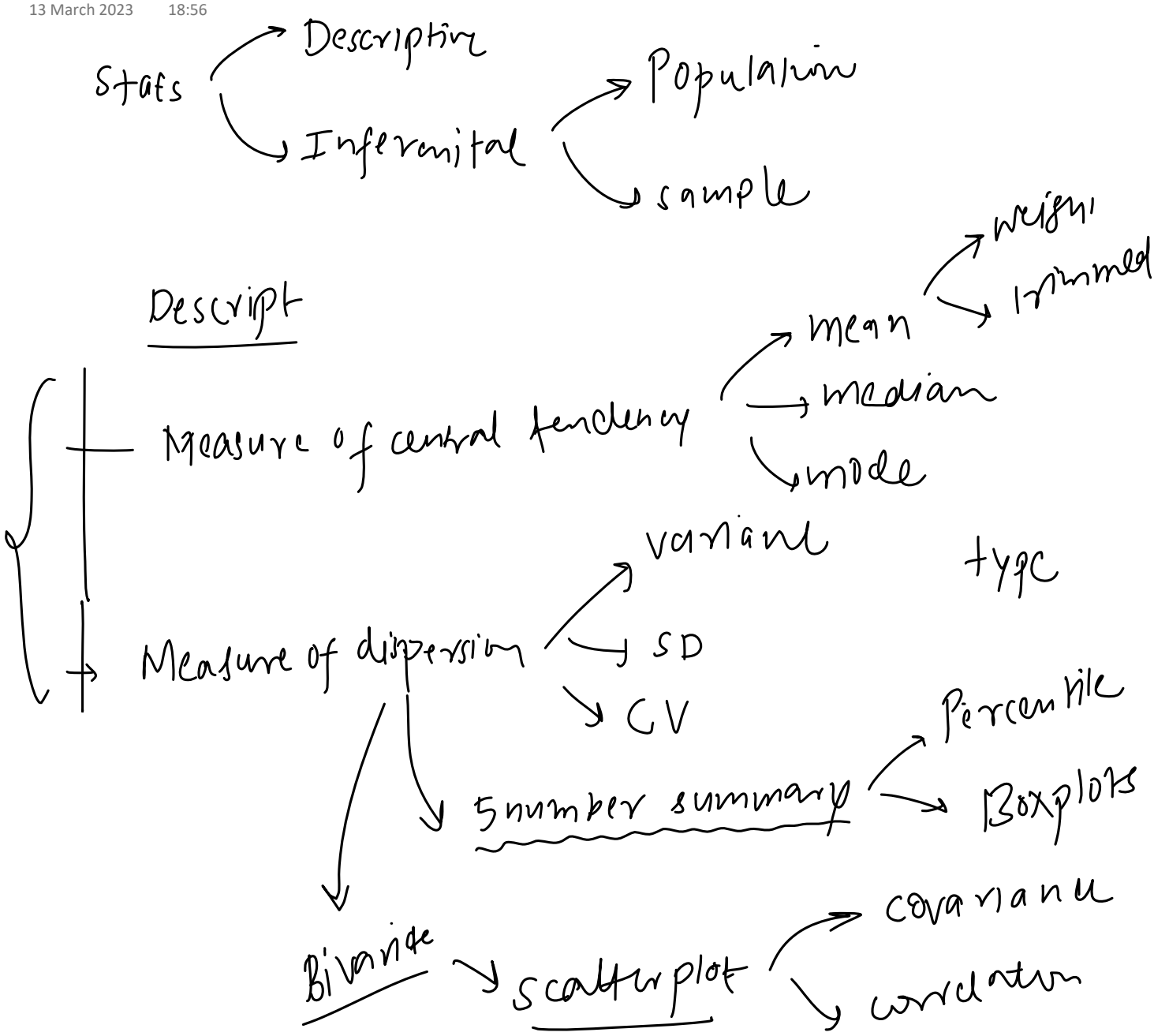
3. Categorical - Numerical

09 March 2023 16:58



Recap

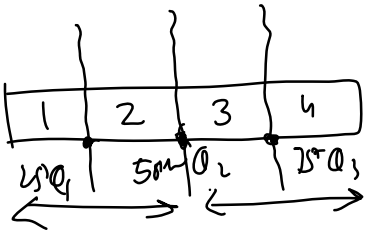
13 March 2023 18:56



Quantiles and Percentiles

13 March 2023 06:57

Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.



Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles used in statistical analysis, including:

a. Quartiles: Divide the data into four equal parts, Q1 (25th percentile), Q2 (50th percentile or median), and Q3 (75th percentile).

b. Deciles: Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile), ..., D9 (90th percentile).

c. Percentiles: Divide the data into 100 equal parts, P1 (1st percentile), P2 (2nd percentile), ..., P99 (99th percentile).

d. Quintiles: Divides the data into 5 equal parts

Things to remember while calculating these measures:

1. Data should be sorted from low to high
2. You are basically finding the location of an observation
3. They are not actual values in the data
4. All other tiles can be easily derived from Percentiles

Percentile

A percentile is a statistical measure that represents the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

Formula to calculate the percentile value:

$$PL = \frac{p}{100}(N+1)$$

where:

- PL = the desired percentile value location
- N = the total number of observations in the dataset
- p = the percentile rank (expressed as a percentage)

Example:

Find the 75th percentile score from the below data

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Step1 - Sort the data (Asc)

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

1 2 3 4 5 6 7 8 9 10

$$PL = \frac{75}{100}(10+1) = \frac{3}{4} \times 11 = \frac{33}{4} = 8.25$$

96 —— 98 Q75 91 —— 93

100

96 —— 98 (0.75) 91 —— 93

8 —— 9 5 6

↑

$96 + 0.25(98 - 96) = 96 + 0.25 \times 2 = 96.5$

↘

75th percentile = 96.5

$91 + 0.5(93 - 91)$

$91 + 0.5 \times 2 = 92$

$P_L = \frac{50}{100}(10+1) = \frac{1}{2} \times 11 = 5.5$

Percentile of a value

Percentile rank = $\frac{x + 0.5y}{n}$ $\left(\frac{n}{N}\right)$

X = number of values below the given value

Y = number of values equal to the given value

N = total number of values in the dataset

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

1 2 3

$= \frac{3 + 0.5 \times 1}{10} = \frac{3.5}{10}$

0.35 → 35th

$\frac{9 + 0.5 \times 1}{10} = \frac{9.5}{10} = 0.95$

$\frac{9}{10}$ 0.9

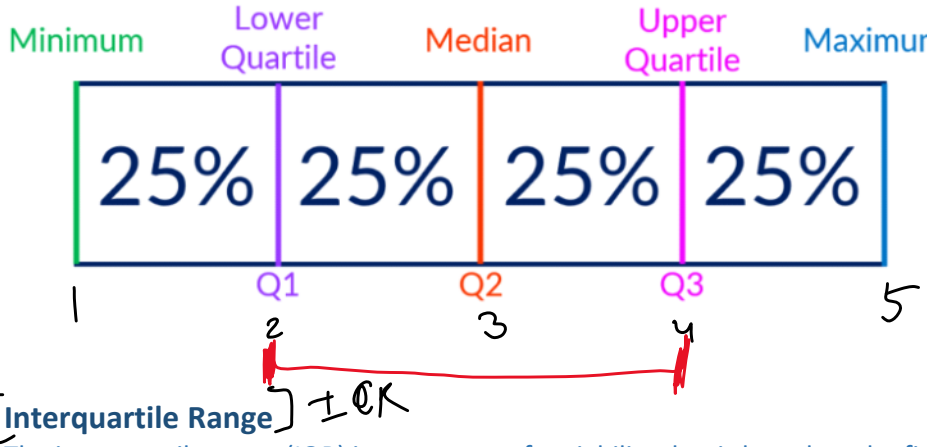
5 number summary

13 March 2023 06:57

The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

1. **Minimum value**: The smallest value in the dataset.
2. **First quartile (Q1)**: The value that separates the lowest 25% of the data from the rest of the dataset.
3. **Median (Q2)**: The value that separates the lowest 50% from the highest 50% of the data.
4. **Third quartile (Q3)**: The value that separates the lowest 75% of the data from the highest 25% of the data.
5. **Maximum value**: The largest value in the dataset.

The five-number summary is often represented visually using a box plot, which displays the range of the dataset, the median, and the quartiles.
The five-number summary is a useful way to quickly summarize the central tendency, variability, and distribution of a dataset.



Interquartile Range

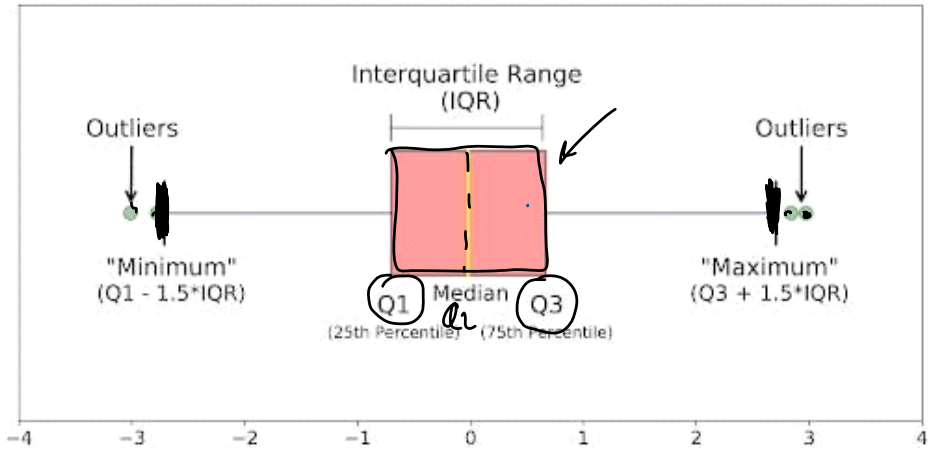
The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.

Boxplots

13 March 2023 06:57

1. What is a boxplot

A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that shows the distribution of the data. The box plot displays a summary of the data, including the minimum and maximum values, the first quartile (Q1), the median (Q2), and the third quartile (Q3).



2. How to create a boxplot with example

6 213 241 260 281 290 314 321 350 1500

1 2 3 4 5 6 7 8 9 10

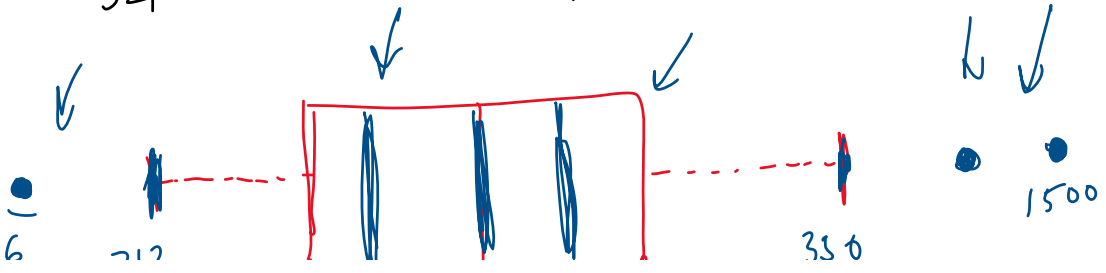
$Q_2 = \frac{50}{100}(11) = 5.5 = 285.5$

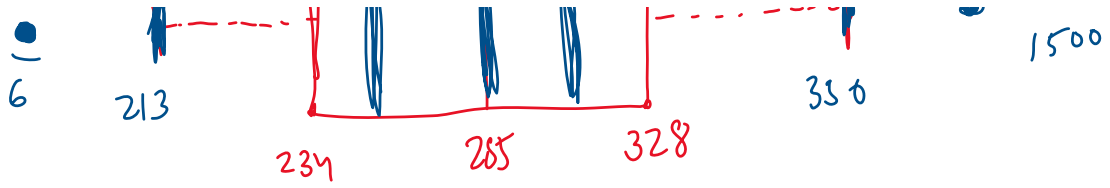
$Q_1 = \frac{25}{100} \times 11 = \frac{11}{4} = 2.75$

$213 + 0.75(241 - 213) = 234$

$Q_3 = \frac{75}{100} \times 11 = \frac{33}{4} = 8.25$

$321 + 0.25(350 - 321) = 328.25$





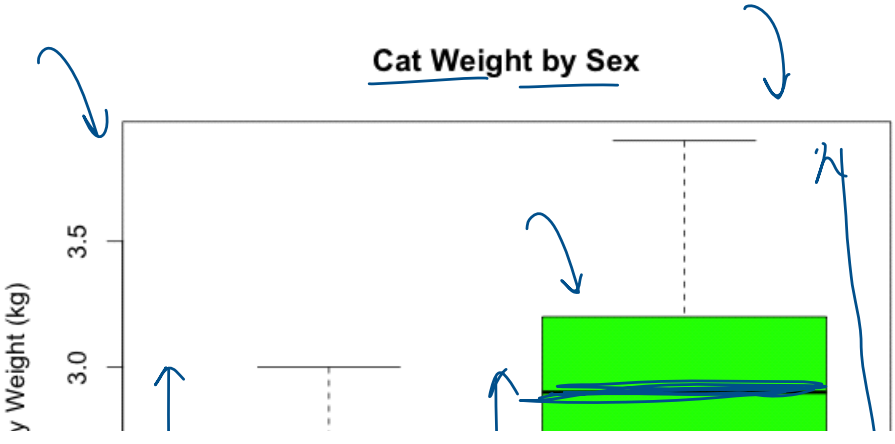
min and max

IQR = 94

$\min = Q_1 - 1.5(IQR) = 93$

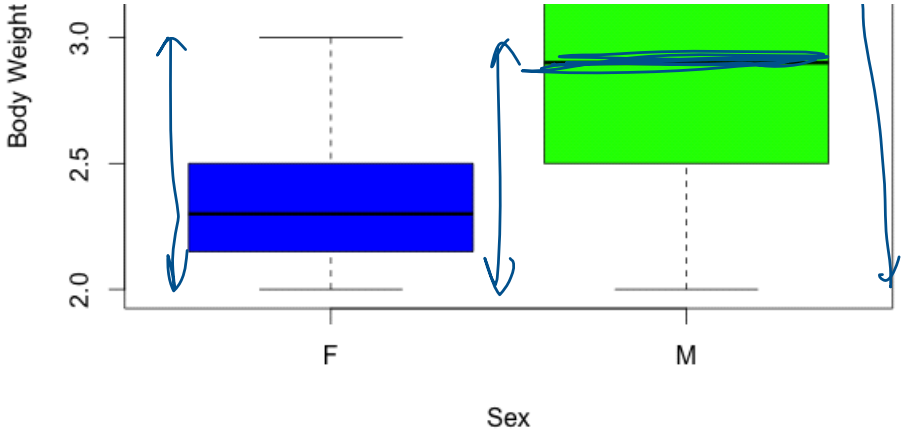
$\max = Q_3 + 1.5(IQR) = 469$

1. Benefits of a Boxplot
   - Easy way to see the distribution of data
   - Tells about skewness of data
   - Can identify outliers
   - Compare 2 categories of data
2. Side by side boxplot



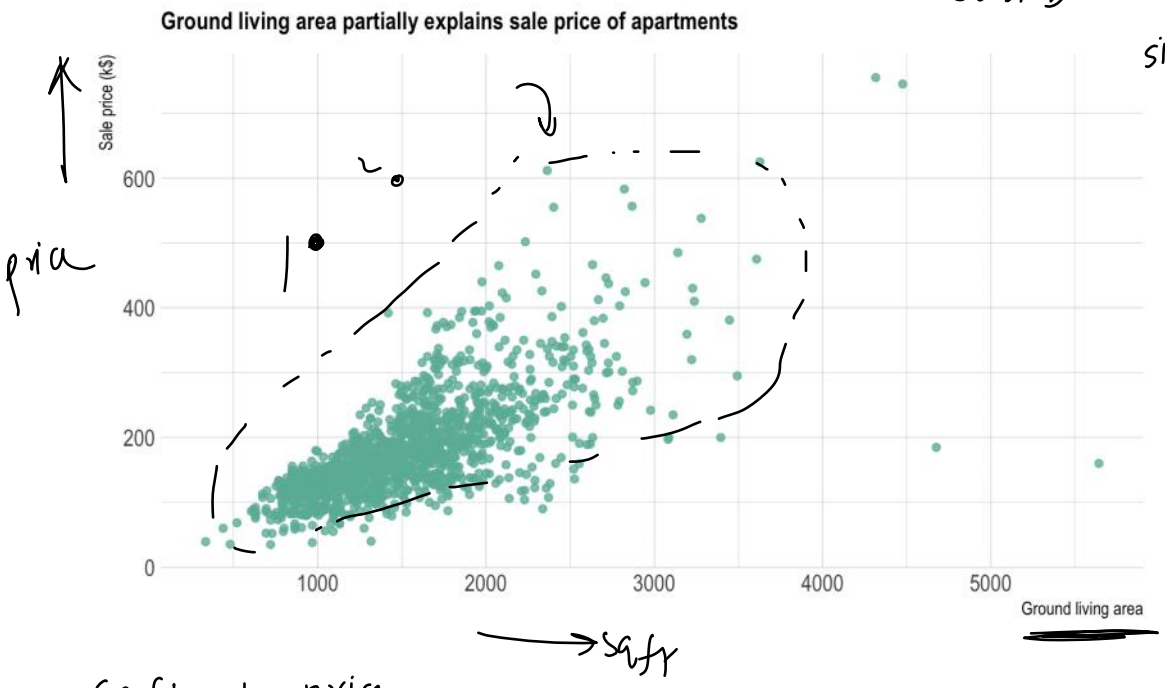
Weight

Age	gender
21	M
34	F
43	M


Body Weight
3.0
2.5
2.0
F
M
Sex

Scatterplots

13 March 2023 06:58

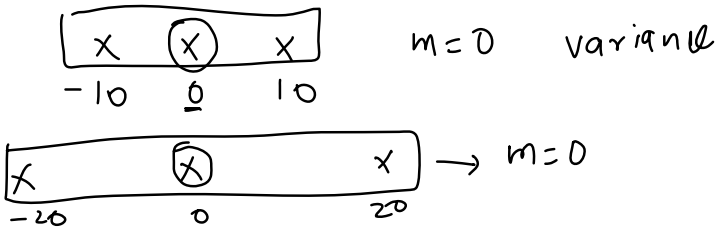


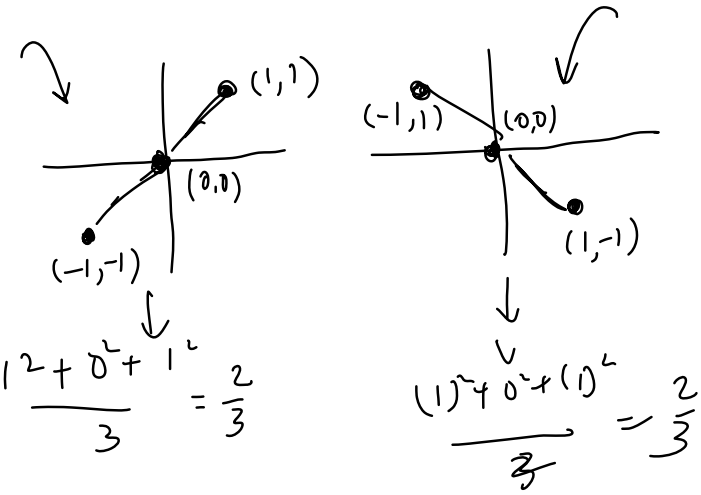
Sqft	price
1500	60L
1700	80L
1200	35L

Covariance

13 March 2023 06:57

- What problem does Covariance solve?

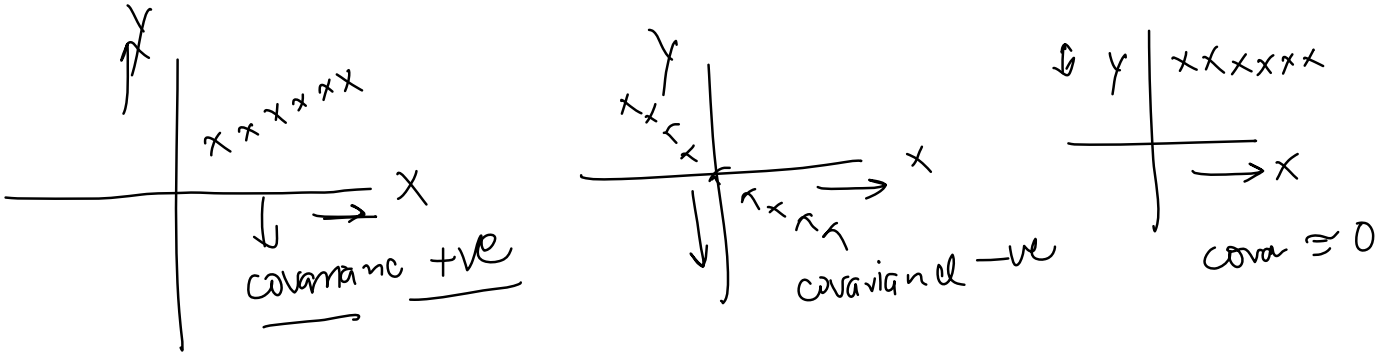




- What is covariance and how is it interpreted?

  Covariance is a statistical measure that describes the degree to which two variables are linearly related. It measures how much two variables change together, such that when one variable increases, does the other variable also increase, or does it decrease?

  If the covariance between two variables is positive, it means that the variables tend to move together in the same direction. If the covariance is negative, it means that the variables tend to move in opposite directions. A covariance of zero indicates that the variables are not linearly related.



- How is it calculated?

Covariance Formula	
Population	**Sample**
$\sigma_{xy} = \frac{\sum(X-\mu_x)(Y-\mu_y)}{N}$	$s_{xy} = \frac{\sum(X-\bar{x})(Y-\bar{y})}{n-1}$
X, Y – The Value of X and Y in the Population μ_x, μ_y *– The population Mean of X and Y* *N – Total Number of Observations*	*X, Y – The Value of X and Y in the Sample Data* $\bar{x}, \bar{y}$ *– The Sample Mean of X and Y* *n - Total Number of Observations*

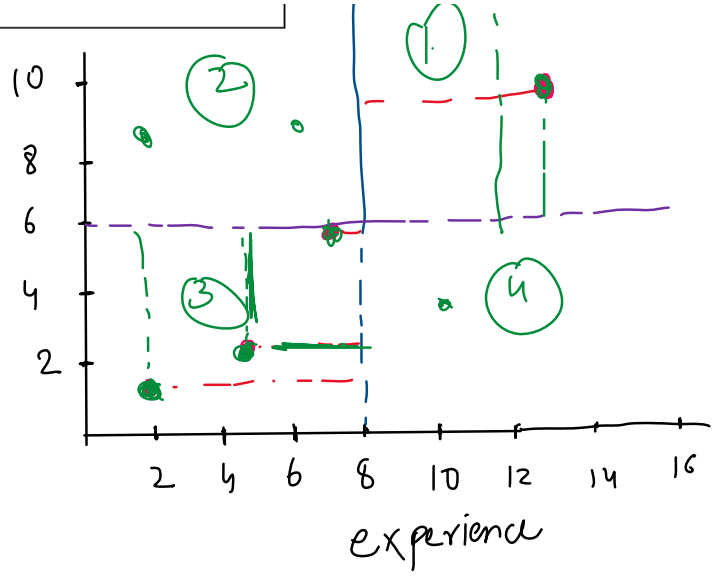
Exp(x)	Salary(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	-1
12	12	4	6	24
13	10	5	4	20

$\bar{X} = 8 \quad \bar{Y} = 6$

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

$n-1$

$\frac{85}{4} =$ COV = 21.5

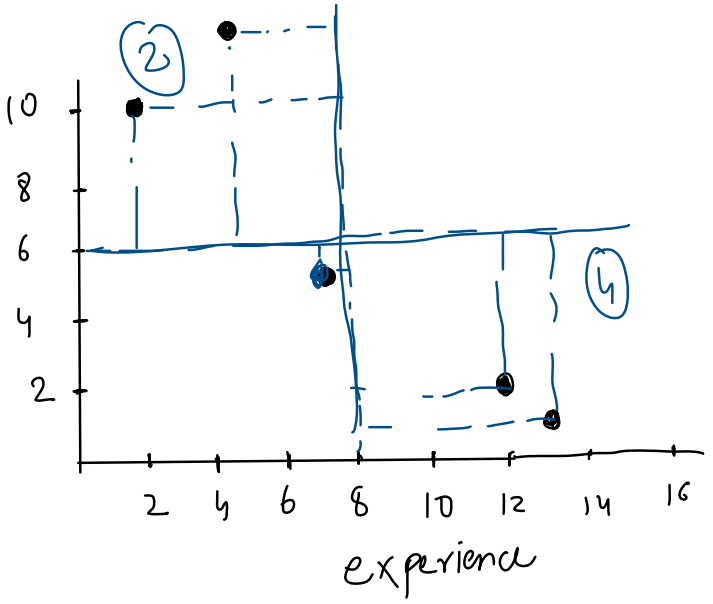


Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10	-6	4	-24
5	12	-3	6	-18
8	5	0	1	0
12	2	4	-4	-16
13	1	5	-5	-25

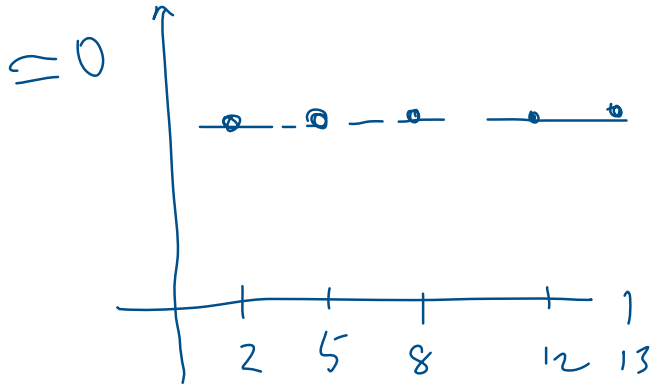
$\bar{X} = 8 \quad \bar{Y} = 6$

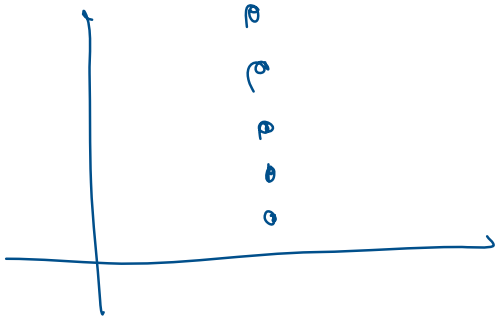
$\frac{-83}{4}$

-21.3



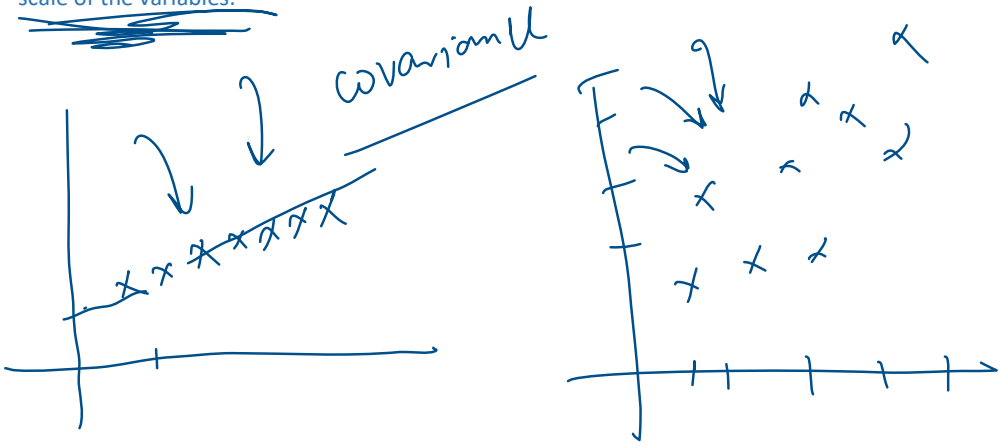
Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10			
5	10			
8	10			
12	10			
13	10			





- Disadvantages of using Covariance

  One limitation of covariance is that it does not tell us about the strength of the relationship between two variables, since the magnitude of covariance is affected by the scale of the variables.



- Covariance of a variable with itself

$$\sum (x-\bar{x})(y-\bar{y})$$

$$\sum_{i=1} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

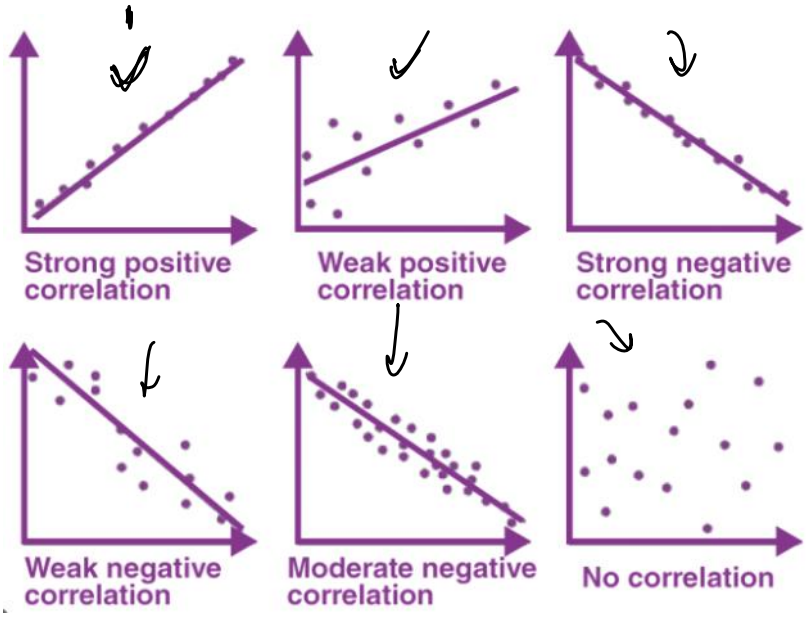
$$= \sum_{i=1}^{n} \frac{(x-\bar{x})(x-\bar{x})}{n-1}$$

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

Correlation

13 March 2023 06:58

1. What problem does Correlation solve?



Can we quantify this weak and strong relationship?

2. What is correlation?

Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together.

Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of -1 indicates a perfect negative correlation, a correlation coefficient of 0 indicates no correlation, and a correlation coefficient of 1 indicates a perfect positive correlation.

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

Correlation and Causation

13 March 2023 18:31

The phrase "**correlation does not imply causation**" means that just because two variables are associated with each other, it does not necessarily mean that one causes the other. In other words, a correlation between two variables does not necessarily imply that one variable is the reason for the other variable's behaviour.

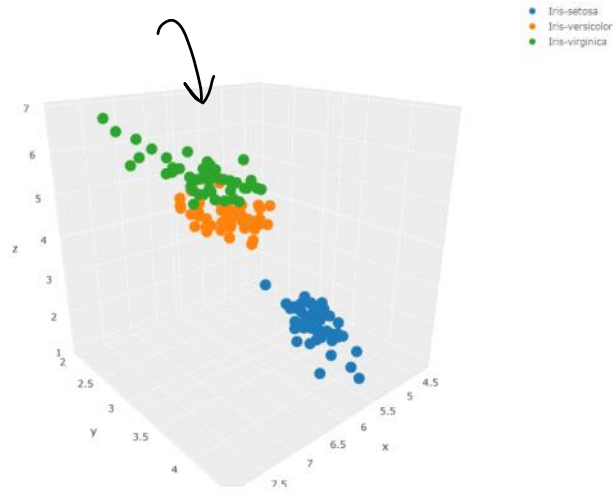
Suppose there is a positive correlation between the number of firefighters present at a fire and the amount of damage caused by the fire. One might be tempted to conclude that the presence of firefighters causes more damage. However, this correlation could be explained by a third variable - the severity of the fire. More severe fires might require more firefighters to be present, and also cause more damage.

Thus, while correlations can provide valuable insights into how different variables are related, they cannot be used to establish causality. Establishing causality often requires additional evidence such as experiments, randomized controlled trials, or well-designed observational studies.

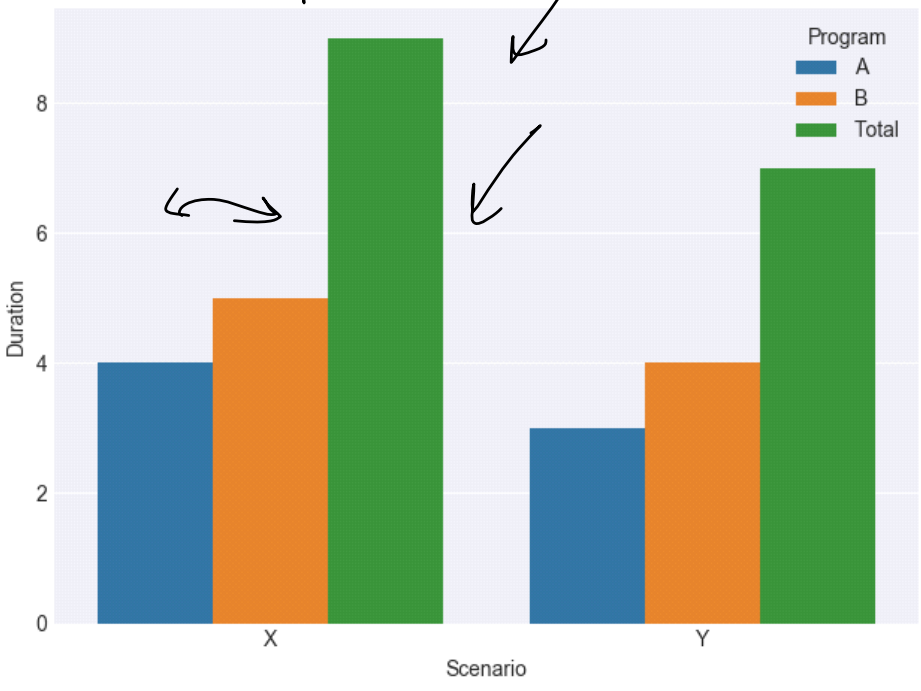
Visualizing Multiple Variables

13 March 2023 06:58

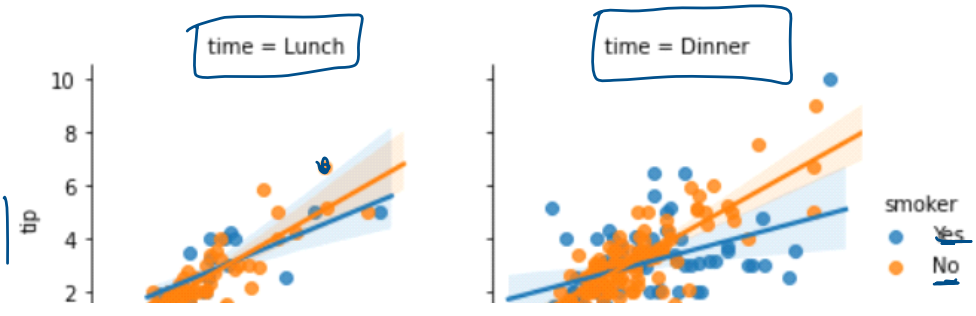
1. 3D Scatter Plots

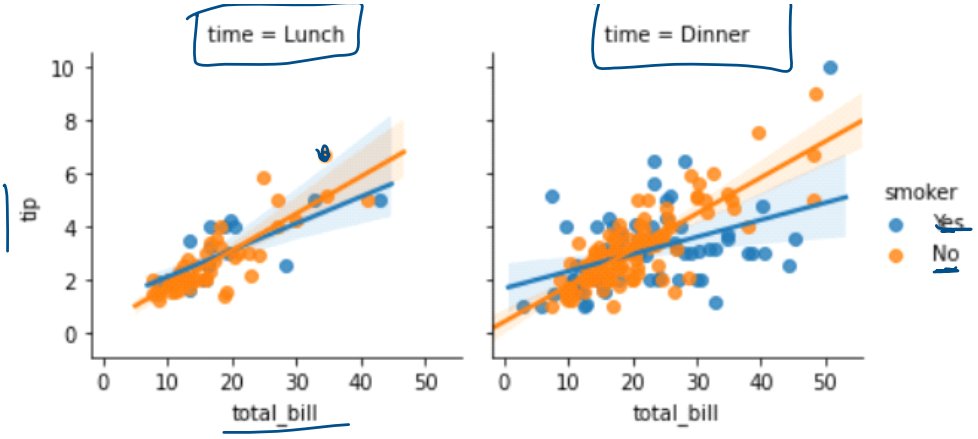


2. Hue Parameter

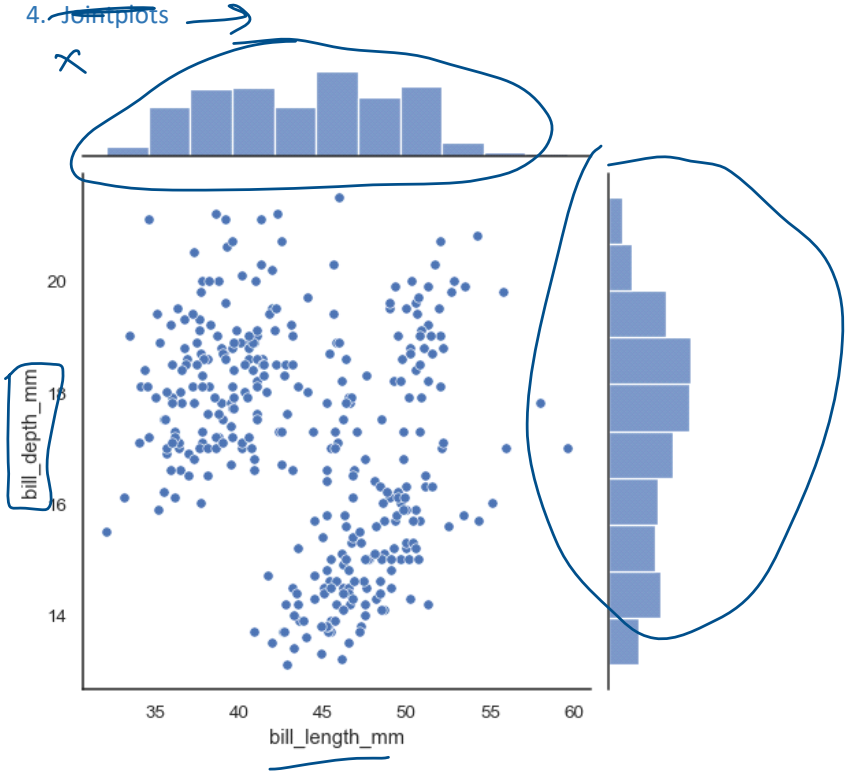


3. Facetgrids

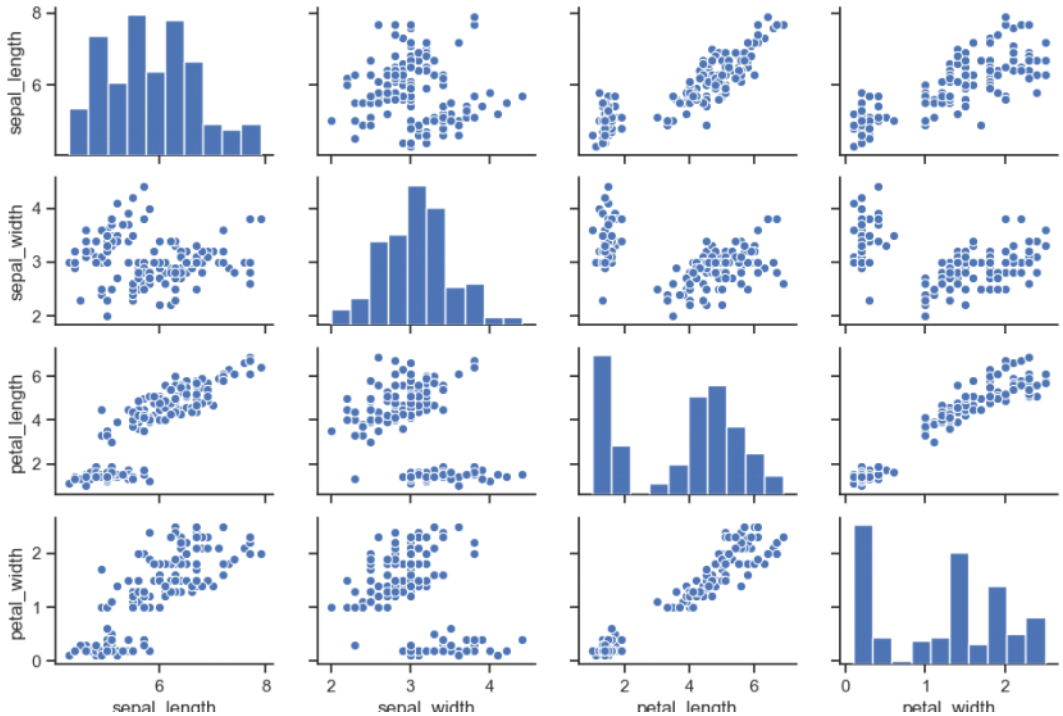




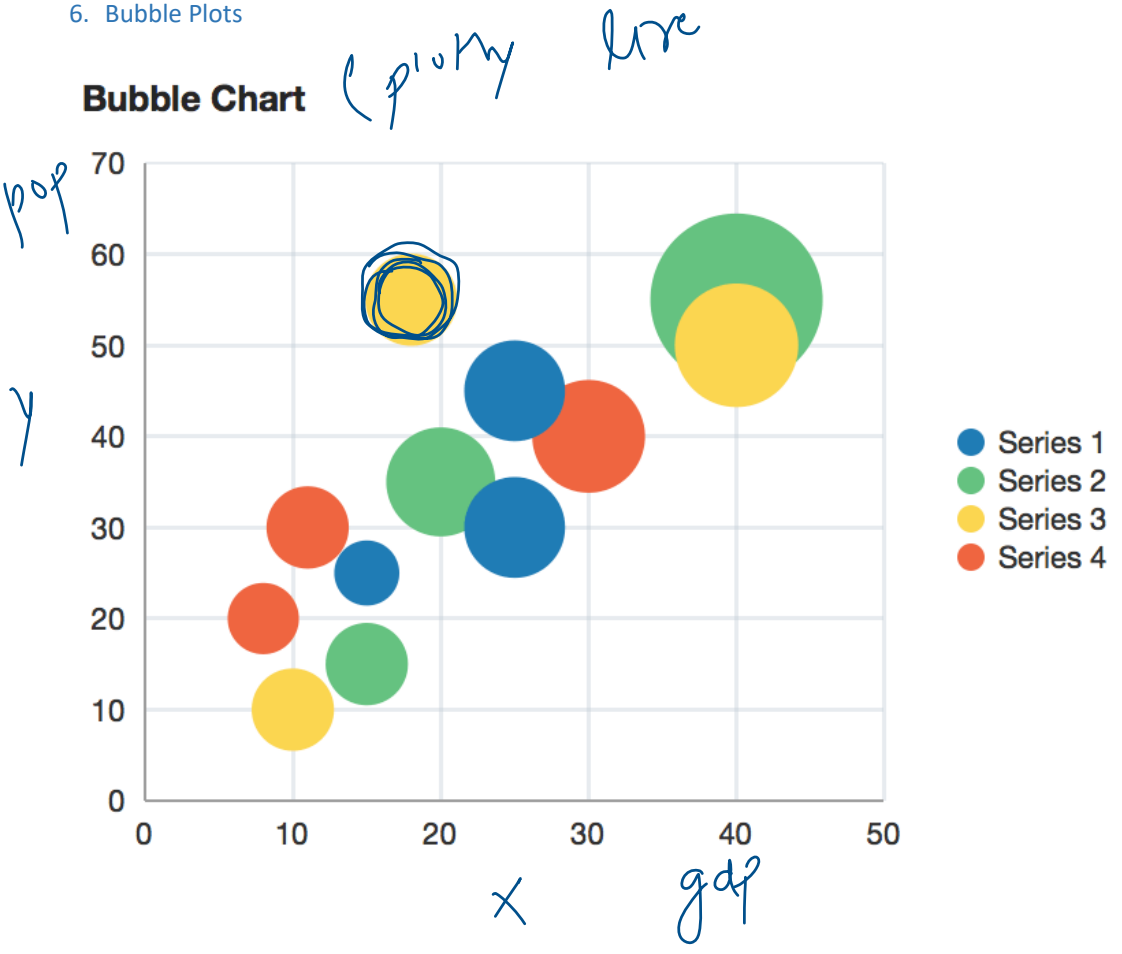
4. Jointplots



5. Pairplots



6. Bubble Plots

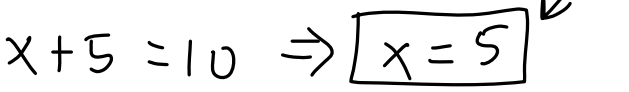


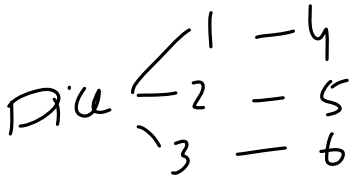
Random Variables

15 March 2023 11:43

- What are Algebraic Variables?

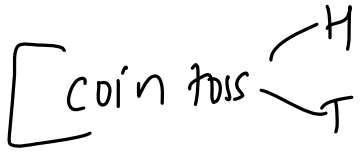
  In Algebra a variable, like x, is an unknown value

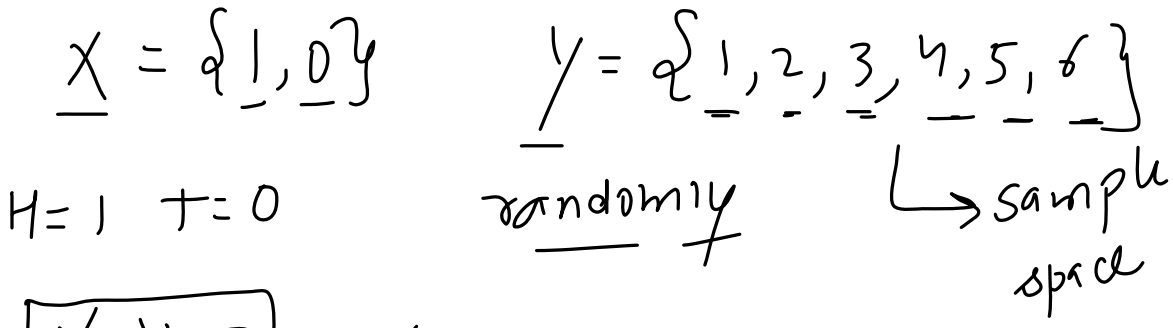




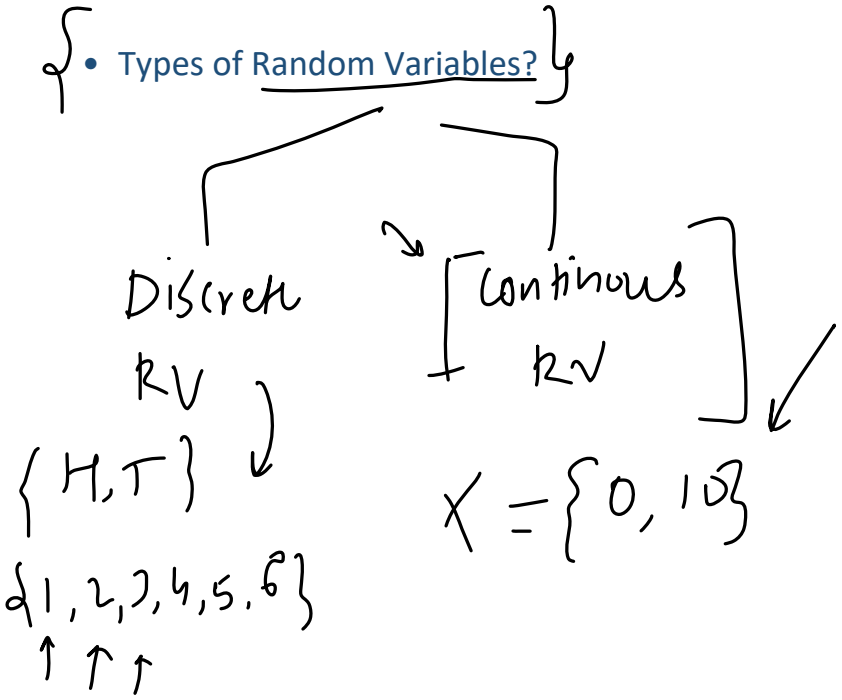
- What are Random Variables in Stats and Probability?

  A Random Variable is a set of possible values from a random experiment.





- Types of Random Variables?



Probability Distributions

15 March 2023 11:53

1. What are Probability Distributions?

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values.

coin toss	1 (H)	0 (T)
probab	$\frac{1}{2}$	$\frac{1}{2}$

2 dice →

2 3 4 5 6 7 8 9

dice

1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

2 → 1/36
3 → 2/36
4 → 3/36
5 → 4/36
6 → 5/36
7 → 6/36
8 → 5/36
9 → 4/36
10 → 3/36
11 → 2/36
12 → 1/36

Problem with Distribution?

In many scenarios, the number of outcomes can be much larger and hence a table would be tedious to write down. Worse still, the number of possible outcomes could be infinite, in which case, good luck writing a table for that.

Example - Height of people, Rolling 10 dice together

→ Solution - Function?

→ What if we use a mathematical function to model the relationship between outcome and probability?

X → outcome
Y → prob

1	1.5	2	3	4	5	6
1/6		1/6	1/6	1/6	1/6	1/6

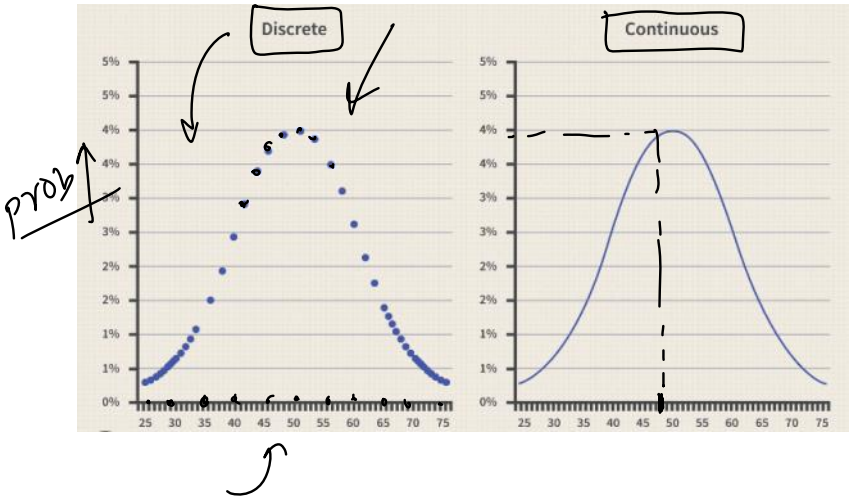
$Y = f(X)$

PDF → table → f(X)

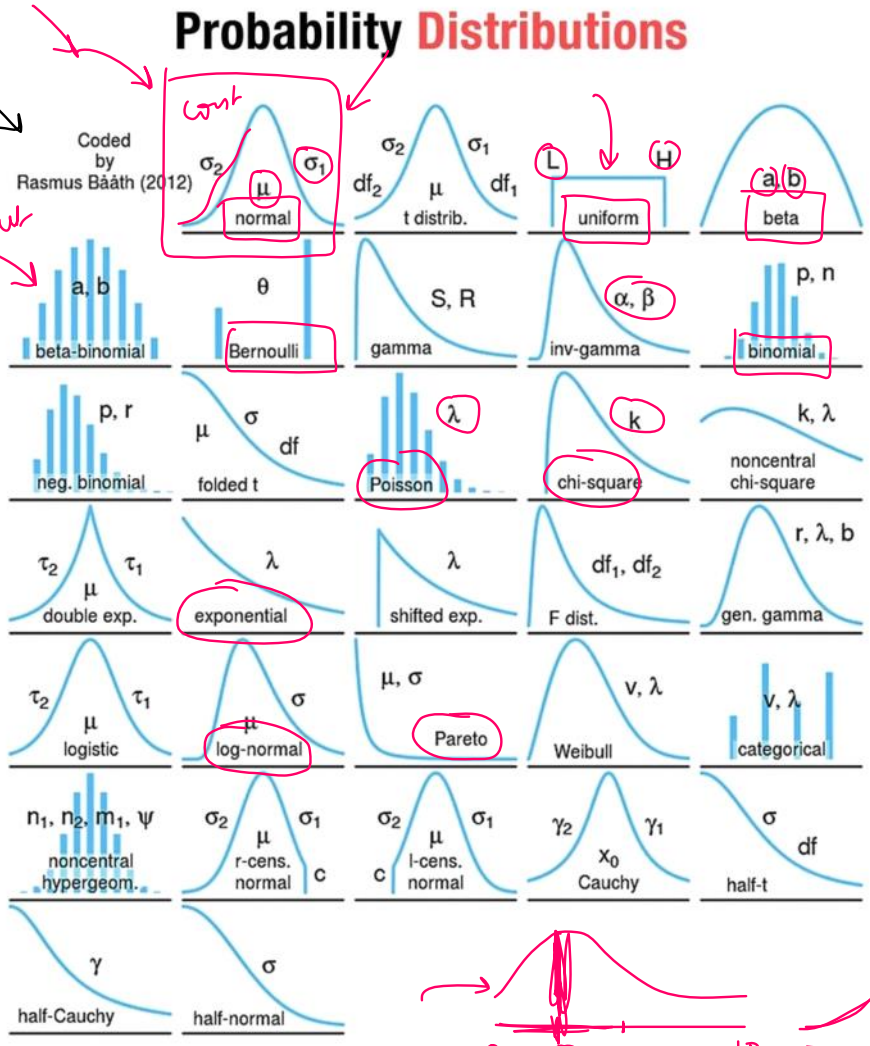
∫ Note - A lot of time Probability Distribution and Probability Distribution Functions are ↴

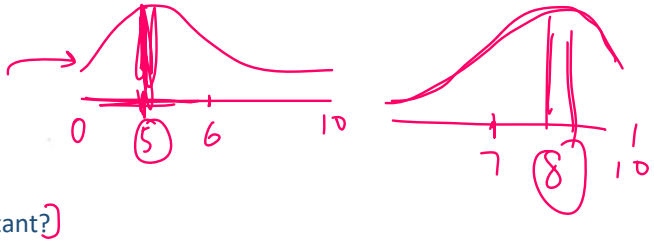
Note - A lot of time Probability Distribution and Probability Distribution Functions are used interchangeably.

1. Types of Probability Distributions (PDF)



Famous Probability Distributions





Why are Probability Distributions important?

- Gives an idea about the shape/distribution of the data.
- And if our data follows a famous distribution then we automatically know a lot about the data.

A note on Parameters (PDF)

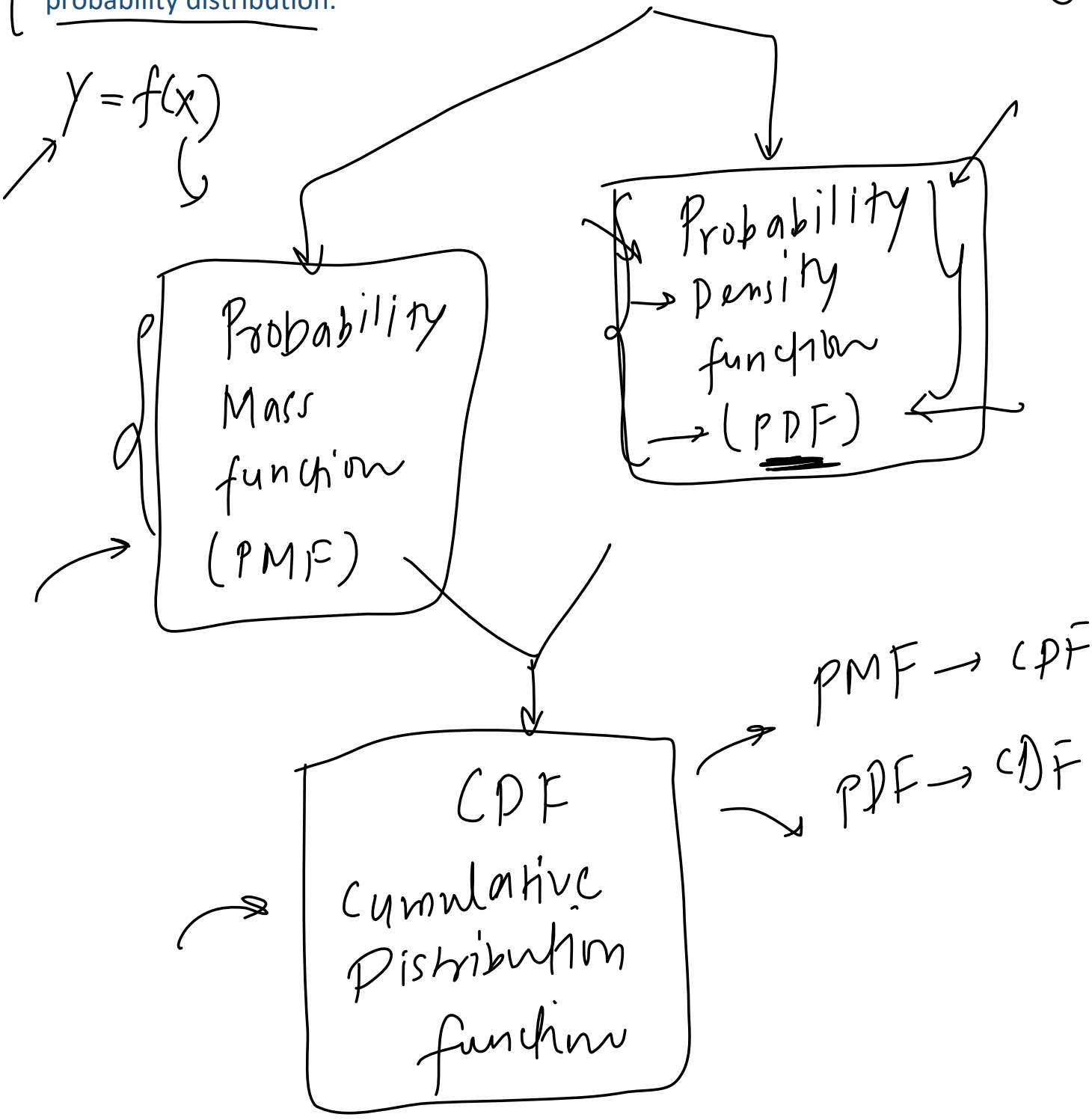
Parameters in probability distributions are numerical values that determine the shape, location, and scale of the distribution.

Different probability distributions have different sets of parameters that determine their shape and characteristics, and understanding these parameters is essential in statistical analysis and inference.

[Probability Distribution Functions] → PDF

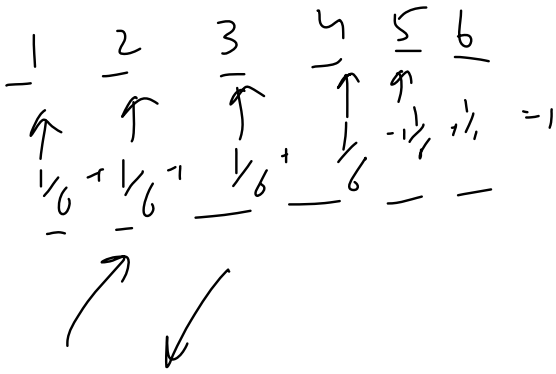
15 March 2023 20:08

A probability distribution function (~~PDF~~) is a mathematical function that describes the probability of obtaining different values of a random variable in a particular probability distribution.



Probability Mass Function (PMF)

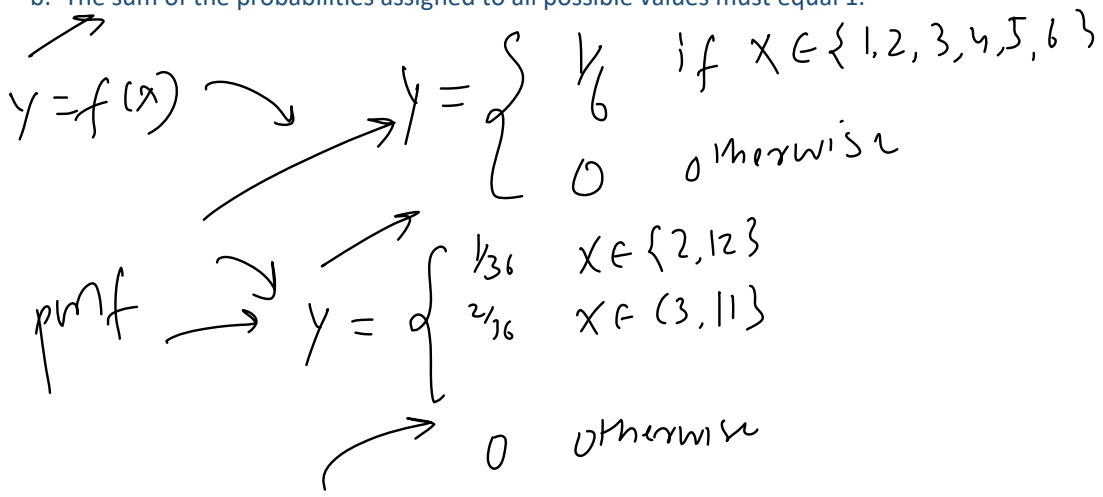
15 March 2023 15:25



PMF stands for Probability Mass Function. It is a mathematical function that describes the probability distribution of a **discrete random variable**.

The PMF of a discrete random variable assigns a probability to each possible value of the random variable. The probabilities assigned by the PMF must satisfy two conditions:

a. The probability assigned to each value must be non-negative (i.e., greater than or equal to zero).
b. The sum of the probabilities assigned to all possible values must equal 1.



Examples

https://en.wikipedia.org/wiki/Bernoulli_distribution

https://en.wikipedia.org/wiki/Binomial_distribution

Cumulative Distribution Function(CDF) of PMF

15 March 2023 20:09

The cumulative distribution function (CDF) F(x) describes the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x

[F(x) = P(X <= x)]

f(x) =

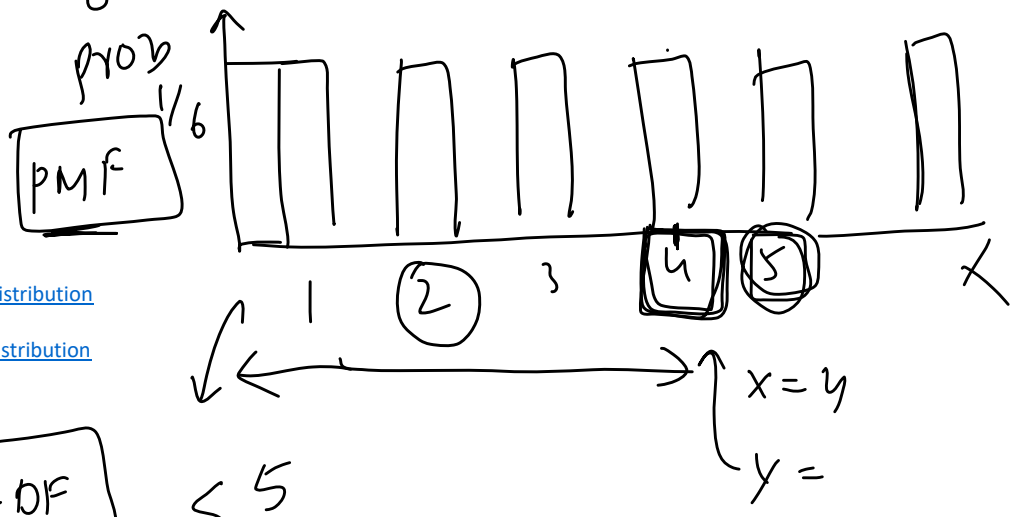
PMF f(x)

f(X=6) = 1/6

f(X=5) = 1/6

f(X≤4) = f(X=4) + f(X=3) + f(X=2) + f(X=1)

1/6 + 1/6 + 1/6 + 1/6



Examples:

https://en.wikipedia.org/wiki/Bernoulli_distribution

https://en.wikipedia.org/wiki/Binomial_distribution

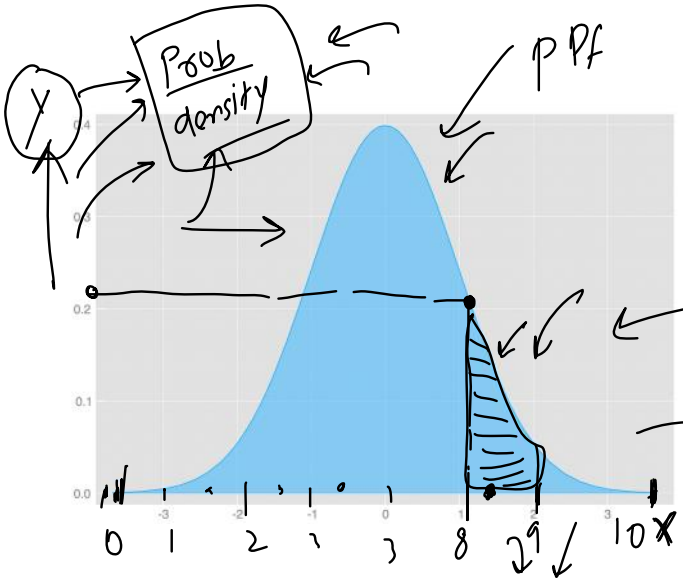
CDF ≤ 5

	PMF	CDF
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6 → 1

Probability Density Function (PDF)

15 March 2023 15:25

PDF stands for Probability Density Function. It is a mathematical function that describes the probability distribution of a **continuous random variable.**



sample PDF

cgpa = 7.912

7.91234567 fg

$\int_8^9 f(x)\,dx$

$P(0 \le x \le 10) = 1$

area

8 → 9

probability between

8 → 8.1

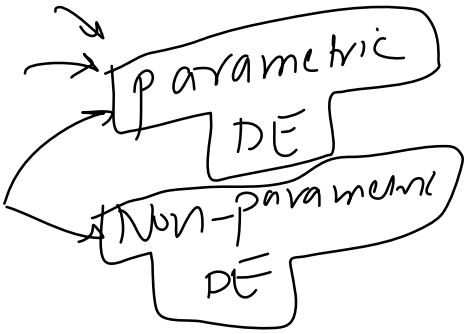
8 → 8.01

8 → 8.001

1. Why Probability Density and why not Probability?
2. What does the area of this graph represents?
3. How to calculate Probability then?
4. Examples of PDF
   a. https://en.wikipedia.org/wiki/Normal_distribution
   b. https://en.wikipedia.org/wiki/Log-normal_distribution
   c. https://en.wikipedia.org/wiki/Poisson_distribution
5. How is graph calculated?

Density Estimation

16 March 2023 06:54



Density estimation is a statistical technique used to estimate the probability density function (PDF) of a random variable based on a set of observations or data. In simpler terms, it involves estimating the underlying distribution of a set of data points.

Density estimation can be used for a variety of purposes, such as **hypothesis testing, data analysis, and data visualization**. It is particularly useful in areas such as **machine learning**, where it is often used to estimate the probability distribution of input data or to model the likelihood of certain events or outcomes.

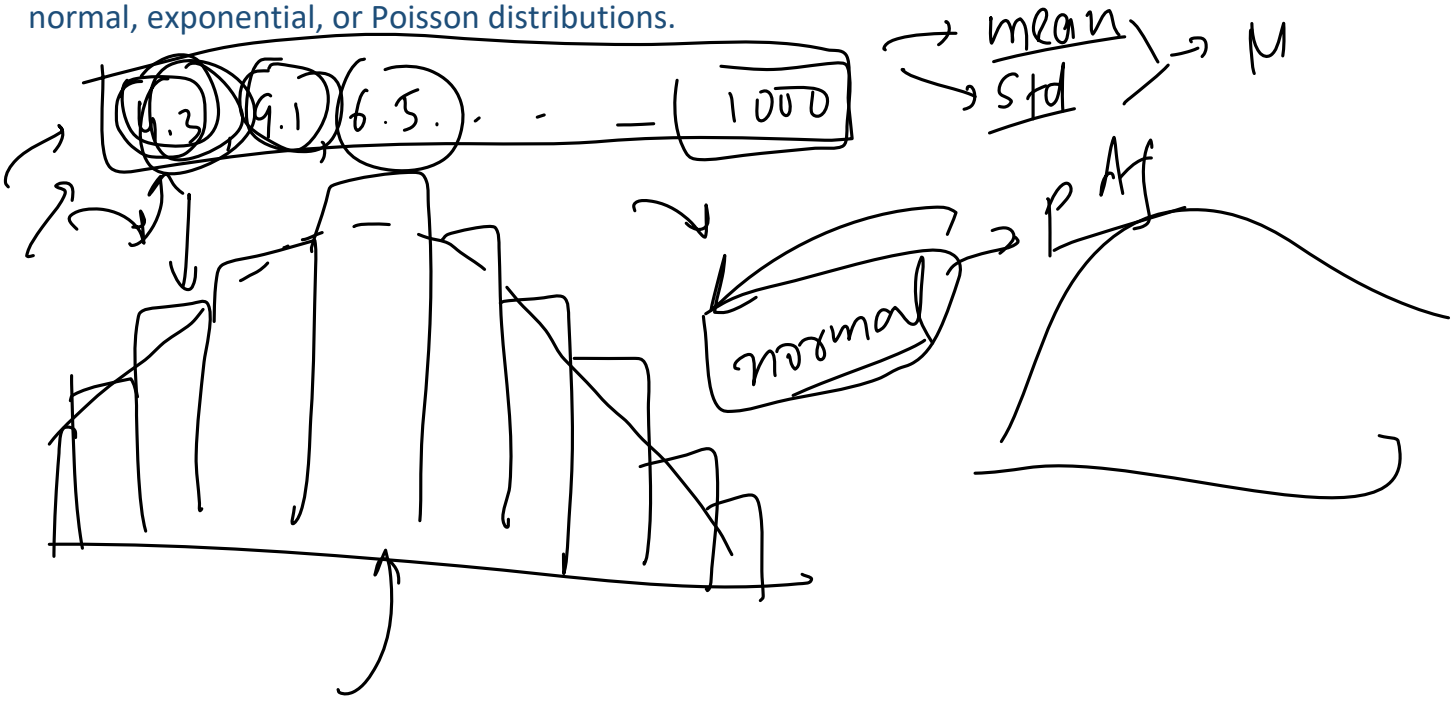
There are various methods for density estimation, including **parametric** and **non-parametric approaches**. Parametric methods assume that the data follows a specific probability distribution (such as a normal distribution), while non-parametric methods do not make any assumptions about the distribution and instead estimate it directly from the data.

Commonly used techniques for density estimation include **kernel density estimation (KDE)**, **histogram estimation**, and **Gaussian mixture models (GMMs)**. The choice of method depends on the specific characteristics of the data and the intended use of the density estimate.

Parametric Density Estimation

16 March 2023 06:54

Parametric density estimation is a method of estimating the probability density function (PDF) of a random variable by assuming that the underlying distribution belongs to a specific parametric family of probability distributions, such as the normal, exponential, or Poisson distributions.



Non-Parametric Density Estimation (KDE)

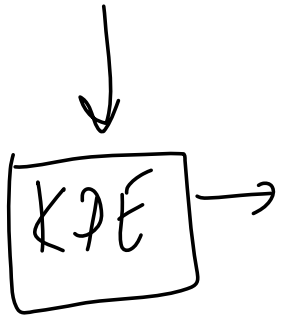
16 March 2023 06:55

But sometimes the distribution is not clear or it's not one of the famous distributions.

Non-parametric density estimation is a statistical technique used to estimate the probability density function of a random variable without making any assumptions about the underlying distribution. It is also referred to as non-parametric density estimation because it does not require the use of a predefined probability distribution function, as opposed to parametric methods such as the Gaussian distribution.

The non-parametric density estimation technique involves constructing an estimate of the probability density function using the available data. This is typically done by creating a **kernel density estimate**

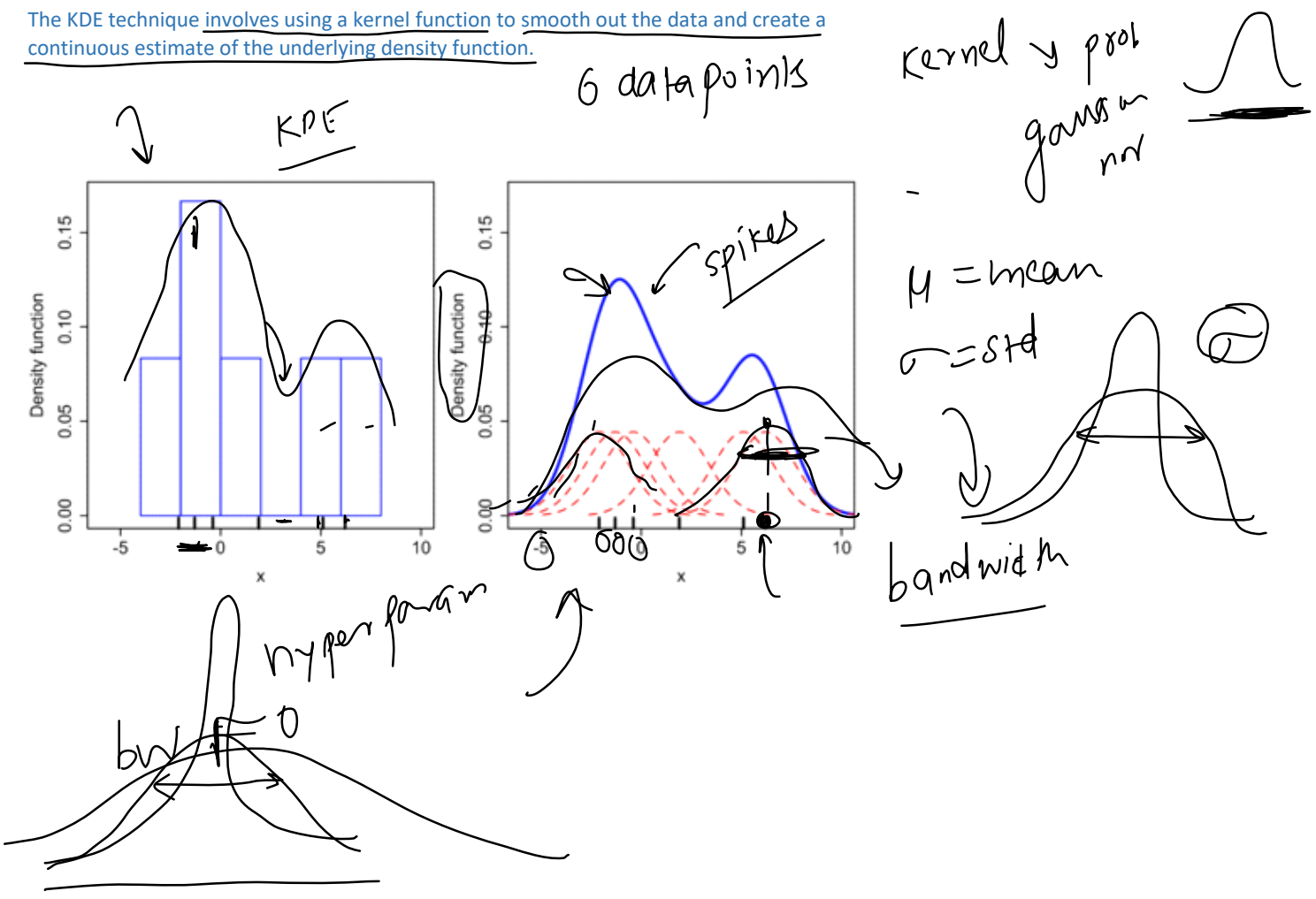
Non-parametric density estimation has several advantages over parametric density estimation. One of the main **advantages** is that it **does not require the assumption of a specific distribution**, which allows for more flexible and accurate estimation in situations where the underlying distribution is unknown or complex. However, non-parametric density estimation can be **computationally intensive and may require more data to achieve accurate estimates** compared to parametric methods.



Kernel Density Estimate(KDE)

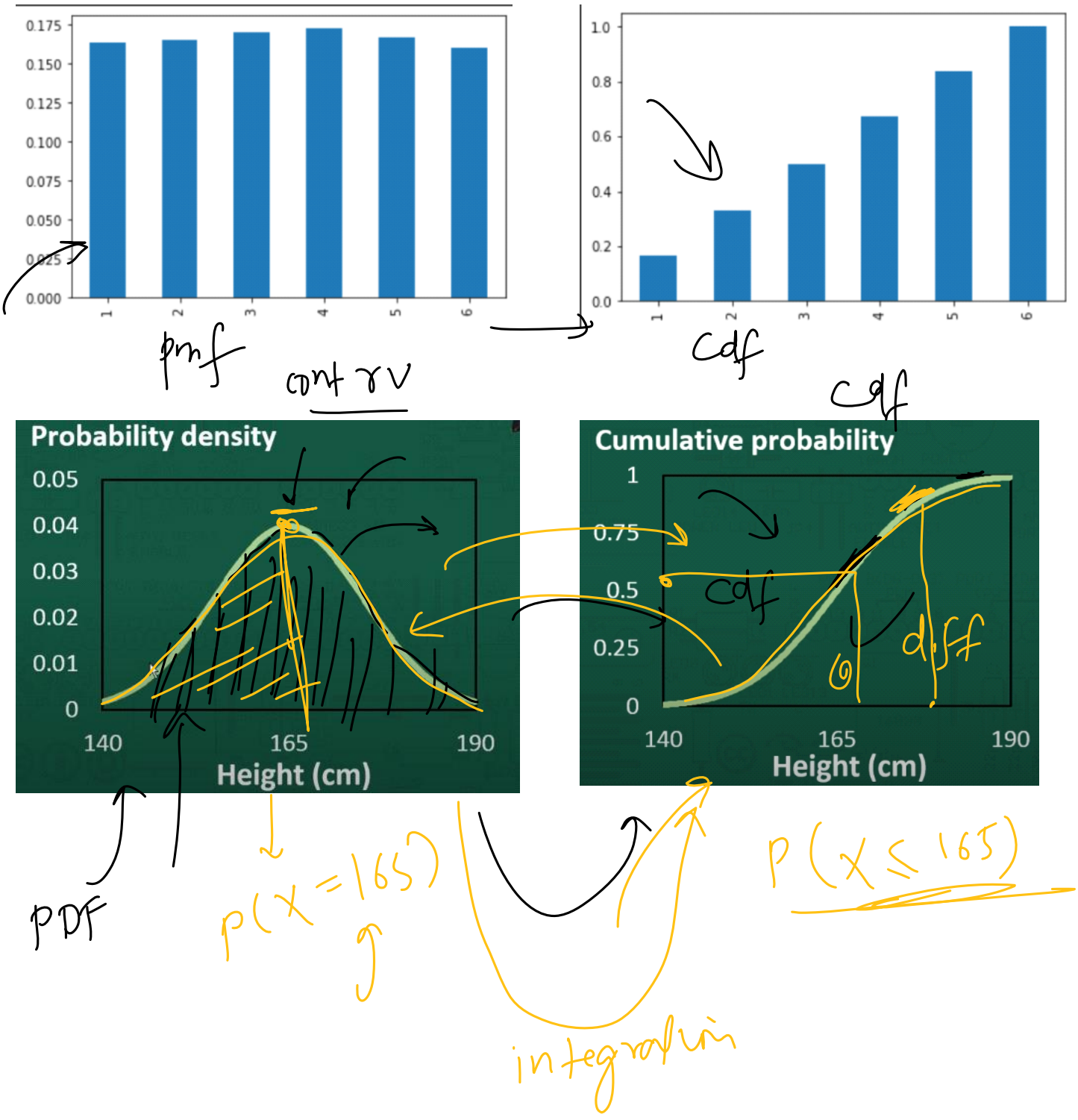
16 March 2023 16:08

The KDE technique involves using a kernel function to smooth out the data and create a continuous estimate of the underlying density function.



Cumulative Distribution Function(CDF) of PDF

15 March 2023 15:25



How to use PDF and CDF in Data Analysis

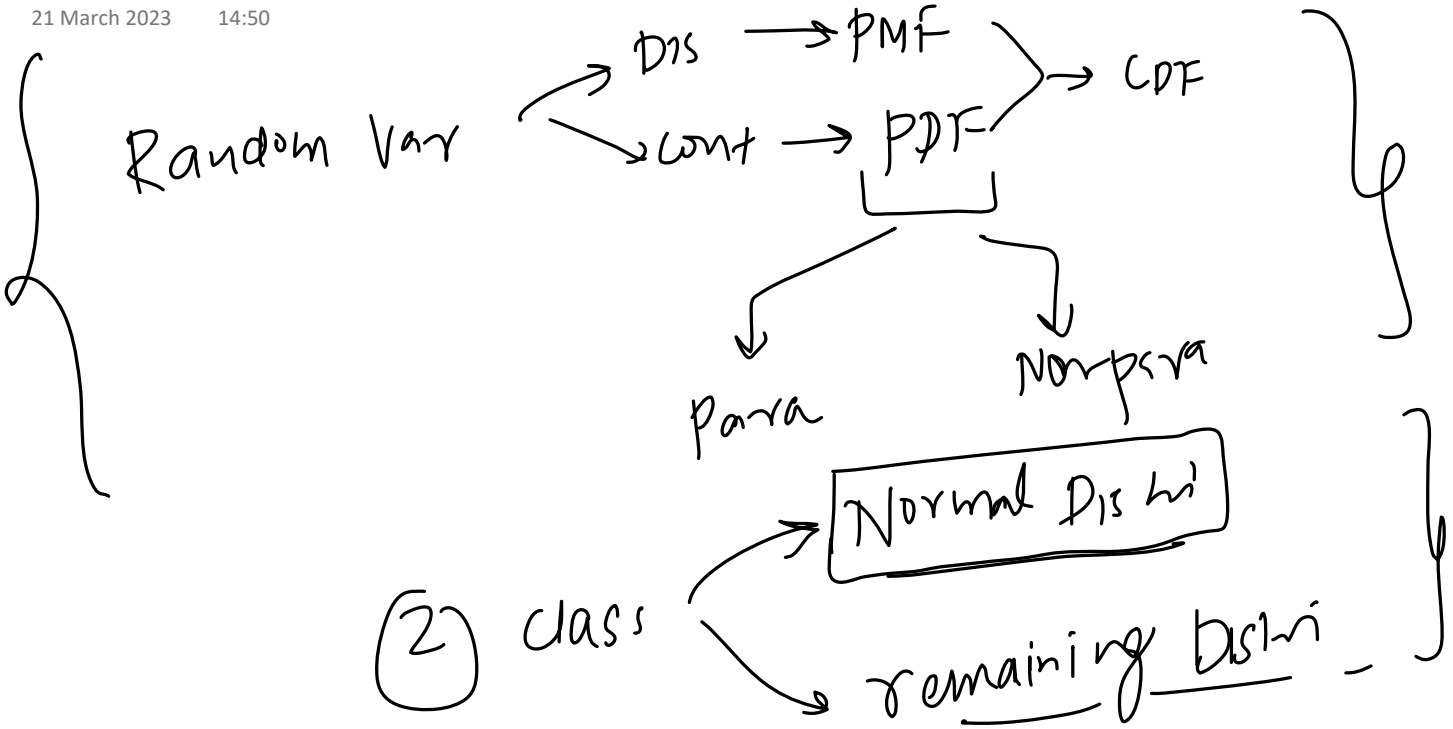
15 March 2023 20:10

2D Probability Density Plots

16 March 2023 06:50

Recap

21 March 2023 14:50



How to use PDF in Data Science

20 March 2023 18:11

PDF → CDF

PDF $P(X = x)$

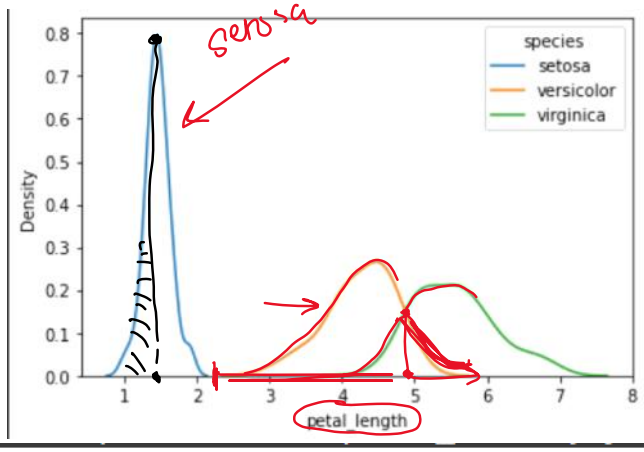
$2.3 < pl < 5$ → versicolor

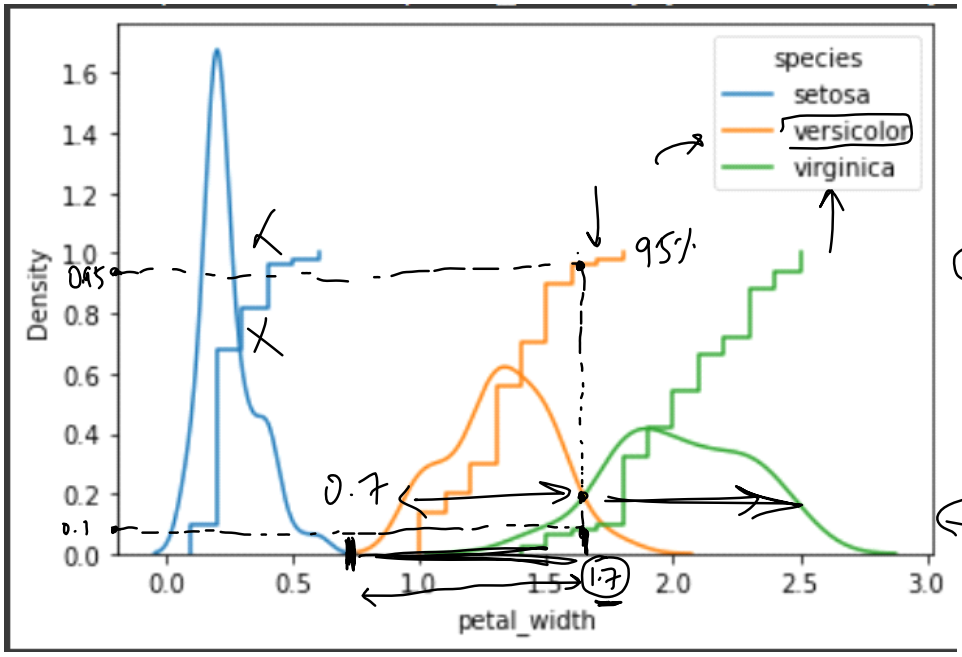
$P(X \leq x)$

$pl < 2.3$

> 2.3

$pl > 5$ virginica





95% of times

If $0.7 < pw < 1.7$ → versicolor

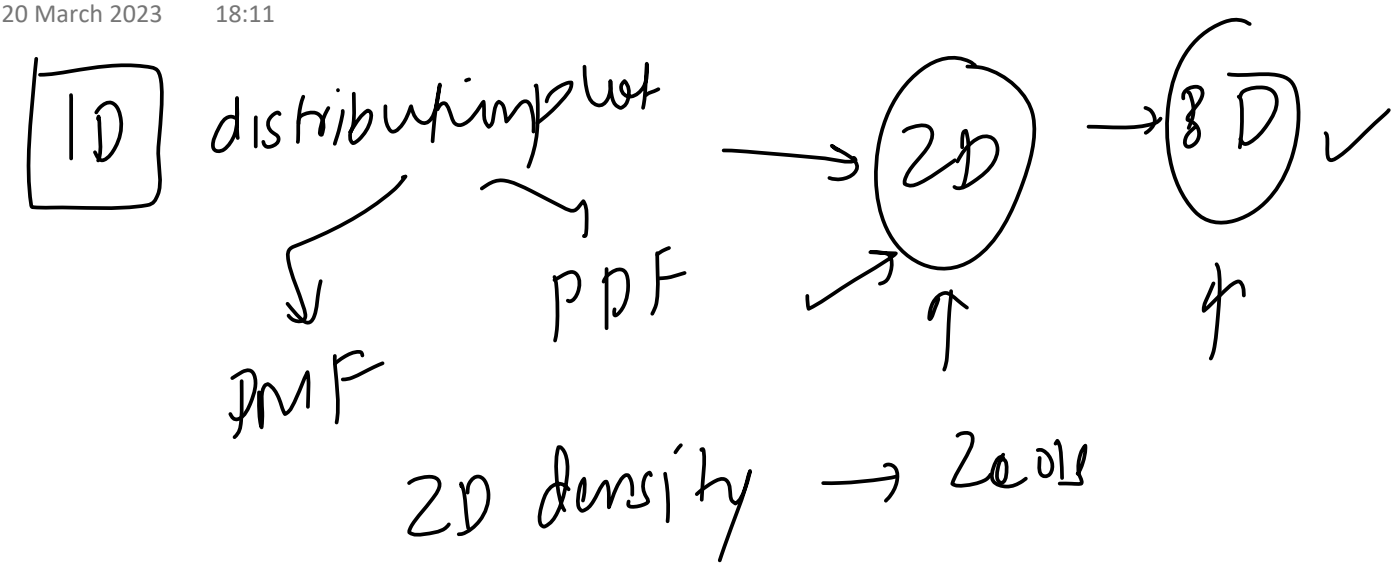
$pw > 1.7$ → virgin

90%

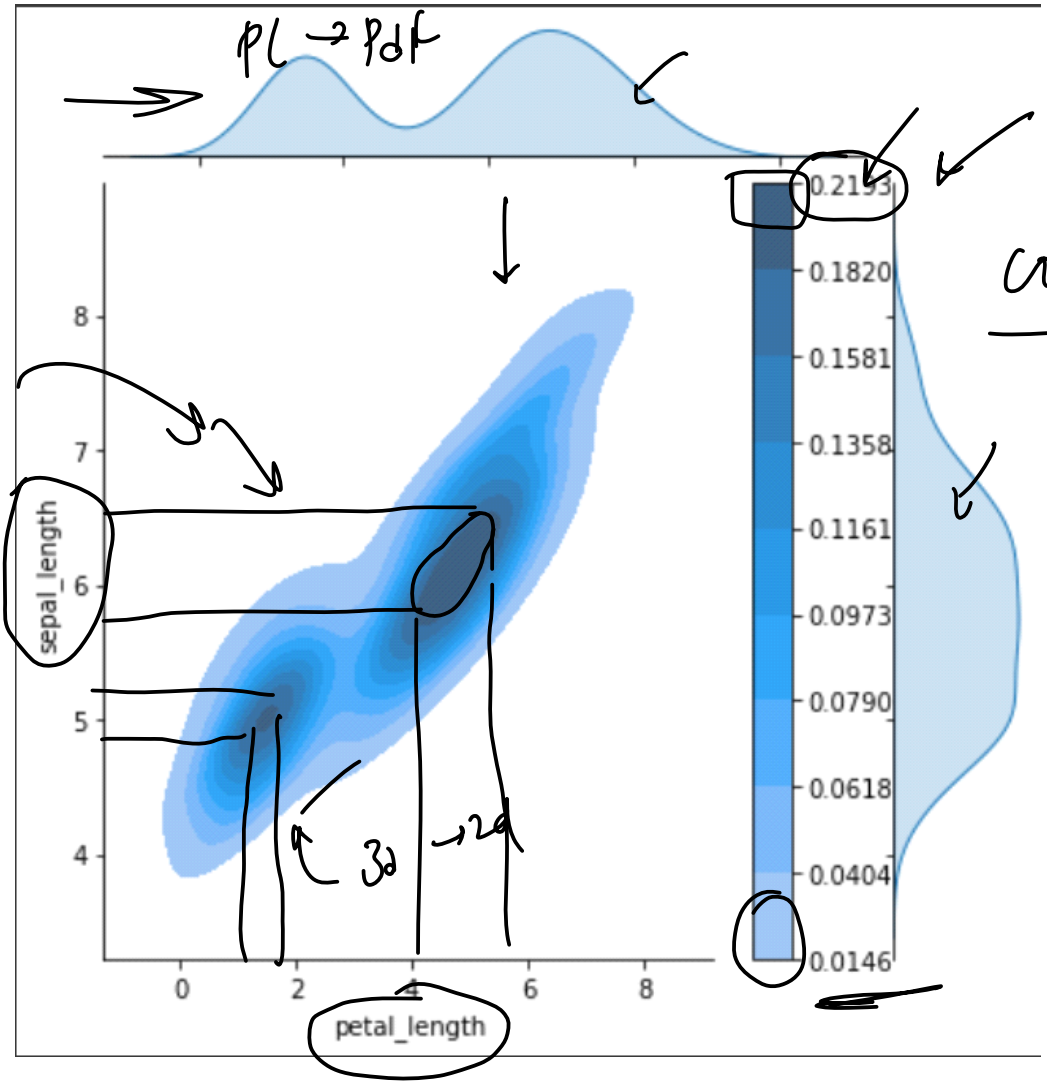
10% virgin

2D Density Plots

20 March 2023 18:11



2D density → 2 cols



2D

Contour

sl - pl

3d → 2d

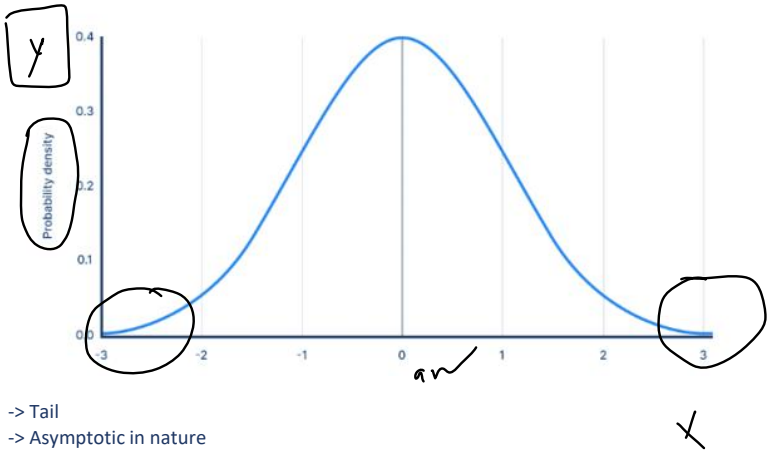
Color

Normal Distribution

20 March 2023 18:06

1. What is normal distribution?

Normal distribution, also known as Gaussian distribution, is a probability distribution that is commonly used in statistical analysis. It is a continuous probability distribution that is symmetrical around the mean, with a bell-shaped curve.



-> Tail
-> Asymptotic in nature
-> Lots of points near the mean and very few far away

The normal distribution is characterized by two parameters: the mean (μ) and the standard deviation (σ). The mean represents the centre of the distribution, while the standard deviation represents the spread of the distribution.

Denoted as:

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

Why is it so important?

Commonality in Nature: Many natural phenomena follow a normal distribution, such as the heights of people, the weights of objects, the IQ scores of a population, and many more. Thus, the normal distribution provides a convenient way to model and analyse such data.

PDF Equation of Normal Distribution

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

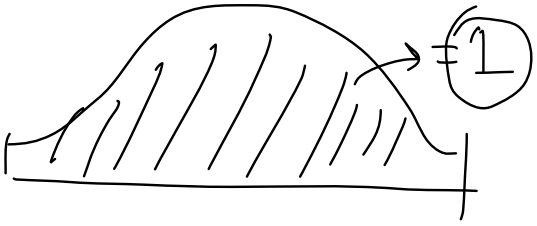
Parameters in Normal Distribution

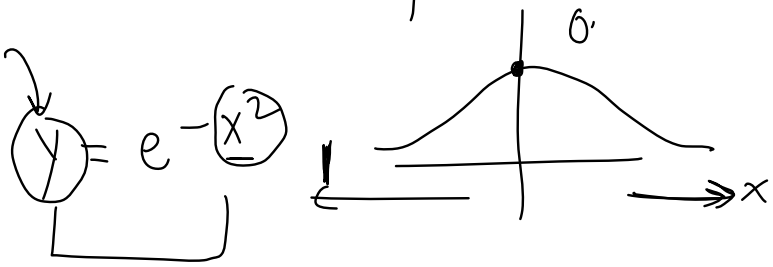
https://samp-suman-normal-dist-visualize-app-lkntug.streamlit.app/

Equation in detail:

$$y = e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\sigma\sqrt{2\pi}$





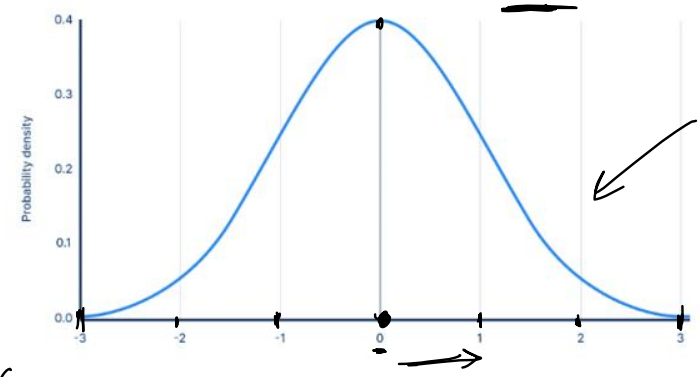
$$y = \frac{1}{e^{x^2}}$$

Standard Normal Variate

20 March 2023 18:08

- What is Standard Normal Variate

A Standard Normal Variate(Z) is a standardized form of the normal distribution with mean = 0 and standard deviation = 1.



Standardizing a normal distribution allows us to compare different distributions with each other, and to calculate probabilities using standardized tables or software.

Equation:

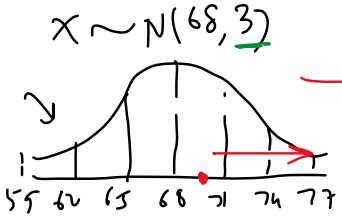
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

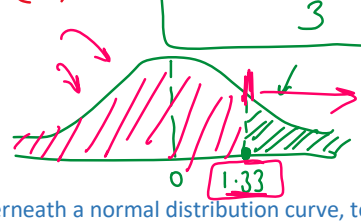
- How to transform a normal distribution to Standard Normal Variate

Refer Python code

Kya Fayda Standardize karne ka?

Suppose the heights of adult males in a certain population follow a normal distribution with a mean of 68 inches and a standard deviation of 3 inches. What is the probability that a randomly selected adult male from this population is taller than 72 inches?





- What are Z-tables

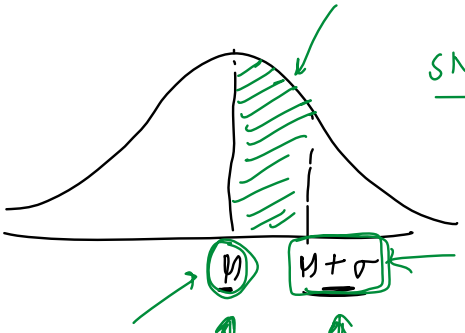
A z-table tells you the area underneath a normal distribution curve, to the left of the z-score

https://www.ztable.net/

For a Normal Distribution X~(u,std) what percent of population lie between mean and 1 standard deviation, 2 std and 3 std?

standard deviation, 2 std and 3 std.

$X \sim N(\mu, \sigma)$ %



SNV(z) →

$\mu \rightarrow$ ↓

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

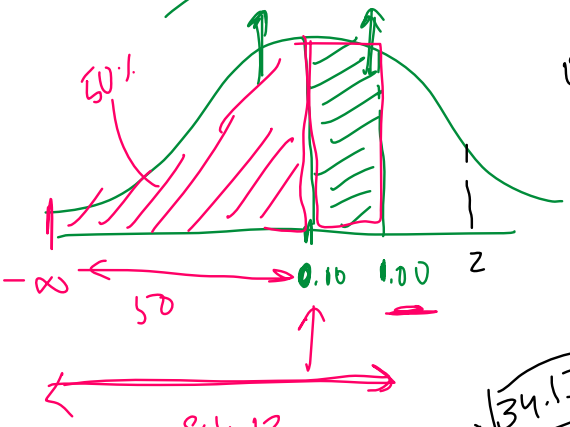
$$z = \frac{\mu-\mu}{\sigma} = 0$$

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

0.9772 = 97%

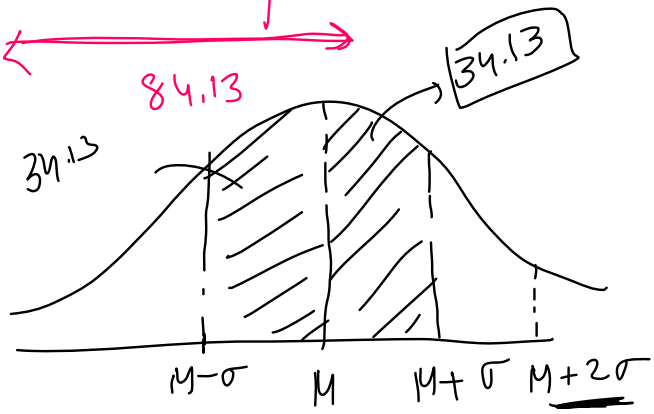
$$\frac{\sigma}{\sigma} = 1$$

84.13 − 50 = 34.13

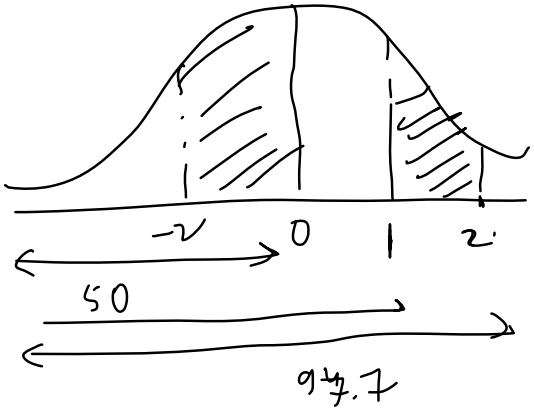


−1σ +1σ

64.28 %



$$z = \frac{\mu+2\sigma-\mu}{\sigma} = 2$$

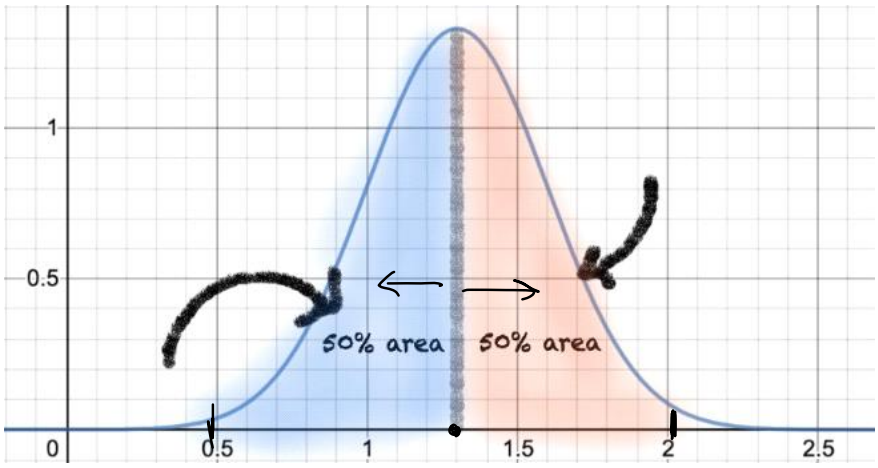


Properties of Normal Distribution

20 March 2023 18:06

1. Symmetricity

The normal distribution is symmetric about its mean which means that the probability of observing a value above the mean is the same as the probability of observing a value below the mean. The bell-shaped curve of the normal distribution reflects this symmetry.



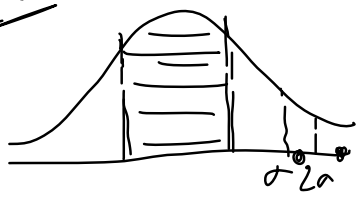
2. Measures of Central Tendencies are equal → mean → median → mode

3. Empirical Rule

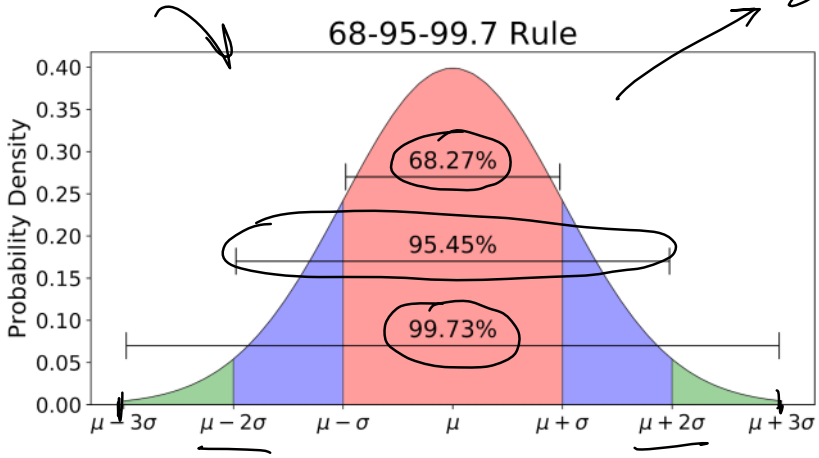
The normal distribution has a well-known empirical rule, also called the 68-95-99.7 rule, which states that approximately 68% of the data falls within one standard deviation of the mean, about 95% of the data falls within two standard deviations of the mean, and about 99.7% of the data falls within three standard deviations of the mean.

Standard normal variate

Z-table

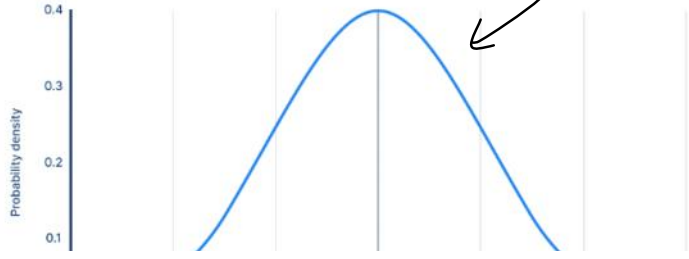


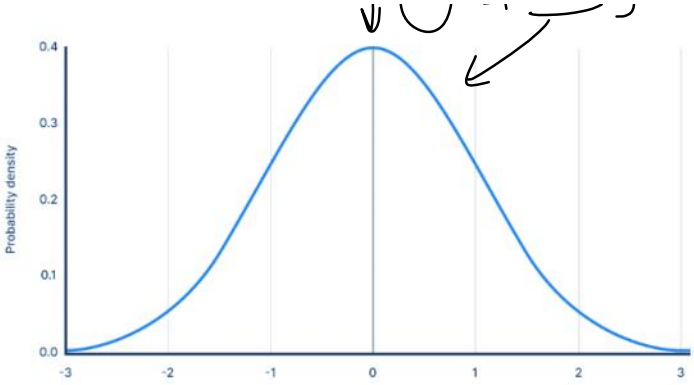
5σ



4. The area under the curve

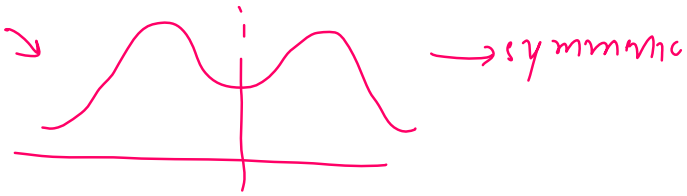
1 → PDF




Probability density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

Skewness

20 March 2023 18:07



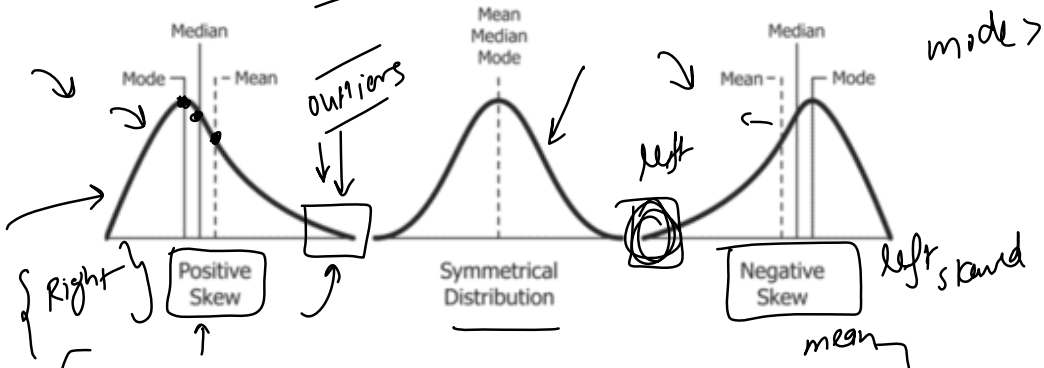
- What is skewness?

A normal distribution is a bell-shaped, symmetrical distribution with a specific mathematical formula that describes how the data is spread out. Skewness indicates that the data is not symmetrical, which means it is not normally distributed.

Skewness is a measure of the asymmetry of a probability distribution. It is a statistical measure that describes the degree to which a dataset deviates from the normal distribution.

In a symmetrical distribution, the mean, median, and mode are all equal. In contrast, in a skewed distribution, the mean, median, and mode are not equal, and the distribution tends to have a longer tail on one side than the other.

Skewness can be positive, negative, or zero. A positive skewness means that the tail of the distribution is longer on the right side, while a negative skewness means that the tail is longer on the left side. A zero skewness indicates a perfectly symmetrical distribution.

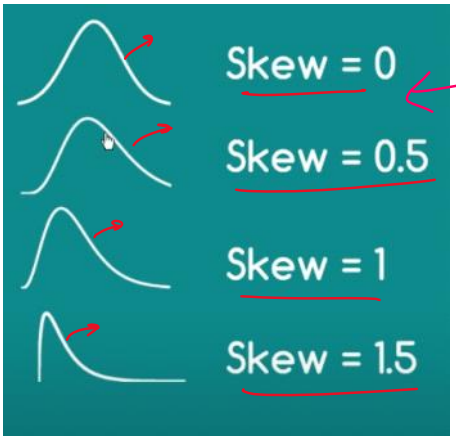


The greater the skew the greater the distance between mode, median and mode

- How skewness is calculated?

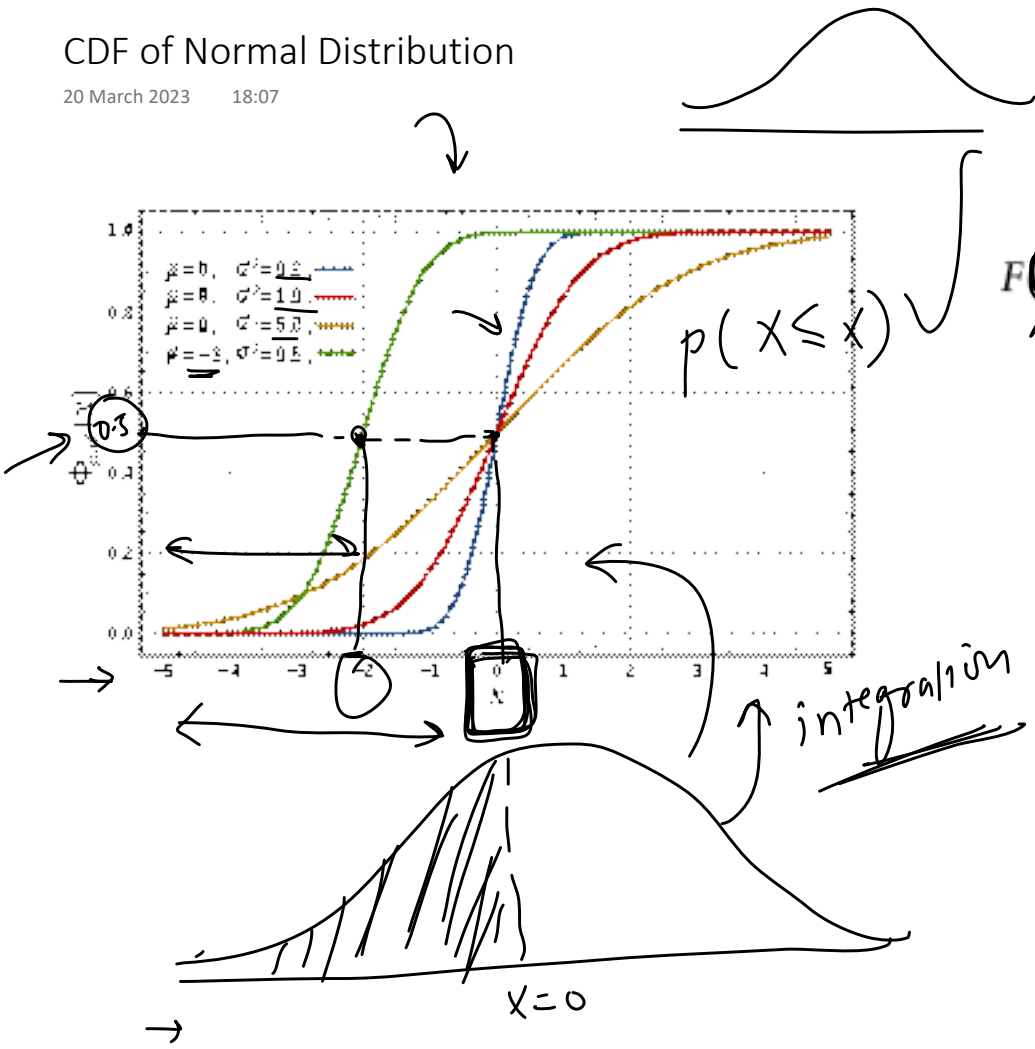
$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{(x-\bar{x})}{s}\right)^3$$

- Python Example
- Interpretation



CDF of Normal Distribution

20 March 2023 18:07



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

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

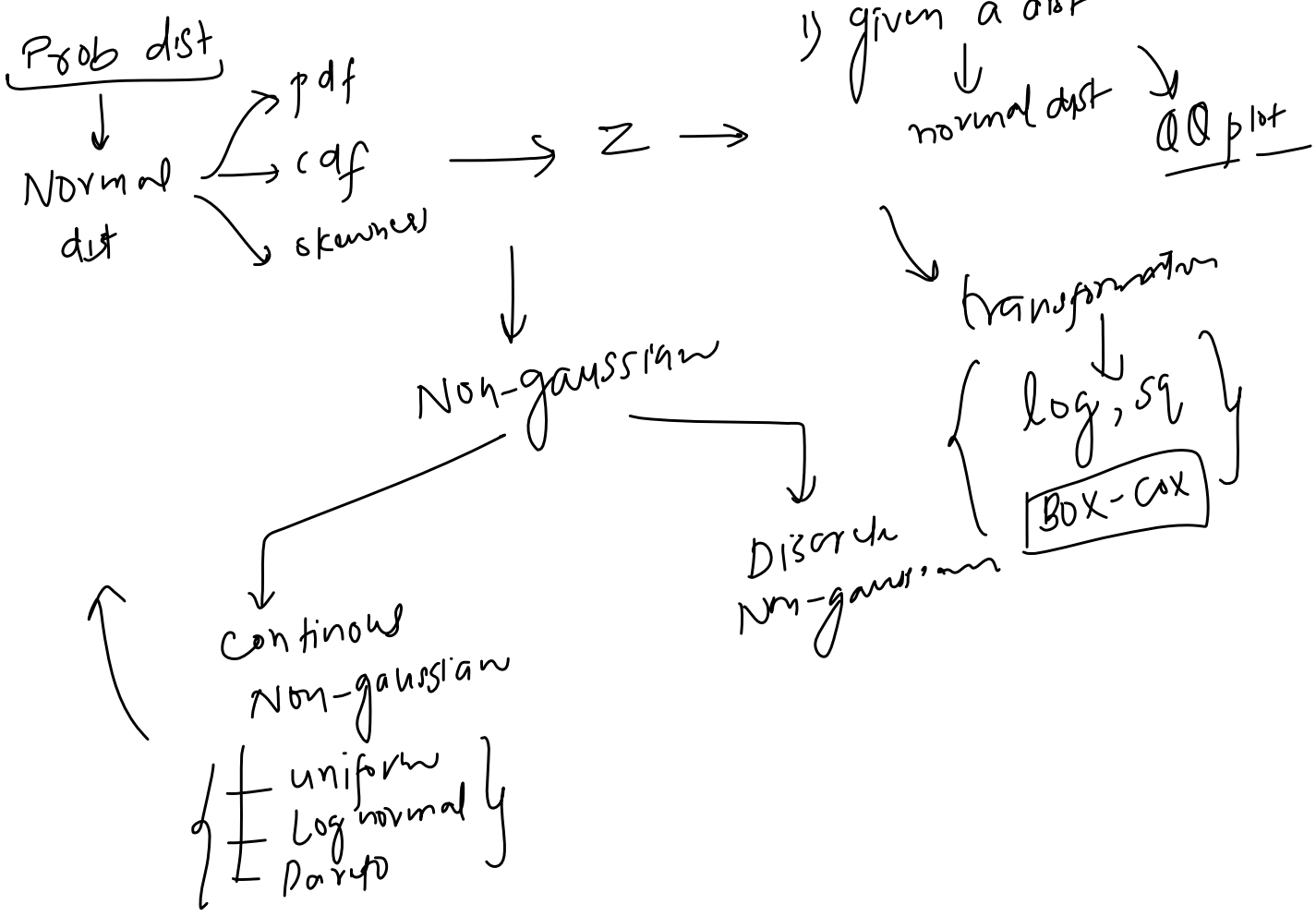
Use in Data Science

20 March 2023 18:08

- Outlier detection
- Assumptions on data for ML algorithms -> Linear Regression and GMM
- Hypothesis Testing
- Central Limit Theorem

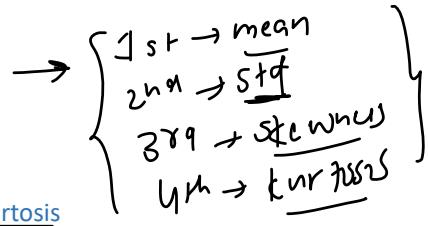
Recap

23 March 2023 18:27



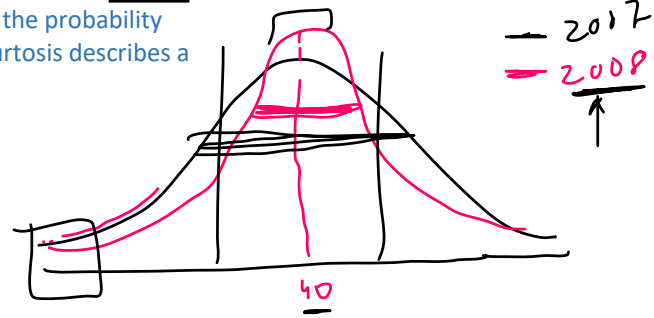
Kurtosis

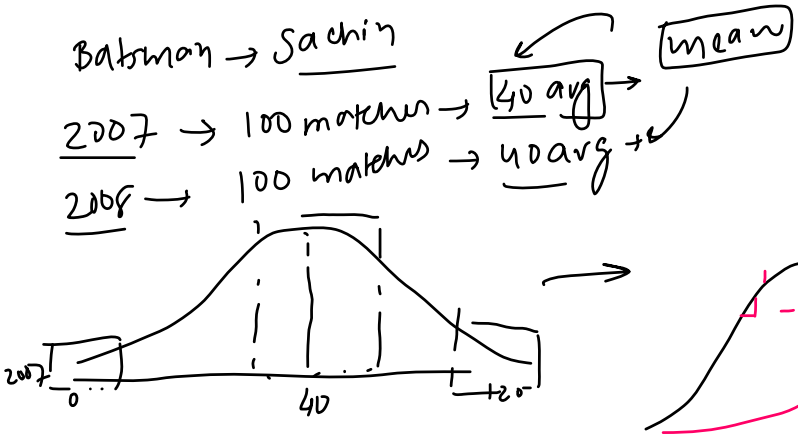
23 March 2023 13:18

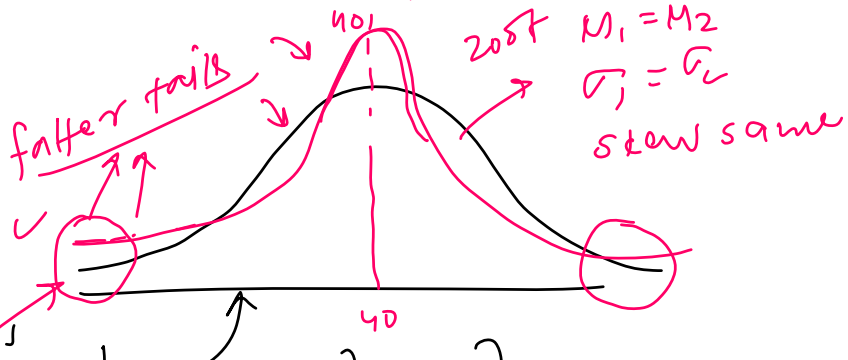


- What is Kurtosis?

Kurtosis is the 4th statistical moment. In probability theory and statistics, kurtosis (meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable. Like skewness, kurtosis describes a particular aspect of a probability distribution.







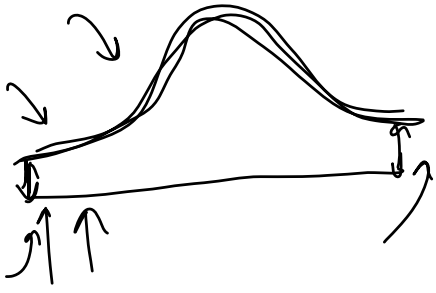
- False notation about Kurtosis

https://en.wikipedia.org/wiki/Kurtosis

- Formula

sample kurtosis

$$\left\{\frac{n*(n+1)}{(n-1)*(n-2)*(n-3)} * \sum_i^n \left(\frac{x_i - \bar{x}}{s}\right)^4\right\} - \frac{3*(n-1)^2}{(n-2)*(n-3)}$$



- Practical Use-case

In finance, kurtosis risk refers to the risk associated with the possibility of extreme outcomes or "fat tails" in the distribution of returns of a particular asset or portfolio.

If a distribution has high kurtosis, it means that there is a higher likelihood of extreme events occurring, either positive or negative, compared to a normal distribution.

In finance, kurtosis risk is important to consider because it indicates that there is a greater probability of large losses or gains occurring, which can have significant implications for investors. As a result, investors may want to adjust their investment strategies to account for kurtosis risk.

- Excess Kurtosis & Types

Excess kurtosis is a measure of how much more peaked or flat a distribution is

Excess Kurtosis & Types

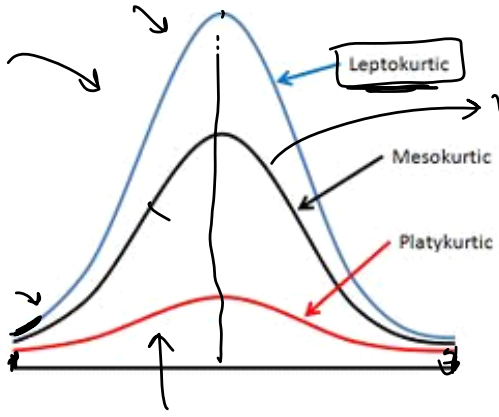
Excess kurtosis is a measure of how much more peaked or flat a distribution is compared to a normal distribution, which is considered to have a kurtosis of 0. It is calculated by subtracting 3 from the sample kurtosis coefficient.

Types of Kurtosis

Leptokurtic

A distribution with positive excess kurtosis is called leptokurtic. "Lepto-" means "slender". In terms of shape, a leptokurtic distribution has fatter tails. This indicates that there are more extreme values or outliers in the distribution.

Example - Assets with positive excess kurtosis are riskier and more volatile than those with a normal distribution, and they may experience sudden price movements that can result in significant gains or losses.



Platykurtic

A distribution with negative excess kurtosis is called platykurtic. "Platy-" means "broad". In terms of shape, a platykurtic distribution has thinner tails. This indicates that there are fewer extreme values or outliers in the distribution.

Assets with negative excess kurtosis are less risky and less volatile than those with a normal distribution, and they may experience more gradual price movements that are less likely to result in large gains or losses.

Mesokurtic

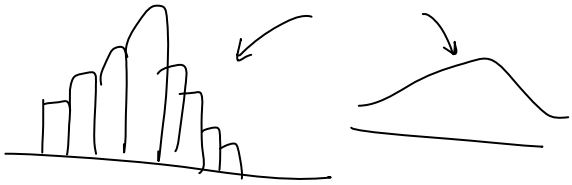
Distributions with zero excess kurtosis are called mesokurtic. The most prominent example of a mesokurtic distribution is the normal distribution family, regardless of the values of its parameters.

Mesokurtic is a term used to describe a distribution with a excess kurtosis of 0, indicating that it has the same degree of "peakedness" or "flatness" as a normal distribution.

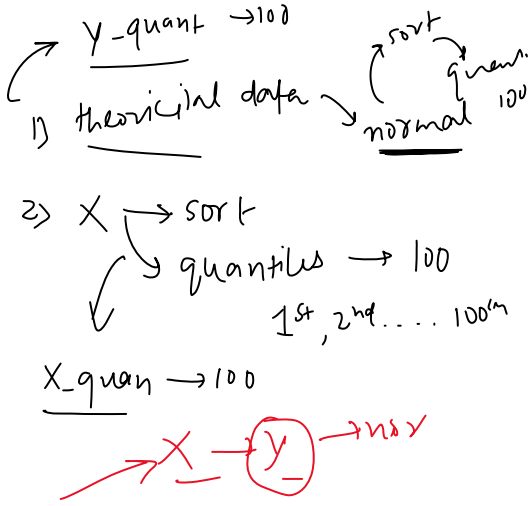
Example -
In finance, a mesokurtic distribution is considered to be the ideal distribution for assets or portfolios, as it represents a balance between risk and return.

QQ Plot

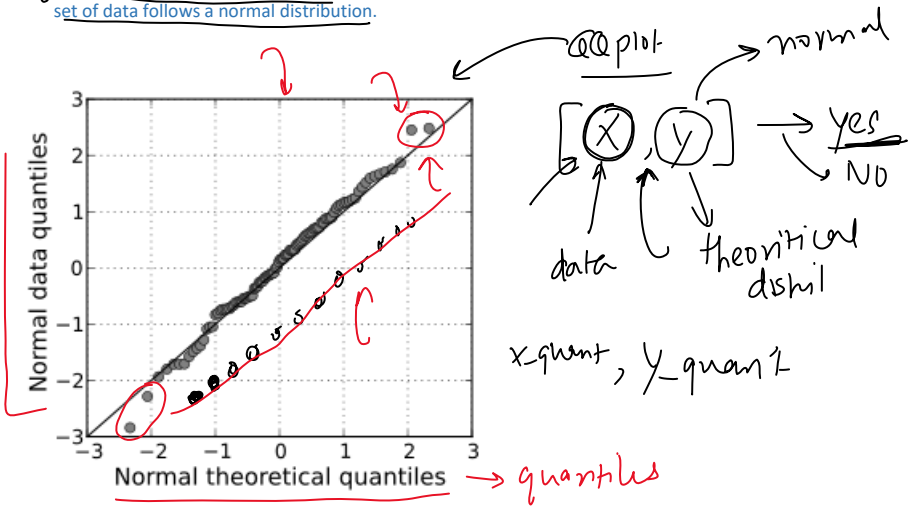
23 March 2023 13:19



- How to find if a given distribution is normal or not?
  - *Visual inspection* One of the easiest ways to check for normality is to visually inspect a histogram or a density plot of the data. A normal distribution has a bell-shaped curve, which means that the majority of the data falls in the middle, and the tails taper off symmetrically. If the distribution looks approximately bell-shaped, it is likely to be normal.
  - *QQ Plot* Another way to check for normality is to create a normal probability plot (also known as a Q-Q plot) of the data. A normal probability plot plots the observed data against the expected values of a normal distribution. If the data points fall along a straight line, the distribution is likely to be normal.
  - *Statistical tests*: There are several statistical tests that can be used to test for normality, such as the Shapiro-Wilk test, the Anderson-Darling test, and the Kolmogorov-Smirnov test. These tests compare the observed data to the expected values of a normal distribution and provide a p-value that indicates whether the data is likely to be normal or not. A p-value less than the significance level (usually 0.05) suggests that the data is not normal.



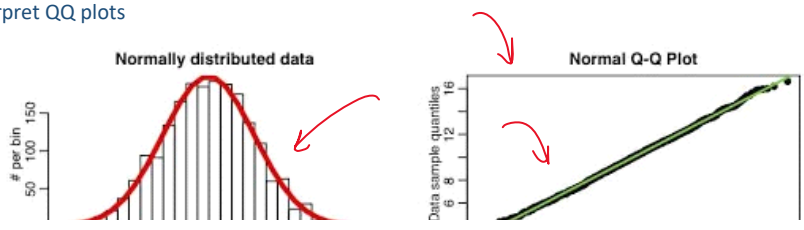
- What is a QQ Plot and how is it plotted?
  - A QQ plot (quantile-quantile plot) is a graphical tool used to assess the similarity of the distribution of two sets of data. It is particularly useful for determining whether a set of data follows a normal distribution.



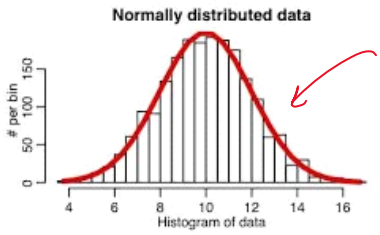
  - In a QQ plot, the quantiles of the two sets of data are plotted against each other. The quantiles of one set of data are plotted on the x-axis, while the quantiles of the other set of data are plotted on the y-axis. If the two sets of data have the same distribution, the points on the QQ plot will fall on a straight line. If the two sets of data do not have the same distribution, the points will deviate from the straight line.

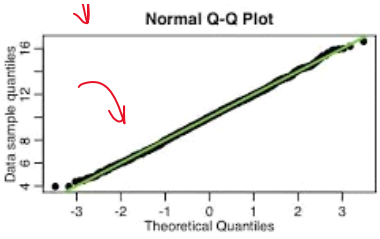
- Python example
- https://www.statsmodels.org/dev/generated/statsmodels.graphics.gofplots.qqplot.html

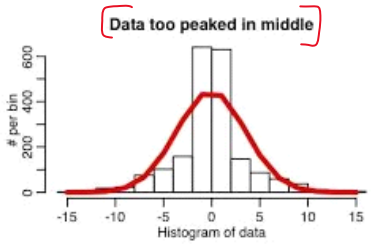
- How to interpret QQ plots

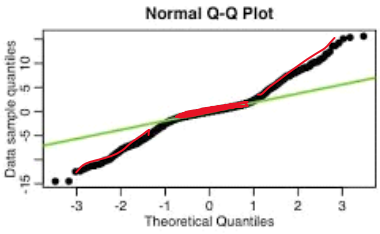


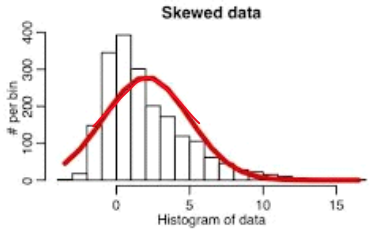
1.

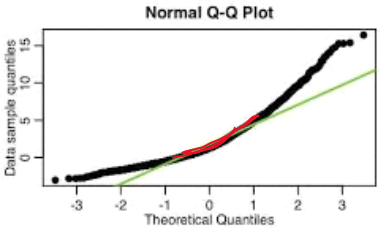

Normally distributed data
per bin
Histogram of data

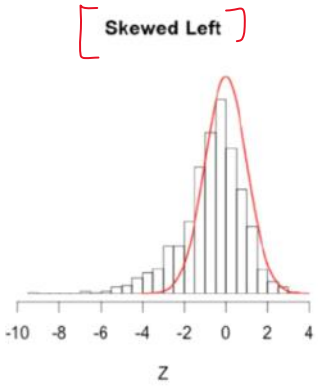

Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles

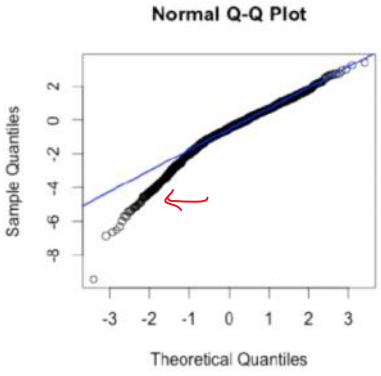

Data too peaked in middle
per bin
Histogram of data


Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles

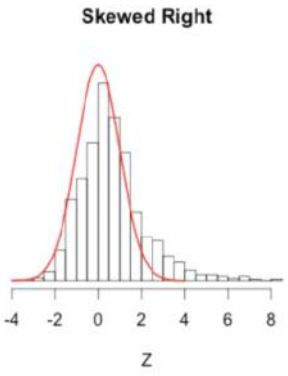

Skewed data
per bin
Histogram of data

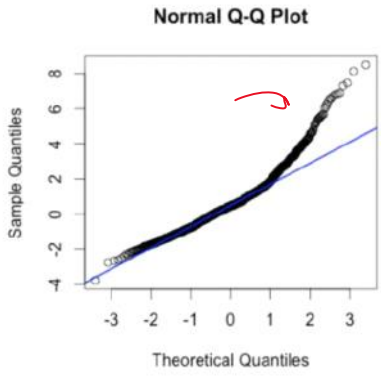

Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles


Skewed Left
Z

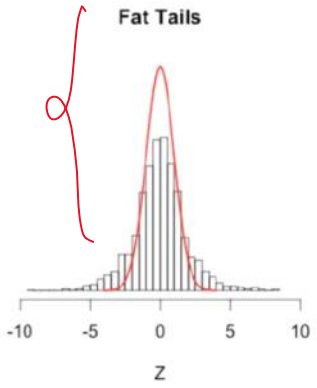

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

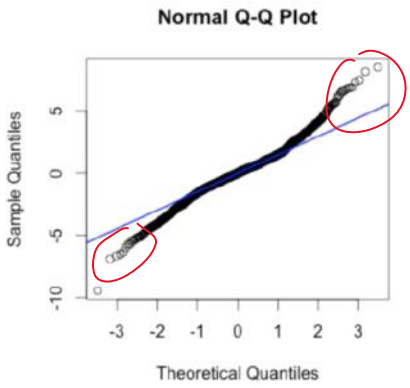
1.

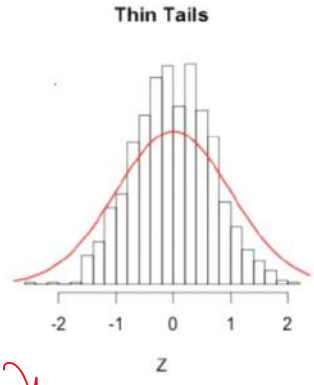

Skewed Right
Z

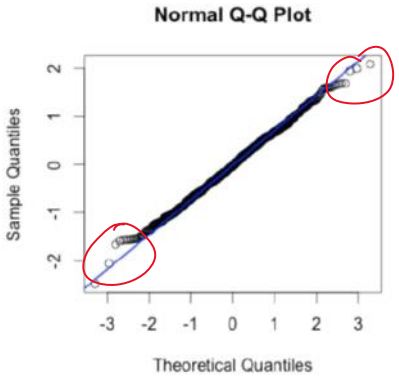

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

1.

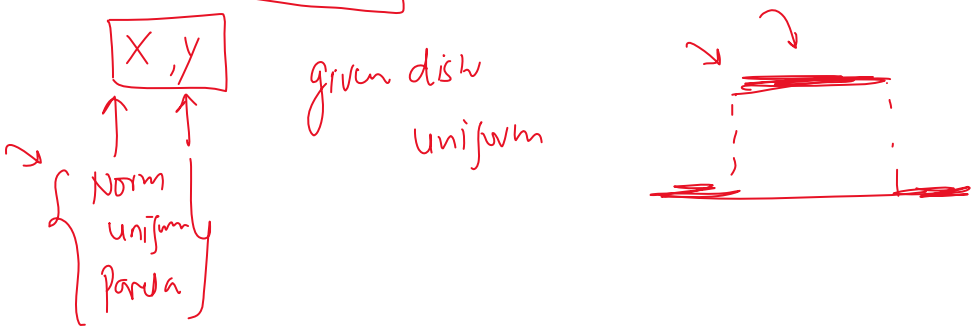








- Does QQ plot only detect normal distribution?



Uniform Distribution

23 March 2023 13:19

- What is Uniform Distribution and it's types

In probability theory and statistics, a uniform distribution is a probability distribution where all outcomes are equally likely within a given range. This means that if you were to select a random value from this range, any value would be as likely as any other value.

Types

Uniform (X)

Discrete — Continuous

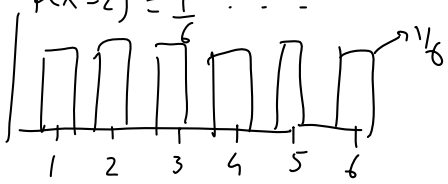
Denoted as $X \sim U(a,b)$ → parameters (a ↔ b; a: lower, b: high)

a=1 b=6

dice roll → {1, 2, 3, 4, 5, 6}

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

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

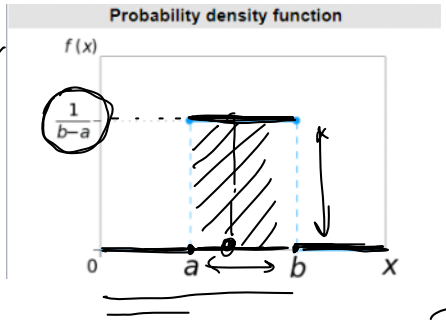


- Examples

a. The height of a person randomly selected from a group of individuals whose heights range from 5'6" to 6'0" would follow a continuous uniform distribution.
b. The time it takes for a machine to produce a product, where the production time ranges from 5 to 10 minutes, would follow a continuous uniform distribution.
c. The distance that a randomly selected car travels on a tank of gas, where the distance ranges from 300 to 400 miles, would follow a continuous uniform distribution.
d. The weight of a randomly selected apple from a basket of apples that weighs between 100 and 200 grams, would follow a continuous uniform distribution.

- PDF CDF and Graphs

PDF

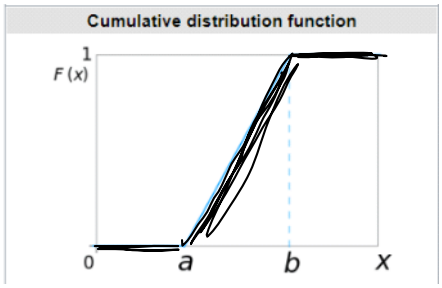


$\frac{1}{b-a}$ ← $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for otherwise} \end{cases}$

$x < a$ or $x > b$

$b-a$

CDF

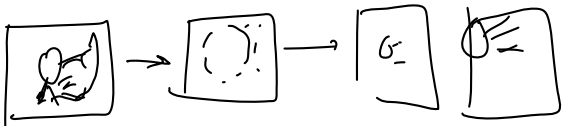


https://en.wikipedia.org/wiki/Continuous_uniform_distribution

- Skewness → 0 → Symmetric → Normal

- Application in Machine learning and Data Science

a. **Random initialization** In many machine learning algorithms, such as neural networks and k-means clustering, the initial values of the parameters can have a significant impact on the final result. Uniform distribution is often used to randomly initialize the parameters, as it ensures that all values in the range have an equal probability of being selected

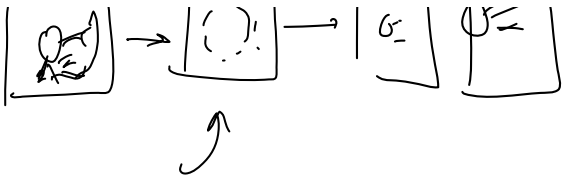


a. **Random initialization**: In many machine learning algorithms, such as neural networks and k-means clustering, the initial values of the parameters can have a significant impact on the final result. Uniform distribution is often used to randomly initialize the parameters, as it ensures that all values in the range have an equal probability of being selected.

b. **Sampling**: Uniform distribution can also be used for sampling. For example, if you have a dataset with an equal number of samples from each class, you can use uniform distribution to randomly select a subset of the data that is representative of all the classes.

c. **Data augmentation**: In some cases, you may want to artificially increase the size of your dataset by generating new examples that are similar to the original data. Uniform distribution can be used to generate new data points that are within a specified range of the original data.

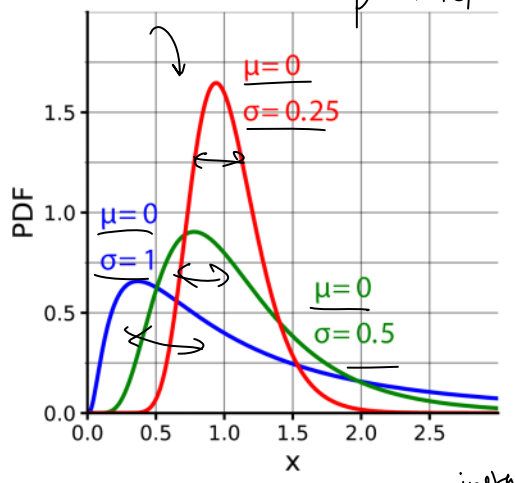
d. **Hyperparameter tuning**: Uniform distribution can also be used in hyperparameter tuning, where you need to search for the best combination of hyperparameters for a machine learning model. By defining a uniform prior distribution for each hyperparameter, you can sample from the distribution to explore the hyperparameter space.



Log Normal Distribution

23 March 2023 13:19

In probability theory and statistics, a lognormal distribution is a heavy tailed continuous probability distribution of a random variable whose logarithm is normally distributed.



Examples

- The length of comments posted in Internet discussion forums follows a log-normal distribution.
- Users' dwell time on online articles (jokes, news etc.) follows a log-normal distribution.
- The length of chess games tends to follow a log-normal distribution.
- In economics, there is evidence that the income of 97%–99% of the population is distributed log-normally.

Denoted as

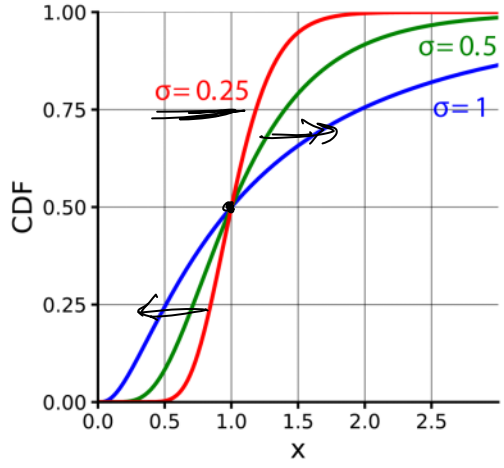
$X \sim \text{lognormal}(\mu, \sigma) \rightarrow \ln(x) \sim N(\mu', \sigma')$

PDF Equation

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

← similar Normal dist.

CDF



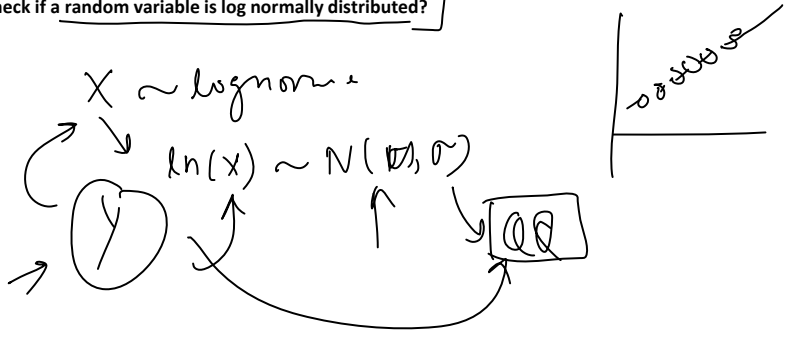
Skewness

Skewed

How to check if a random variable is log normally distributed?

Skewness

How to check if a random variable is log normally distributed?



Pareto Distribution

23 March 2023 13:19

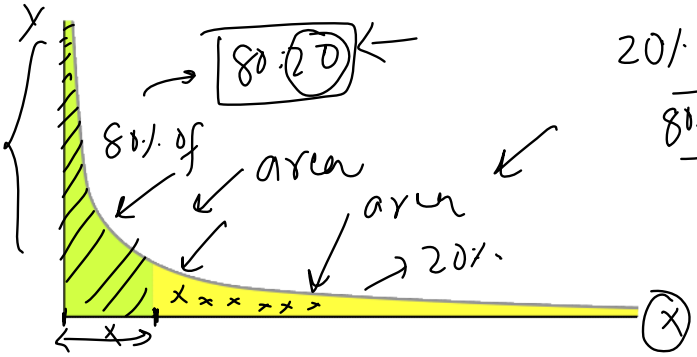
Pareto Distribution

The Pareto distribution is a type of probability distribution that is commonly used to model the distribution of wealth, income, and other quantities that exhibit a similar power-law behaviour

What is Power Law

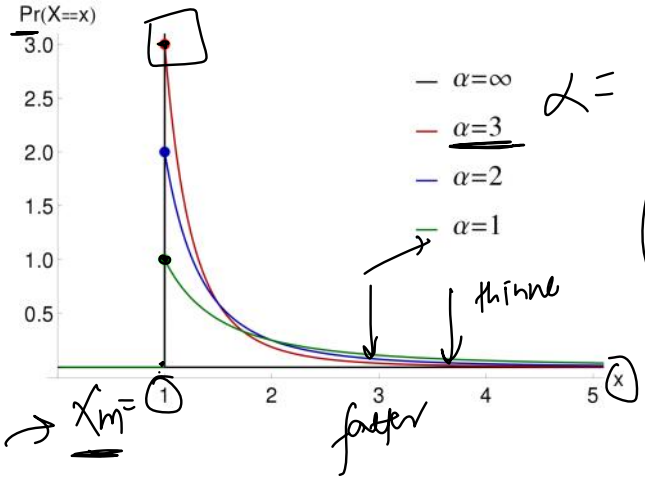
In mathematics, a power law is a functional relationship between two variables, where one variable is proportional to a power of the other. Specifically, if y and x are two variables related by a power law, then the relationship can be written as:

y = k * x^a



Vilfredo Pareto originally used this distribution to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. He also used it to describe distribution of income. This idea is sometimes expressed more simply as the Pareto principle or the "80-20 rule" which says that 20% of the population controls 80% of the wealth

Graph & Parameters

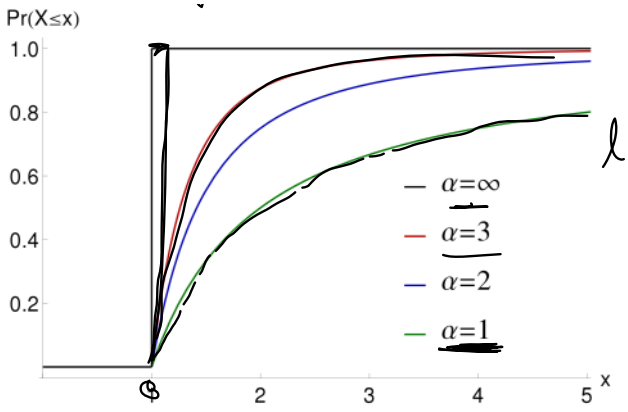


Examples

- The sizes of human settlements (few cities, many hamlets/villages)
- File size distribution of Internet traffic which uses the TCP protocol (many smaller files, few larger ones)

CDF

Pr(X<x)



log (y)

pareto

x log

Skewness → skewed

log-log plot → log

log x

[How to detect if a distribution is Pareto Distribution?]

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

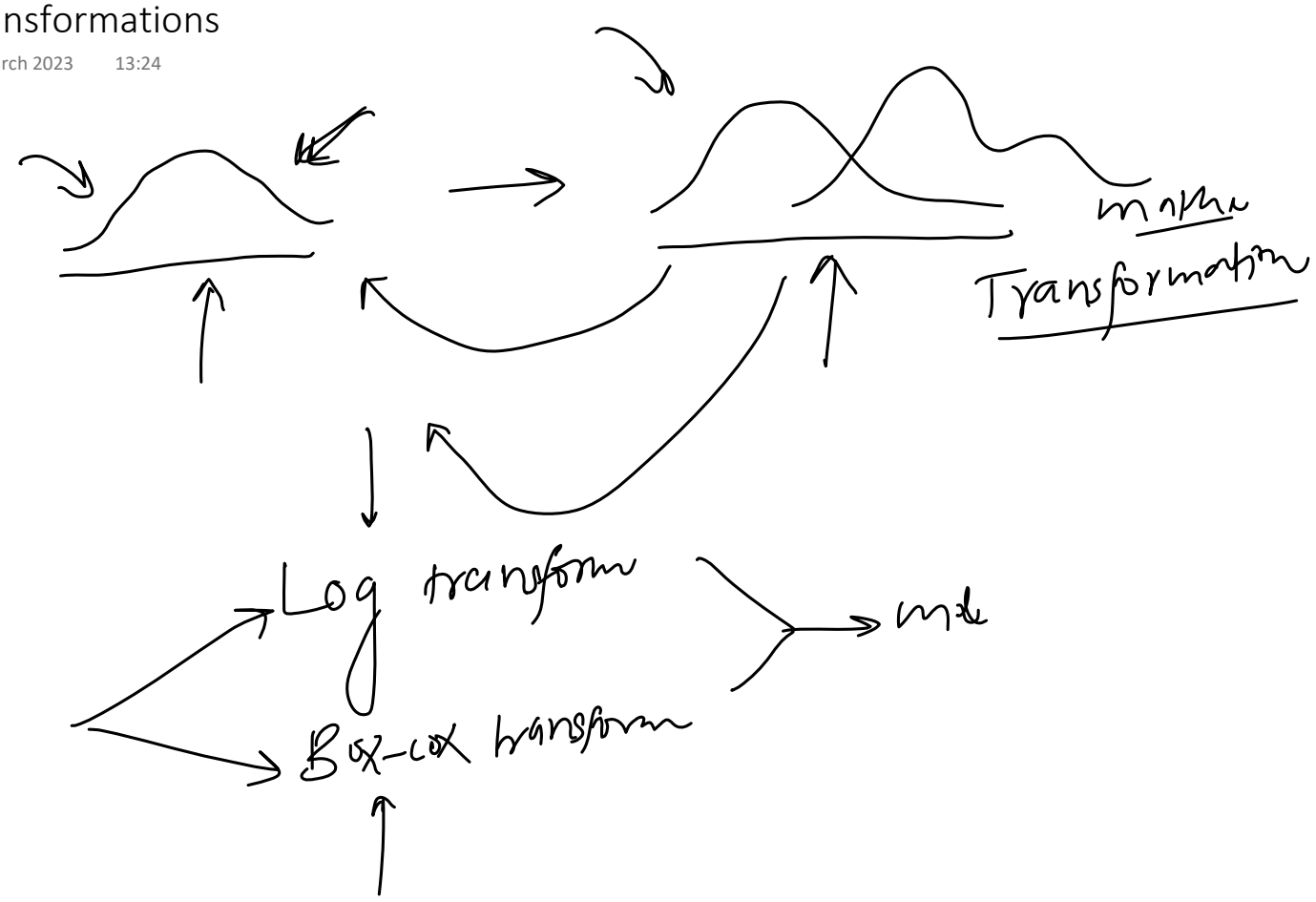
QQ plot

x, y

pareto

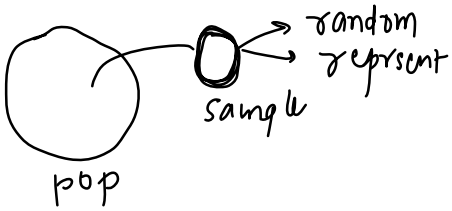
Transformations

23 March 2023 13:24



Some Terms

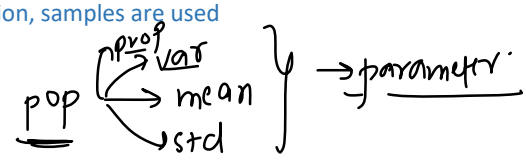
30 March 2023 07:09



Population Vs Sample

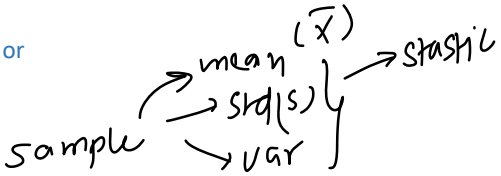
Population: A population is the entire group or set of individuals, objects, or events that a researcher wants to study or draw conclusions about. It can be people, animals, plants, or even inanimate objects, depending on the context of the study. The population usually represents the complete set of possible data points or observations.

Sample: A sample is a subset of the population that is selected for study. It is a smaller group that is intended to be representative of the larger population. Researchers collect data from the sample and use it to make inferences about the population as a whole. Since it is often impractical or impossible to collect data from every member of a population, samples are used as an efficient and cost-effective way to gather information.



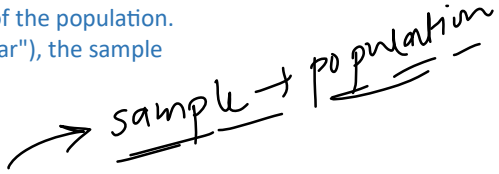
Parameter Vs ~~Estima~~te

Parameter: A parameter is a numerical value that describes a characteristic of a population. Parameters are usually denoted using Greek letters, such as μ (mu) for the population mean or σ (sigma) for the population standard deviation. Since it is often difficult or impossible to obtain data from an entire population, parameters are usually unknown and must be estimated based on available sample data.



$\bar{x} \rightarrow \mu$

Statistic: A statistic is a numerical value that describes a characteristic of a sample, which is a subset of the population. By using statistics calculated from a representative sample, researchers can make inferences about the unknown respective parameter of the population. Common statistics include the sample mean (denoted by x̄, pronounced "x-bar"), the sample median, and the sample standard deviation (denoted by s).



Inferential Statistics

Inferential statistics is a branch of statistics that focuses on making predictions, estimations, or generalizations about a larger population based on a sample of data taken from that population. It involves the use of probability theory to make inferences and draw conclusions about the characteristics of a population by analysing a smaller subset or sample.

The key idea behind inferential statistics is that it is often impractical or impossible to collect data from every member of a population, so instead, we use a representative sample to make inferences about the entire group. Inferential statistical techniques include hypothesis testing, confidence intervals, and regression analysis, among others.

These methods help researchers answer questions like:

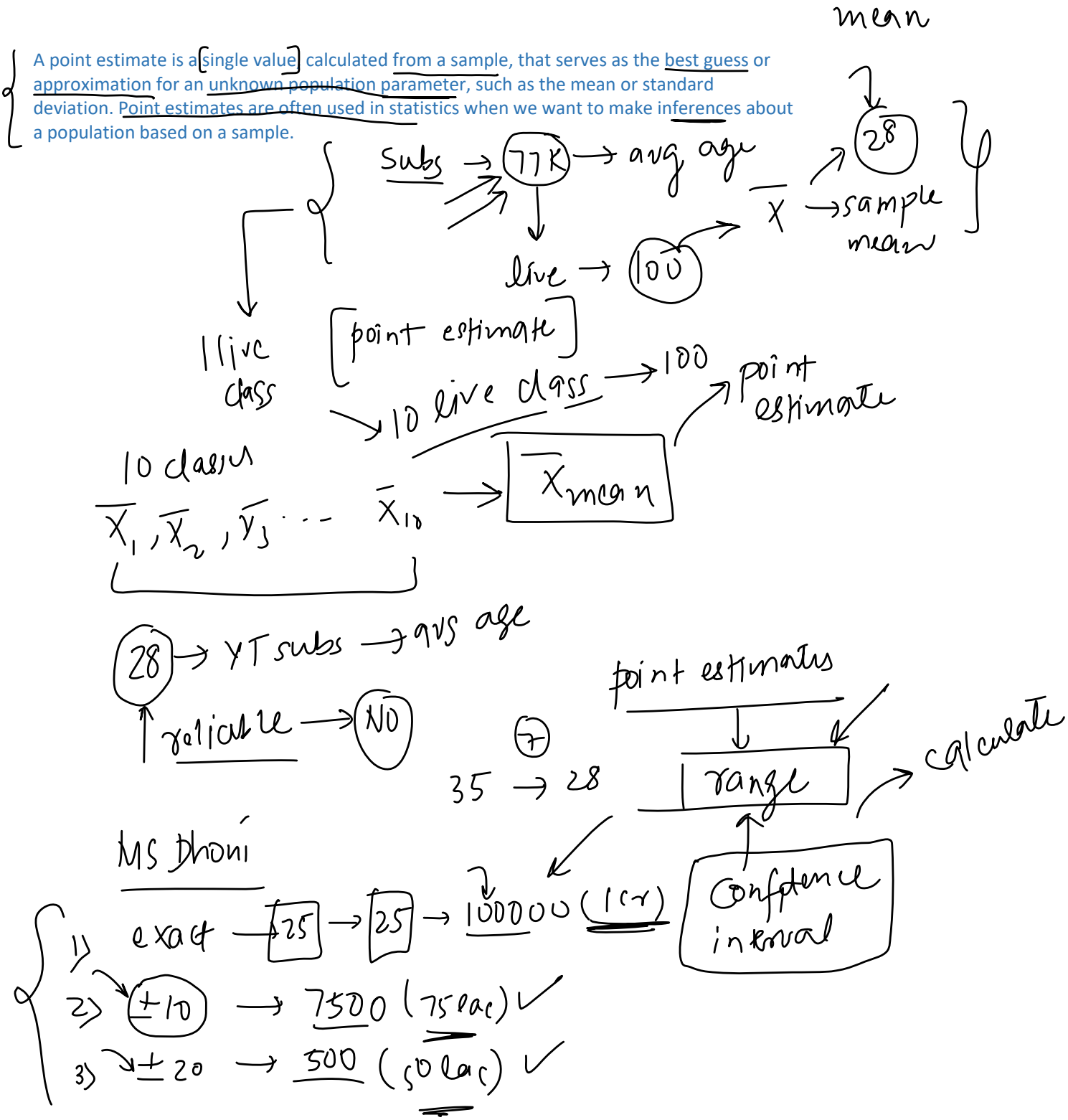
a. Is there a significant difference between two groups?
b. Can we predict the outcome of a variable based on the values of other variables?
c. What is the relationship between two or more variables?

Inferential statistics are widely used in various fields, such as economics, social sciences, medicine, and natural sciences, to make informed decisions and guide policy based on limited data.

Point Estimate

30 March 2023 07:19

A point estimate is a single value calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.



Confidence Interval

30 March 2023 07:18

Confidence interval, in simple words, is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

Confidence level, usually expressed as a percentage like 95%, indicates how sure we are that the true value lies within the interval.

Confidence Interval = [Point Estimate] ± [Margin of Error]

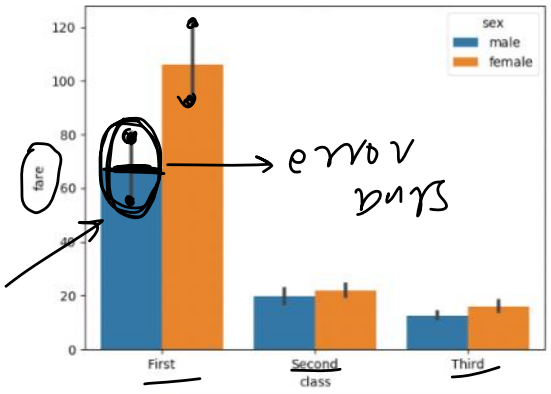
Ways to calculate CI:

Confidence Interval is created for Parameters and not statistics. Statistics help us get the confidence interval for a parameter.

Examples of CT usage

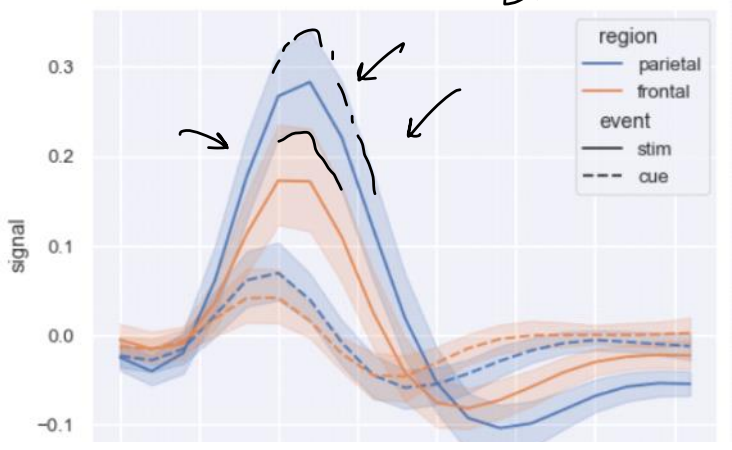
seaborn

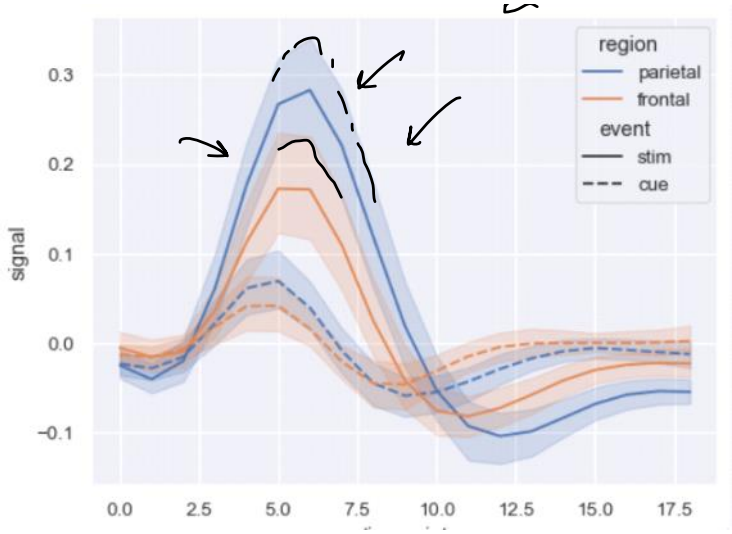
→ bar plot



error bars

(CI)



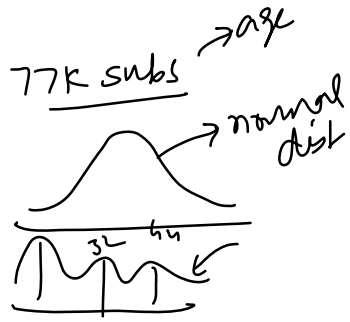

region
parietal
frontal
event
stim
cue
signal
0.3
0.2
0.1
0.0
−0.1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

Confidence Interval (Sigma Known)

30 March 2023 07:13

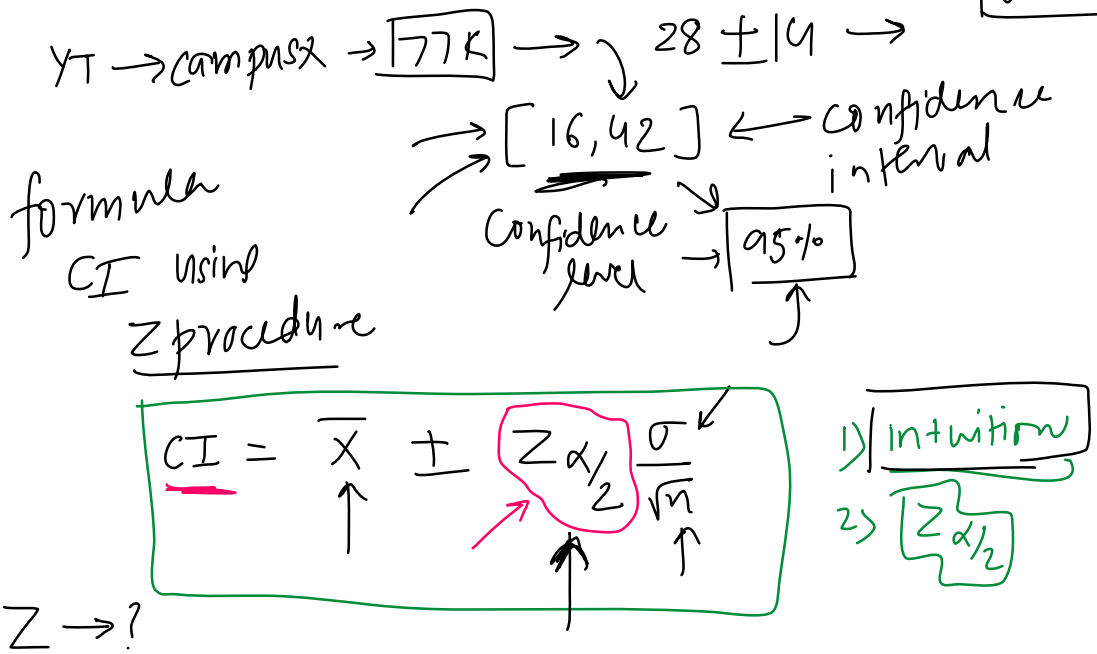
Assumptions

1. Random sampling: The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. Known population standard deviation: The population standard deviation (σ) must be known or accurately estimated. In practice, the population standard deviation is often unknown, and the sample standard deviation (s) is used as an estimate. However, if the sample size is large enough, the sample standard deviation can provide a reasonably accurate approximation.
3. Normal distribution or large sample size: The Z-procedure assumes that the underlying population is normally distributed. However, if the population distribution is not normal, the Central Limit Theorem can be applied when the sample size is large (usually, sample size $n \geq 30$ is considered large enough). According to the Central Limit Theorem, the sampling distribution of the sample mean will approach a normal distribution as the sample size increases, regardless of the shape of the population distribution.



A (1 - alpha)*100% Confidence Interval for mu:

$\sigma = 15$



$$CI = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

1) intuition
2) $Z_{\alpha/2}$

$Z \rightarrow ?$

$(1-\alpha) \rightarrow$ confidence level

$(1-\alpha) \rightarrow 95\%$

$\sigma \rightarrow$ std pop

$n \rightarrow$ sample size $\rightarrow 100$

Intuition

Point estimate $(\bar{X}) \rightarrow$ CLT

point estimate $(\bar{x})$ → CI

10 live class → 50 → avg class

$[\bar{x}_1, \bar{x}_2 \ldots \bar{x}_{10}]$ → sampling dist of sample mean

normally dist → (z) → N(0,1)

$\bar{x}$ μ $\sigma/\sqrt{n}$

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

95% = (1−α) sure

95% (z)

1−α prob → 95% z lie

α/2 α/2

$-z_{\alpha/2}$ $z_{\alpha/2}$

$-z_{2.5} \longleftrightarrow z_{2.5}$ → 95%

$\alpha = 5$

$CI = \bar{x} \pm (z_{\alpha/2}) \frac{\sigma}{\sqrt{n}}$ (95)

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

(μ → CI)

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

fixed

95%

$\bar{X}$ → sample

$1-\alpha$ 0.95 $\alpha = 0.50$

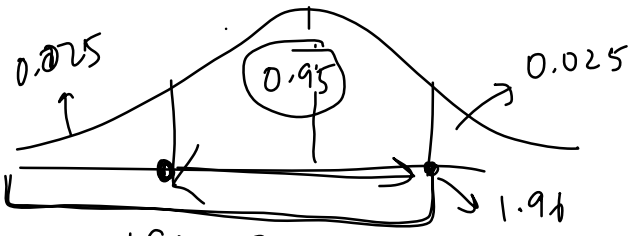
$\frac{\alpha}{2}$ 0.250

confidence $(1-\alpha)$ → 95%

CI $\mu = \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$\mu = \bar{X} \pm z_{0.250}\frac{\sigma}{\sqrt{n}}$ (z)

950
025
975



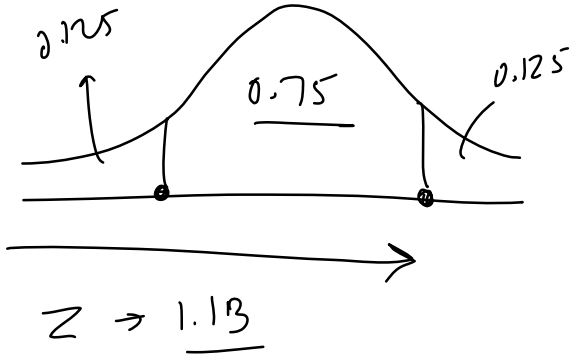
range → confidence interval

$\mu = \bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ → CI with a confidence of 95%.

50% 75% 99%

$z_{\alpha/2} \rightarrow 1.96$

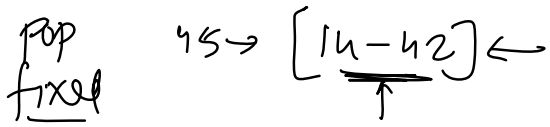
0.87 $z_{\alpha/2}$



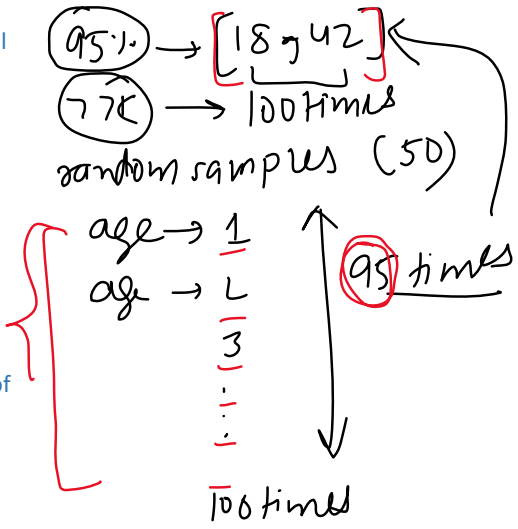
$z \rightarrow 1.15$

Interpreting Confidence Interval

30 March 2023 08:33

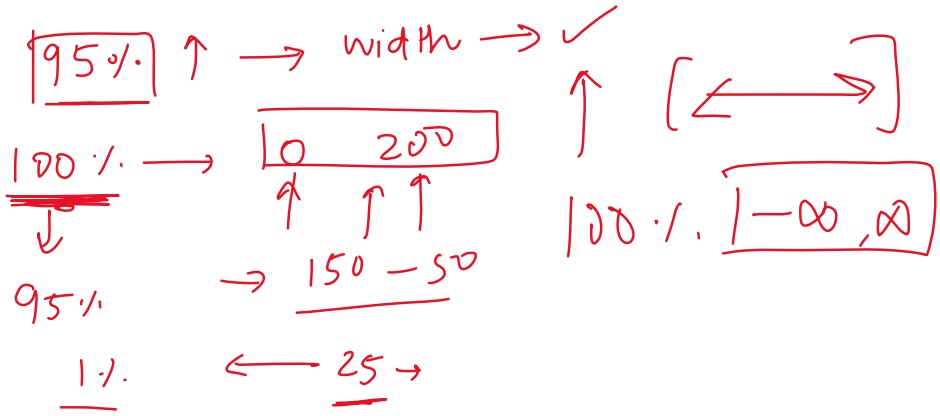


A confidence interval is a range of values within which a population parameter, such as the population mean, is estimated to lie with a certain level of confidence. The confidence interval provides an indication of the precision and uncertainty associated with the estimate. To interpret the confidence interval values, consider the following points:



1. **Confidence level**: The confidence level (commonly set at 90%, 95%, or 99%) represents the probability that the confidence interval will contain the true population parameter if the sampling and estimation process were repeated multiple times. For example, a 95% confidence interval means that if you were to draw 100 different samples from the population and calculate the confidence interval for each, approximately 95 of those intervals would contain the true population parameter.
2. **Interval range**: The width of the confidence interval gives an indication of the precision of the estimate. A narrower confidence interval suggests a more precise estimate of the population parameter, while a wider interval indicates greater uncertainty. The width of the interval depends on the sample size, variability in the data, and the desired level of confidence.
3. **Interpretation**: To interpret the confidence interval values, you can say that you are "X% confident that the true population parameter lies within the range (lower limit, upper limit)." Keep in mind that this statement is about the interval, not the specific point estimate, and it refers to the confidence level you chose when constructing the interval.

What is the trade-off



Factors Affecting Margin of Error

30 March 2023 07:15

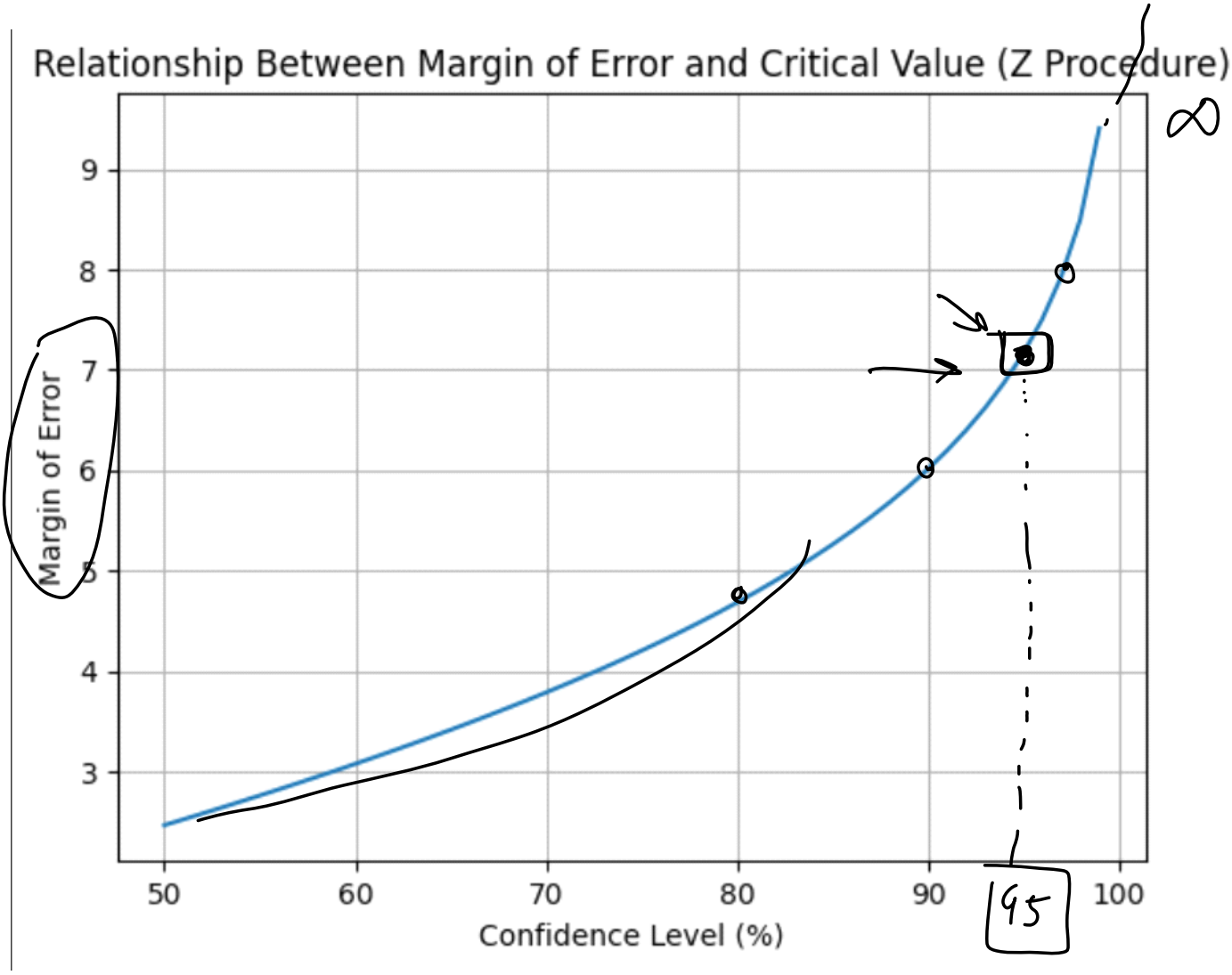
1. Confidence Level (1-alpha)
2. Sample Size
3. Population Standard Deviation

$$CI = \text{point estimate} \pm \text{margin of error}$$

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

$\bar{x}$: sample mean ①; $z_{\alpha/2}$: critical value ①; σ: pop std ②; $\sqrt{n}$: sample size ③

$$\frac{\text{upper} - \text{lower}}{2} \rightarrow \text{margin}$$



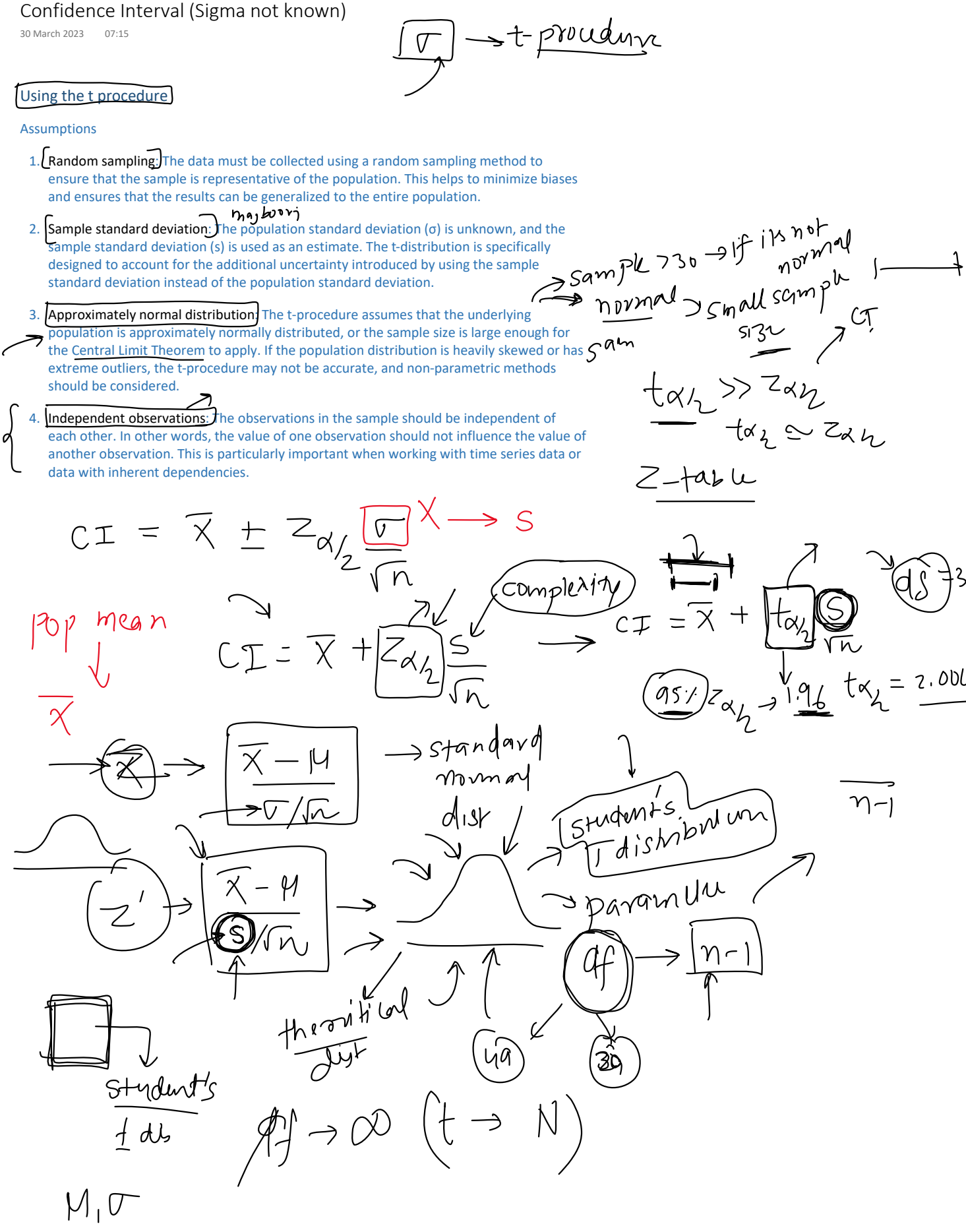
Confidence Interval (Sigma not known)

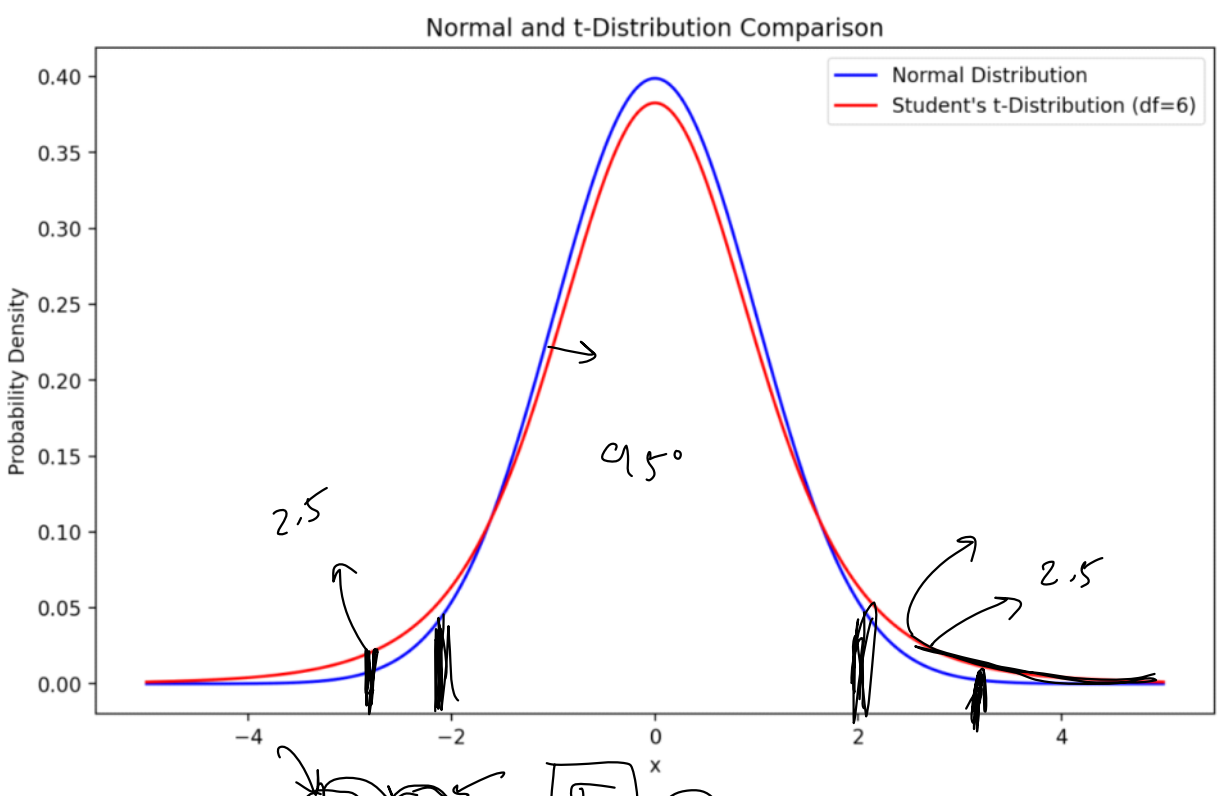
30 March 2023 07:15

Using the t procedure

Assumptions

1. Random sampling: The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. Sample standard deviation: The population standard deviation (σ) is unknown, and the sample standard deviation (s) is used as an estimate. The t-distribution is specifically designed to account for the additional uncertainty introduced by using the sample standard deviation instead of the population standard deviation.
3. Approximately normal distribution: The t-procedure assumes that the underlying population is approximately normally distributed, or the sample size is large enough for the Central Limit Theorem to apply. If the population distribution is heavily skewed or has extreme outliers, the t-procedure may not be accurate, and non-parametric methods should be considered.
4. Independent observations: The observations in the sample should be independent of each other. In other words, the value of one observation should not influence the value of another observation. This is particularly important when working with time series data or data with inherent dependencies.





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

σ

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

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

1 Nitish 2 CLT

2 Ankit

count

$\bar{x}_1$ s_1

$\bar{x}_2$ s_2

s_3

s_{10}

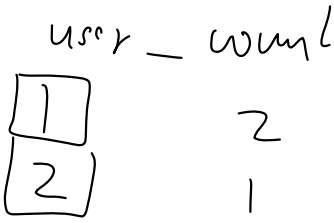
s_{avg}

user_id | name | event | dim

→ Nitish like

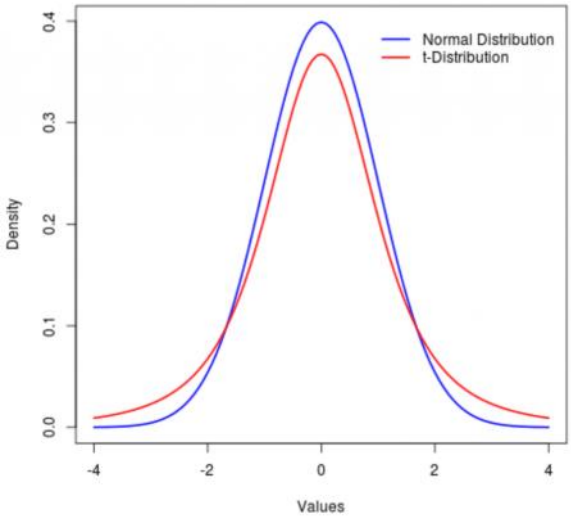
NIL comm

2 Ankit like


user_wml
1
2
2
1

Student's T Distribution

30 March 2023 07:16



Student's t-distribution, or simply the t-distribution, is a probability distribution that arises when estimating the mean of a normally distributed population when the sample size is small and the population standard deviation is unknown. It was introduced by William Sealy Gosset, who published under the pseudonym "Student."

The t-distribution is similar to the normal distribution (also known as the Gaussian distribution or the bell curve) but has heavier tails. The shape of the t-distribution is determined by the degrees of freedom, which is closely related to the sample size (degrees of freedom = sample size - 1). As the degrees of freedom increase (i.e., as the sample size increases), the t-distribution approaches the normal distribution.

In hypothesis testing and confidence interval estimation, the t-distribution is used in place of the normal distribution when the sample size is small (usually less than 30) and the population standard deviation is unknown. The t-distribution accounts for the additional uncertainty that arises from estimating the population standard deviation using the sample standard deviation.

To use the t-distribution in practice, you look up critical t-values from a t-distribution table, which provides values corresponding to specific degrees of freedom and confidence levels (e.g., 95% confidence). These critical t-values are then used to calculate confidence intervals or perform hypothesis tests.

Titanic Case Study

31 March 2023 18:00

Pop → 1360

μ → ×

σ → ×

CLT → 10 times → 30 size

95% confidence level

inference

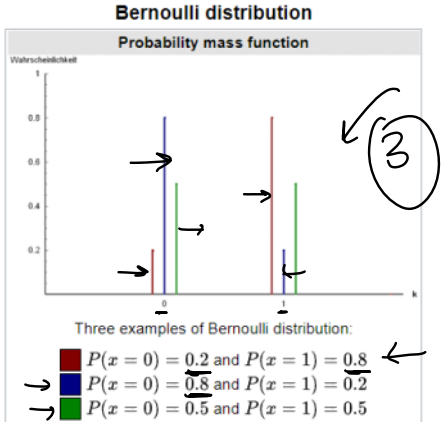
Bernoulli Distribution

27 March 2023 16:06

Bernoulli distribution is a probability distribution that models a binary outcome, where the outcome can be either success (represented by the value 1) or failure (represented by the value 0). The Bernoulli distribution is named after the Swiss mathematician Jacob Bernoulli, who first introduced it in the late 1600s.

The Bernoulli distribution is characterized by a single parameter, which is the probability of success, denoted by p. The probability mass function (PMF) of the Bernoulli distribution is:

$$\text{pmf} = P(X=x) = p^{x}(1-p)^{1-x}$$



The Bernoulli distribution is commonly used in machine learning for modelling binary outcomes, such as whether a customer will make a purchase or not, whether an email is spam or not, or whether a patient will have a certain disease or not.

Binomial Distribution

27 March 2023 16:36

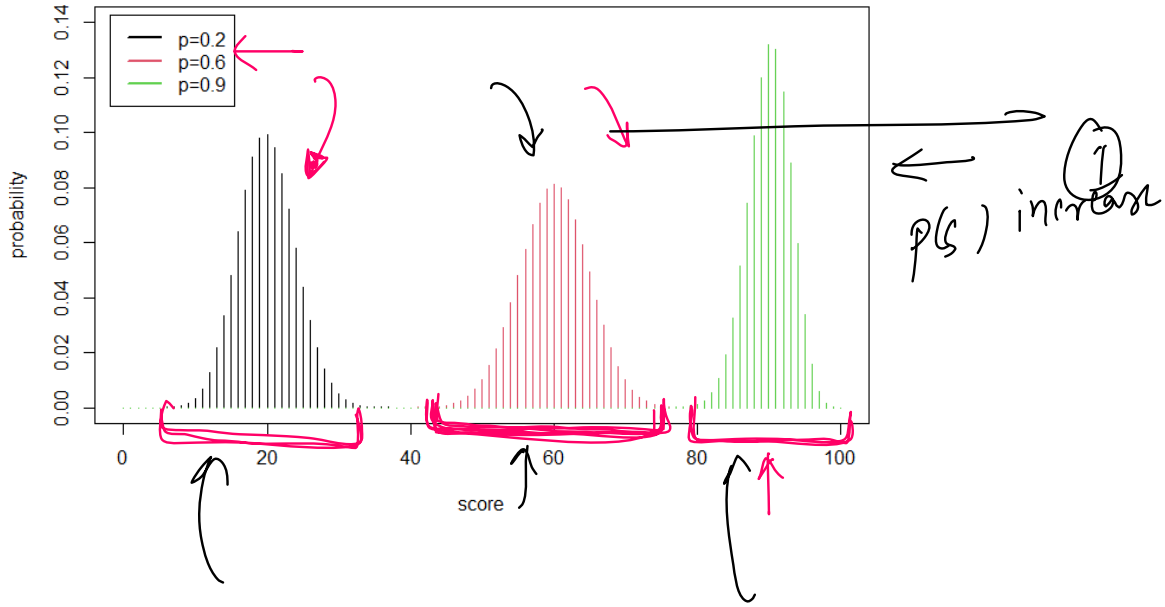
Binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials with two possible outcomes (often called "success" and "failure"), where the probability of success is constant for each trial. The binomial distribution is characterized by two parameters: the number of trials n and the probability of success p.

The Probability of anyone watching this lecture in the future and then liking it is 0.5. What is the probability that:

1. No-one out of 3 people will like it
1. 1 out of 3 people will like it
1. 2 out of 3 people will like it
1. 3 out of 3 people will like it

PDF Formula:

Graph of PDF:



Criteria:

Criteria:

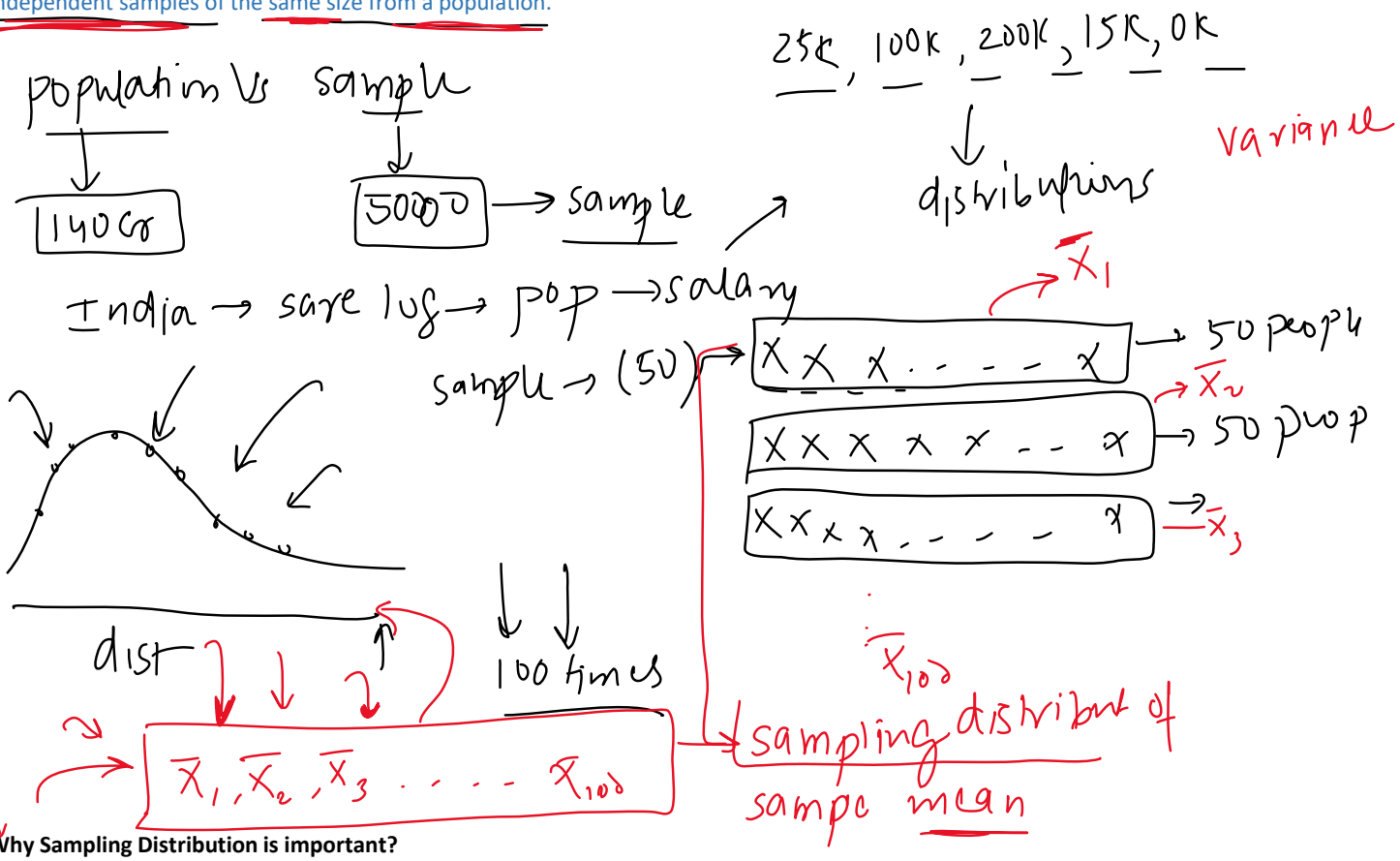
1. The process consists of n trials
2. Only 2 exclusive outcomes are possible, a success and a failure.
3. P(success) = p and P(failure) = 1-p and it is fixed from trial to trial
4. The trials are independent.

1. **Binary classification problems**: In binary classification problems, we often model the probability of an event happening as a binomial distribution. For example, in a spam detection system, we may model the probability of an email being spam or not spam using a binomial distribution.
2. **Hypothesis testing**: In statistical hypothesis testing, we use the binomial distribution to calculate the probability of observing a certain number of successes in a given number of trials, assuming a null hypothesis is true. This can be used to make decisions about whether a certain hypothesis is supported by the data or not.
3. **Logistic regression**: Logistic regression is a popular machine learning algorithm used for classification problems. It models the probability of an event happening as a logistic function of the input variables. Since the logistic function can be viewed as a transformation of a linear combination of inputs, the output of logistic regression can be thought of as a binomial distribution.
4. **A/B testing**: A/B testing is a common technique used to compare two different versions of a product, web page, or marketing campaign. In A/B testing, we randomly assign individuals to one of two groups and compare the outcomes of interest between the groups. Since the outcomes are often binary (e.g., click-through rate or conversion rate), the binomial distribution can be used to model the distribution of outcomes and test for differences between the groups.

Sampling Distribution

27 March 2023 17:10

Sampling distribution is a probability distribution that describes the statistical properties of a sample statistic (such as the sample mean or sample proportion) computed from multiple independent samples of the same size from a population.



Why Sampling Distribution is important?

Sampling distribution is important in statistics and machine learning because it allows us to estimate the variability of a sample statistic, which is useful for making inferences about the population. By analysing the properties of the sampling distribution, we can compute confidence intervals, perform hypothesis tests, and make predictions about the population based on the sample data.

central limit theorem

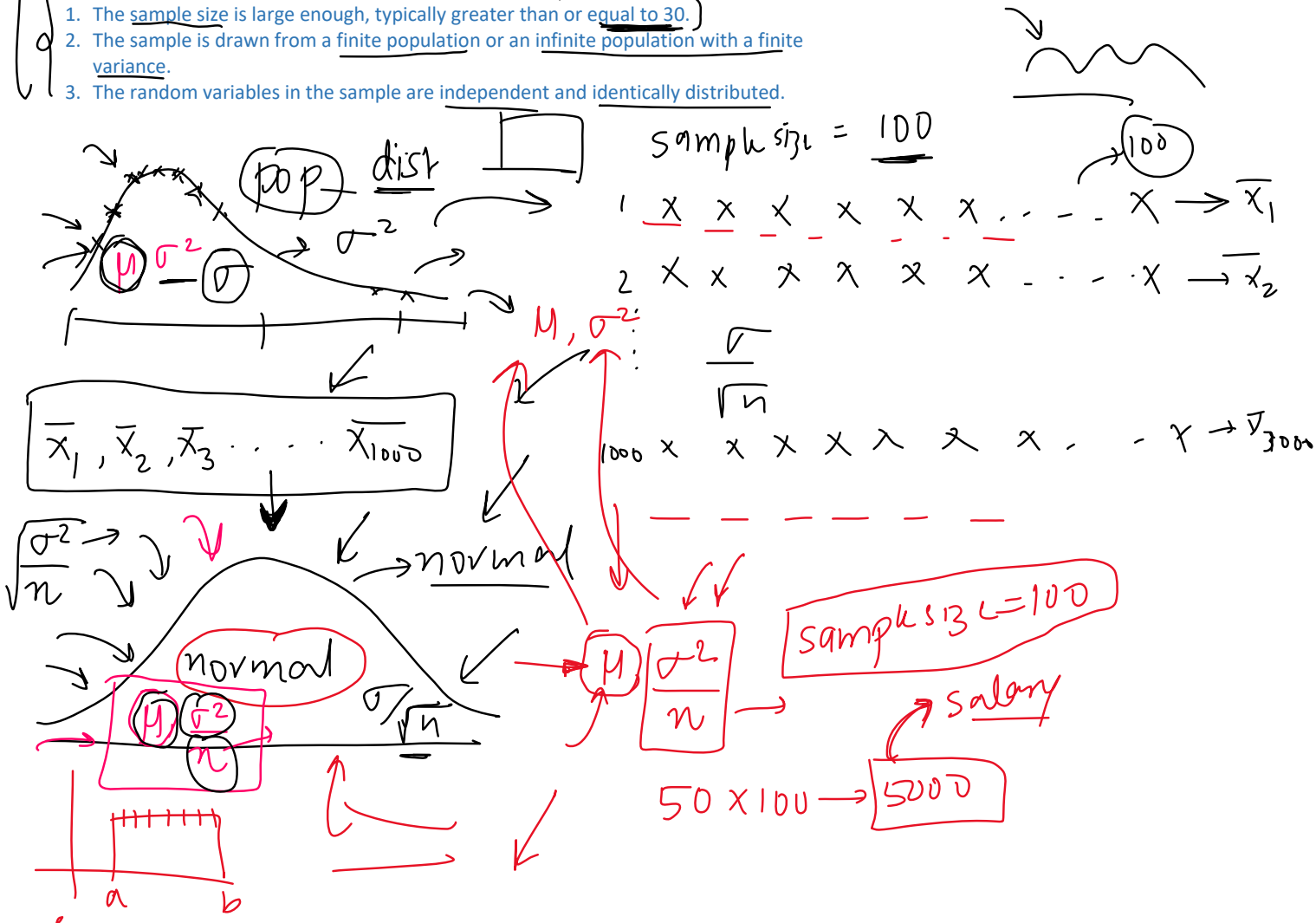
Central Limit Theorem

27 March 2023 17:10

The Central Limit Theorem (CLT) states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of the variables.

The conditions required for the CLT to hold are:

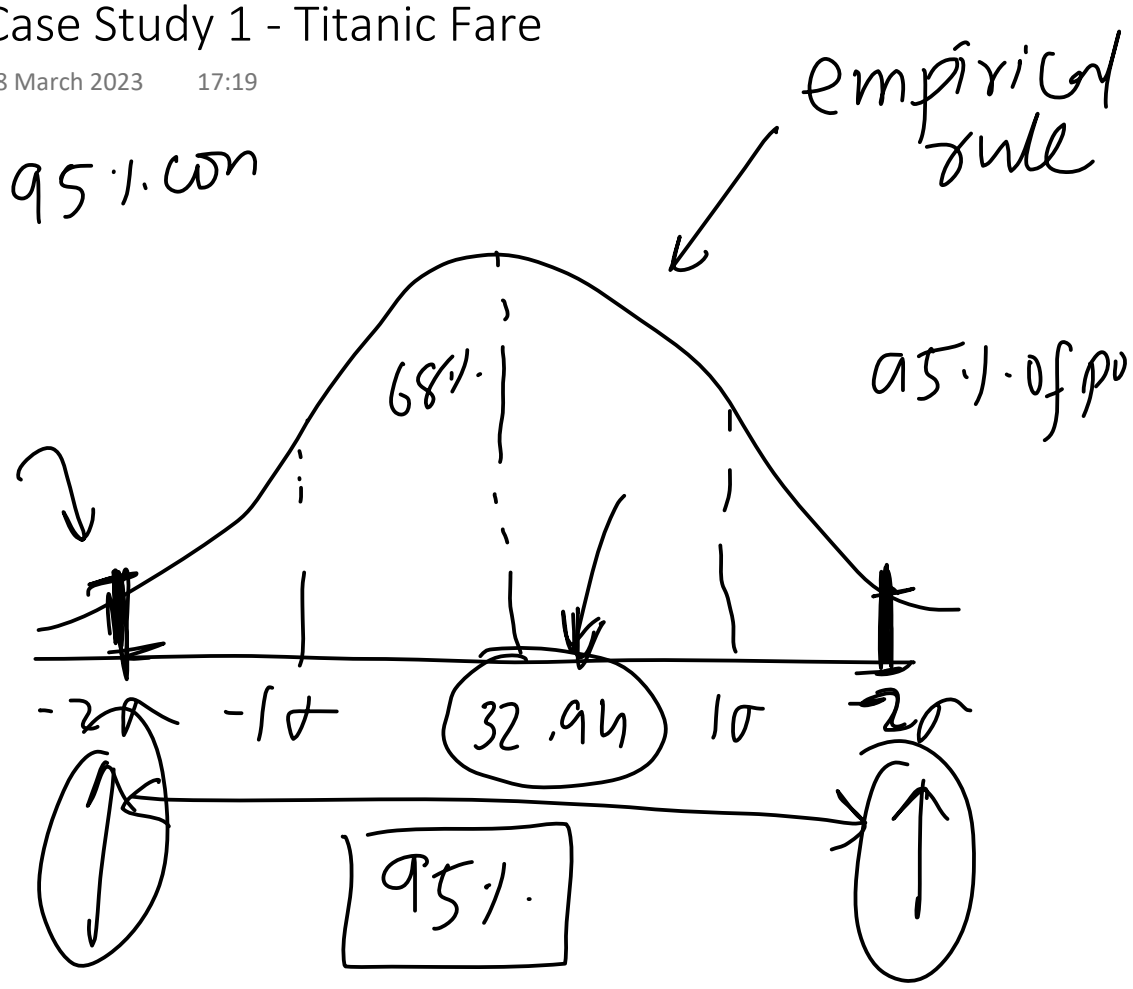
1. The sample size is large enough, typically greater than or equal to 30.
2. The sample is drawn from a finite population or an infinite population with a finite variance.
3. The random variables in the sample are independent and identically distributed.



The CLT is important in statistics and machine learning because it allows us to make probabilistic inferences about a population based on a sample of data. For example, we can use the CLT to construct confidence intervals, perform hypothesis tests, and make predictions about the population mean based on the sample data. The CLT also provides a theoretical justification for many commonly used statistical techniques, such as t-tests, ANOVA, and linear regression.

Case Study 1 - Titanic Fare

28 March 2023 17:19



Case Study - What is the average income of Indians

28 March 2023 15:49

Step-by-step process:

1. Collect multiple random samples of salaries from a representative group of Indians. Each sample should be large enough (usually, n > 30) to ensure the CLT holds. Make sure the samples are representative and unbiased to avoid skewed results.
2. Calculate the sample mean (average salary) and sample standard deviation for each sample.
3. Calculate the average of the sample means. This value will be your best estimate of the population mean (average salary of all Indians).
4. Calculate the standard error of the sample means, which is the standard deviation of the sample means divided by the square root of the number of samples.
5. Calculate the confidence interval around the average of the sample means to get a range within which the true population mean likely falls. For a 95% confidence interval:

   lower_limit = average_sample_means - 1.96 * standard_error
   upper_limit = average_sample_means + 1.96 * standard_error

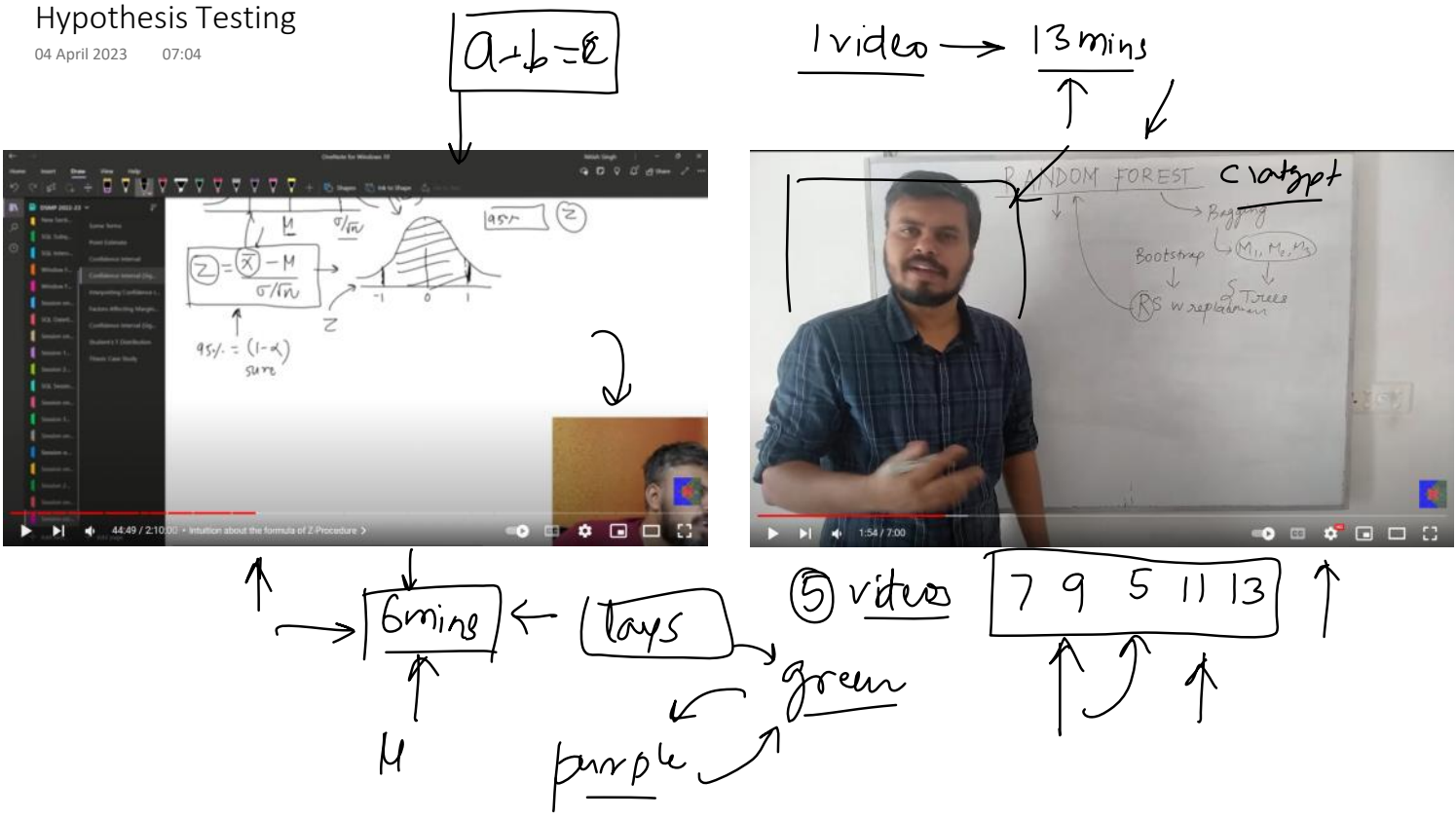
6. Report the estimated average salary and the confidence interval.

Python code

Remember that the validity of your results depends on the quality of your data and the representativeness of your samples. To obtain accurate results, it's crucial to ensure that your samples are unbiased and representative.

Hypothesis Testing

04 April 2023 07:04



A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.

Null and Alternate Hypothesis

04 April 2023 07:09

1. **Null hypothesis (H0)**:

In simple terms, the null hypothesis is a statement that assumes there is no significant effect or relationship between the variables being studied. It serves as the starting point for hypothesis testing and represents the status quo or the assumption of no effect until proven otherwise. The purpose of hypothesis testing is to gather evidence (data) to either reject or fail to reject the null hypothesis in favour of the alternative hypothesis, which claims there is a significant effect or relationship.

2. **Alternative hypothesis (H1 or Ha):**

The alternative hypothesis, is a statement that contradicts the null hypothesis and claims there is a significant effect or relationship between the variables being studied. It represents the research hypothesis or the claim that the researcher wants to support through statistical analysis.

Important Points

- How to decide what will be Null hypothesis and what will be Alternate Hypothesis(Typically the Null hypothesis says nothing new is happening)
- We try to gather evidence to reject the null hypothesis
- It's important to note that failing to reject the null hypothesis doesn't necessarily mean that the null hypothesis is true; it just means that there isn't enough evidence to support the alternative hypothesis.

Hypothesis tests are similar to jury trials, in a sense. In a jury trial, H0 is similar to the not-guilty verdict, and Ha is the guilty verdict. You assume in a jury trial that the defendant isn't guilty unless the prosecution can show beyond a reasonable doubt that he or she is guilty. If the jury says the evidence is beyond a reasonable doubt, they reject H0, not guilty, in favour of Ha , guilty.

Steps involved in Hypothesis Testing

04 April 2023 13:25

Rejection Region Approach → basic → p value approach

H_0: $\mu = 6$ mins → right

H_a: $\mu > 6$ mins

1. Formulate a Null and Alternate hypothesis
2. Select a significance level (This is the probability of rejecting the null hypothesis when it is actually true, usually set at 0.05 or 0.01)
3. Check assumptions (example distribution) → σ
4. Decide which test is appropriate(Z-test, T-test, Chi-square test, ANOVA)
5. State the relevant test statistic
6. Conduct the test
7. Reject or not reject the Null Hypothesis.
8. Interpret the result

0.05 or 0.01 → 5% → 1%

normally distributed, σ given ← t-test

Z-test

Z-test → Z-score

t-test → t-value

Performing a Z test Example 1

04 April 2023 07:15

$\mu = 50 \quad \sigma = 5$

$n = 30 \quad \bar{X} = 53$

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was 50 units per day with a known population standard deviation of 5 units. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day. The company wants to know if the new training program has significantly increased productivity.

1) $H_0: \mu = 50 \quad H_a: \mu > 50$

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

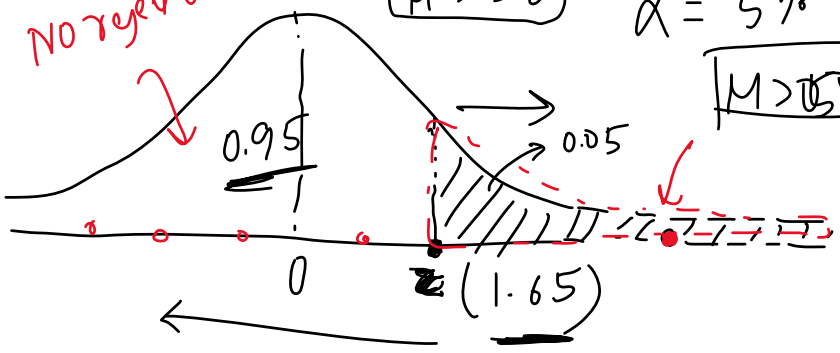
3) normality valid / pop std (σ) known → Z test

4) Z test

5) (Z)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{53 - 50}{5/\sqrt{30}} = \frac{3}{5/\sqrt{30}} = 3.28$$

Rejection



$\mu > 50$

$\alpha = 5\%$

$\mu > 50$

Reject the null hypothesis

> 50

Example 2

04 April 2023 16:06

Z test

Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a known population standard deviation of 4 grams.

$\mu = 50 \quad n = 40 \quad \bar{x} = 49 \quad \sigma = 4$

1) $H_0: \mu = 50 \quad H_a: \mu \neq 50$

2) $\alpha = 0.05$

3) Normality ✓ σ ✓ → Z test

4) Z test

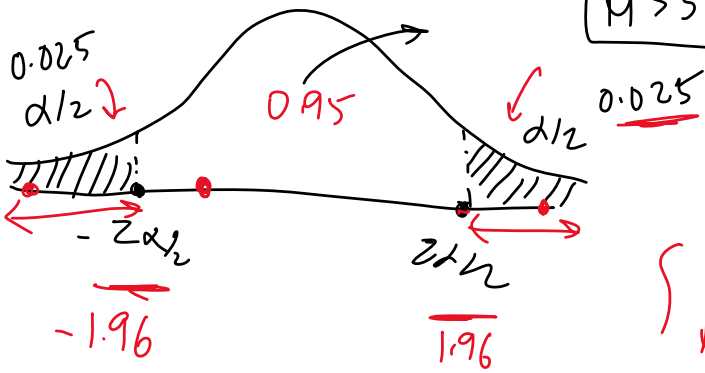
5) Z

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

$\alpha = 5\%$

$\mu > 50$ $\mu \neq 50$

can't reject the NULL hypothesis



$\mu \neq 50$

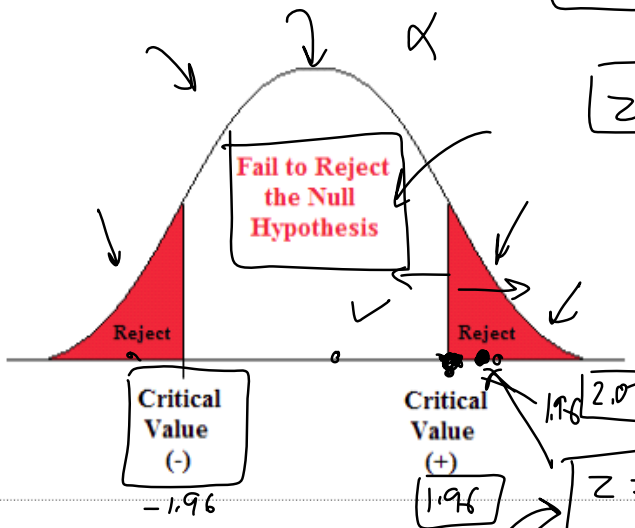
$\{\mu > 50, \mu < 50\}$

Rejection Region

04 April 2023 16:21

Significance level - denoted as α (alpha), is a predetermined threshold used in hypothesis testing to determine whether the null hypothesis should be rejected or not. It represents the probability of rejecting the null hypothesis when it is actually true, also known as Type 1 error.

The critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.



Problem with Rejection Region Approach

Type 1 vs Type 2 Error

04 April 2023 13:29

In hypothesis testing, there are two types of errors that can occur when making a decision about the null hypothesis: Type I error and Type II error.

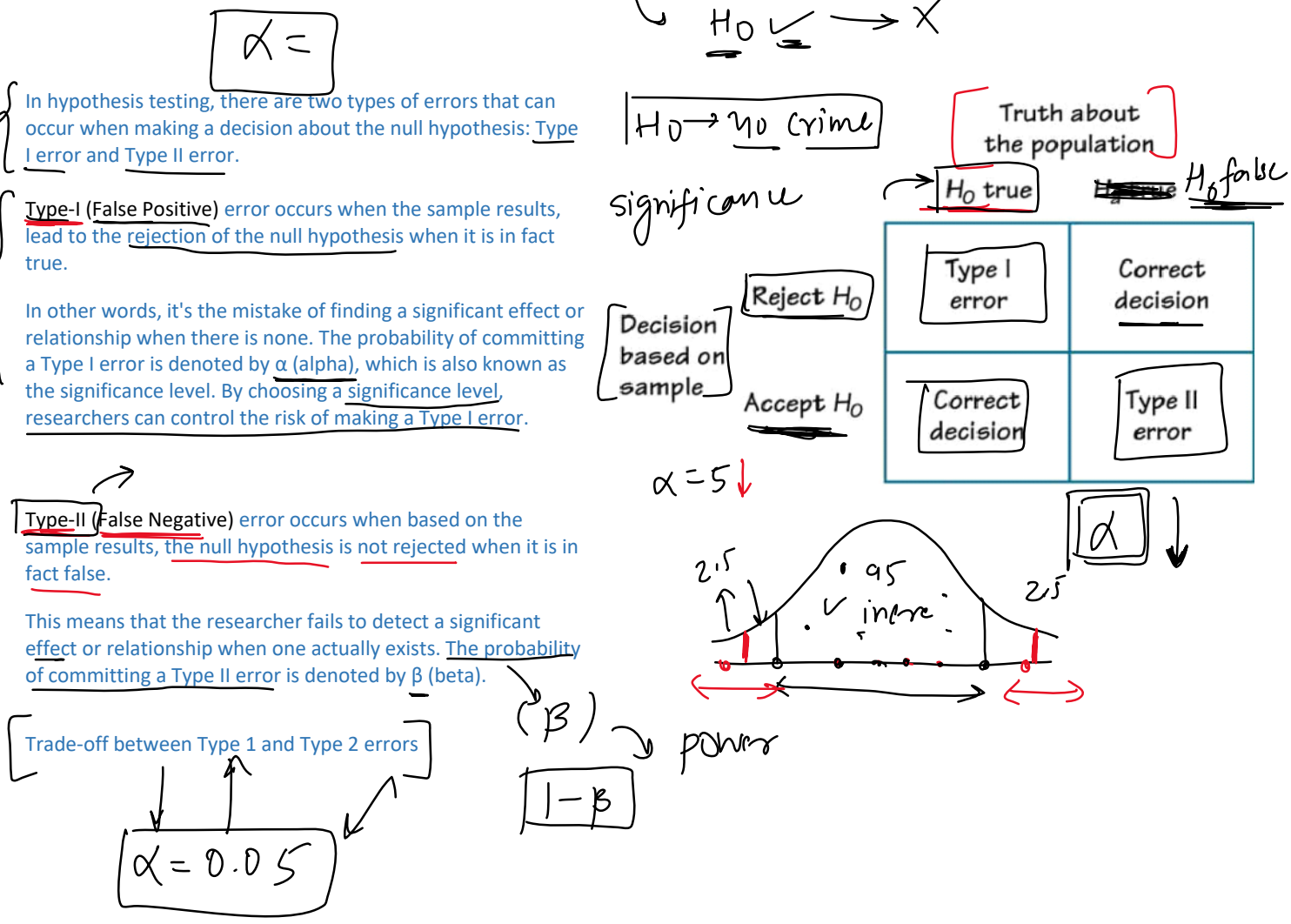
Type-I (False Positive) error occurs when the sample results, lead to the rejection of the null hypothesis when it is in fact true.

In other words, it's the mistake of finding a significant effect or relationship when there is none. The probability of committing a Type I error is denoted by α (alpha), which is also known as the significance level. By choosing a significance level, researchers can control the risk of making a Type I error.

Type-II (False Negative) error occurs when based on the sample results, the null hypothesis is not rejected when it is in fact false.

This means that the researcher fails to detect a significant effect or relationship when one actually exists. The probability of committing a Type II error is denoted by β (beta).

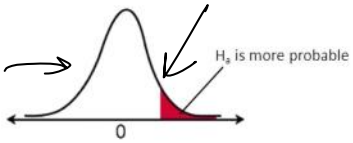
Trade-off between Type 1 and Type 2 errors



One sided vs two sided test

04 April 2023 13:29

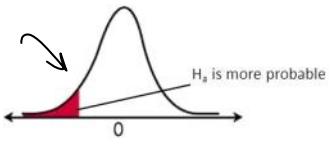
One-sided (one-tailed) test: A one-sided test is used when the researcher is interested in testing the effect in a specific direction (either greater than or less than the value specified in the null hypothesis). The alternative hypothesis in a one-sided test contains an inequality (either ">" or "<").



Right-tail test

$H_a: \mu > \text{value}$

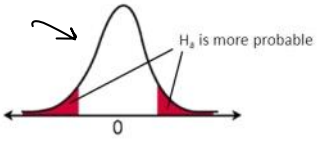
Example: A researcher wants to test whether a new medication increases the average recovery rate compared to the existing medication.



Left-tail test

$H_a: \mu < \text{value}$

Two-sided (two-tailed) test: A two-sided test is used when the researcher is interested in testing the effect in both directions (i.e., whether the value specified in the null hypothesis is different, either greater or lesser). The alternative hypothesis in a two-sided test contains a "not equal to" sign (≠).

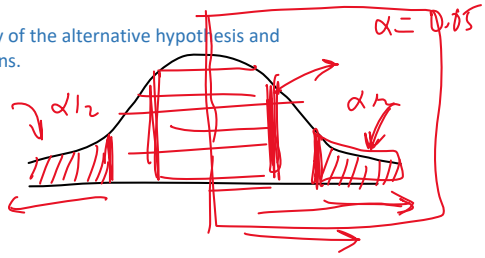


Two-tail test

$H_a: \mu \neq \text{value}$

Example: A researcher wants to test whether a new medication has a different average recovery rate compared to the existing medication.

The main difference between them lies in the directionality of the alternative hypothesis and how the significance level is distributed in the critical regions.



Advantages and Disadvantages?

Two-tailed test (two-sided):

Advantages:

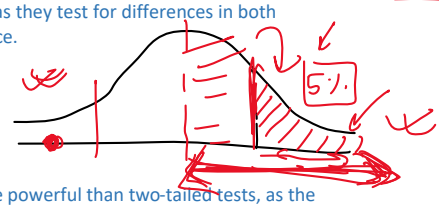
1. Detects effects in both directions: Two-tailed tests can detect effects in both directions, which makes them suitable for situations where the direction of the effect is uncertain or when researchers want to test for any difference between the groups or variables.
2. More conservative: Two-tailed tests are more conservative because the significance level (α) is split between both tails of the distribution. This reduces the risk of Type I errors in cases where the direction of the effect is uncertain.

Disadvantages:

1. Less powerful: Two-tailed tests are generally less powerful than one-tailed tests because the significance level (α) is divided between both tails of the distribution. This means the test requires a larger effect size to reject the null hypothesis, which could lead to a higher risk of Type II errors (failing to reject the null hypothesis when it is false).
2. Not appropriate for directional hypotheses: Two-tailed tests are not ideal for cases where the research question or hypothesis is directional, as they test for differences in both directions, which may not be of interest or relevance.

power = 1 − β

$\mu \neq 100$ gms



One-tailed test (one-sided):

Advantages:

1. More powerful: One-tailed tests are generally more powerful than two-tailed tests, as the entire significance level (α) is allocated to one tail of the distribution. This means that the test is more likely to detect an effect in the specified direction, assuming the effect exists.
2. Directional hypothesis: One-tailed tests are appropriate when there is a strong theoretical or practical reason to test for an effect in a specific direction.

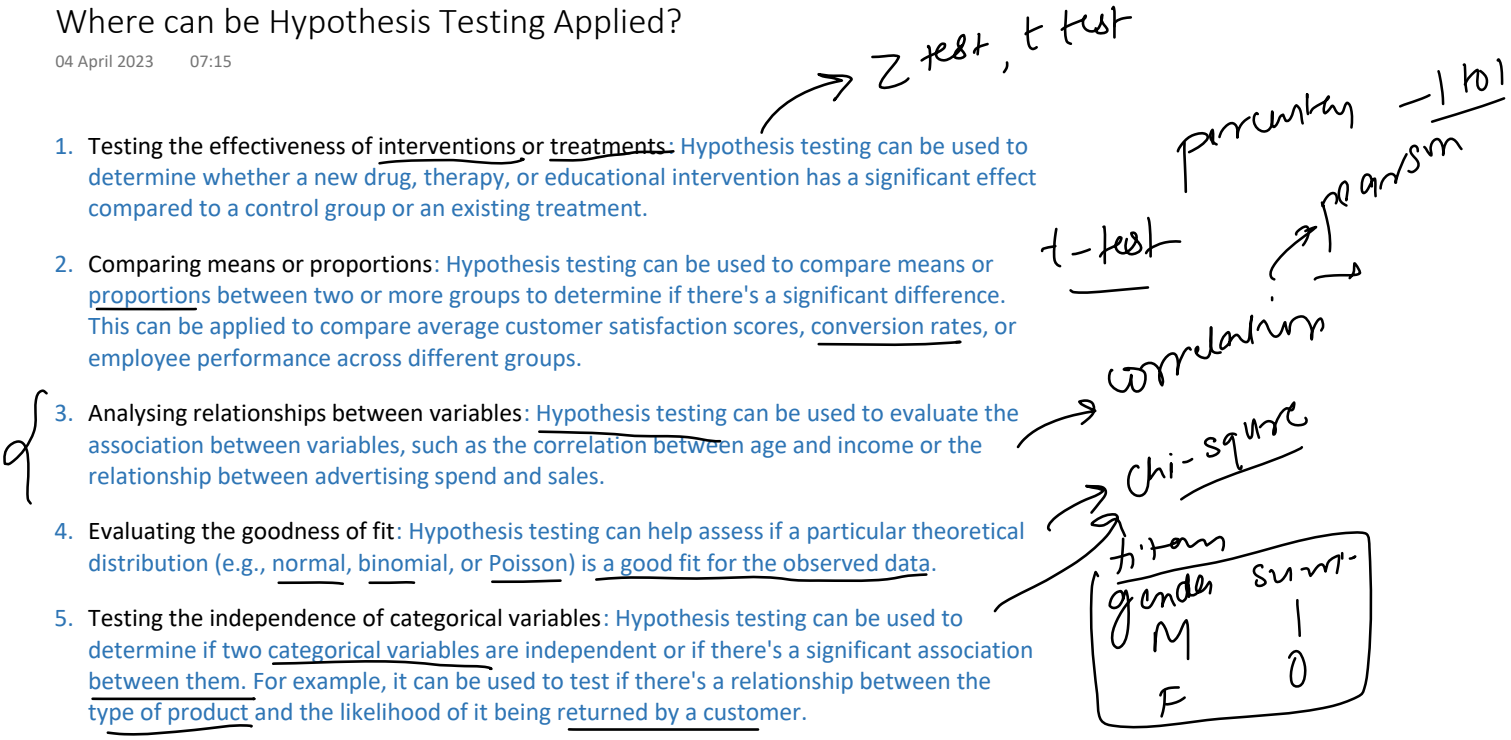
Disadvantages:

1. Missed effects: One-tailed tests can miss effects in the opposite direction of the specified alternative hypothesis. If an effect exists in the opposite direction, the test will not be able to detect it, which could lead to incorrect conclusions.
2. Increased risk of Type I error: One-tailed tests can be more prone to Type I errors if the effect is actually in the opposite direction than the one specified in the alternative hypothesis.

Where can be Hypothesis Testing Applied?

04 April 2023 07:15

1. Testing the effectiveness of interventions or treatments: Hypothesis testing can be used to determine whether a new drug, therapy, or educational intervention has a significant effect compared to a control group or an existing treatment.
2. Comparing means or proportions: Hypothesis testing can be used to compare means or proportions between two or more groups to determine if there's a significant difference. This can be applied to compare average customer satisfaction scores, conversion rates, or employee performance across different groups.
3. Analysing relationships between variables: Hypothesis testing can be used to evaluate the association between variables, such as the correlation between age and income or the relationship between advertising spend and sales.
4. Evaluating the goodness of fit: Hypothesis testing can help assess if a particular theoretical distribution (e.g., normal, binomial, or Poisson) is a good fit for the observed data.
5. Testing the independence of categorical variables: Hypothesis testing can be used to determine if two categorical variables are independent or if there's a significant association between them. For example, it can be used to test if there's a relationship between the type of product and the likelihood of it being returned by a customer.
6. A/B testing: In marketing, product development, and website design, hypothesis testing is often used to compare the performance of two different versions (A and B) to determine which one is more effective in terms of conversion rates, user engagement, or other metrics.



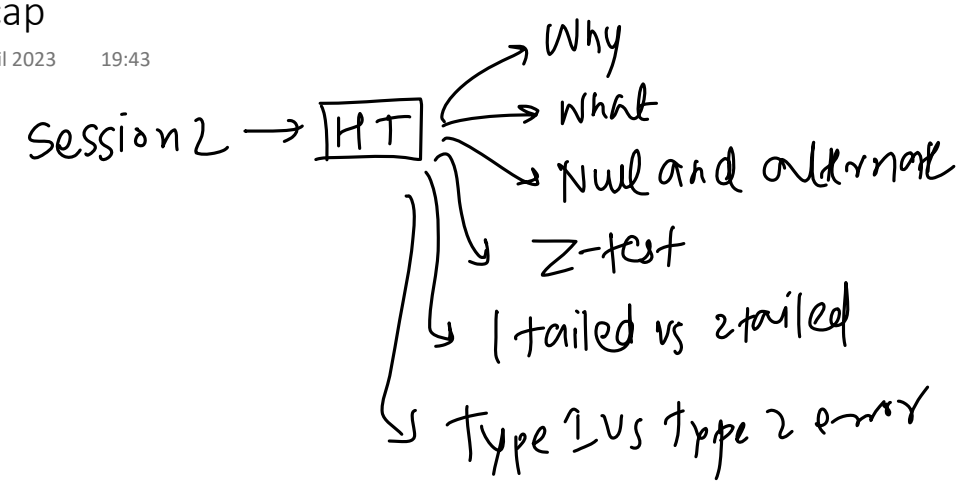
Hypothesis Testing ML Applications

04 April 2023 16:50

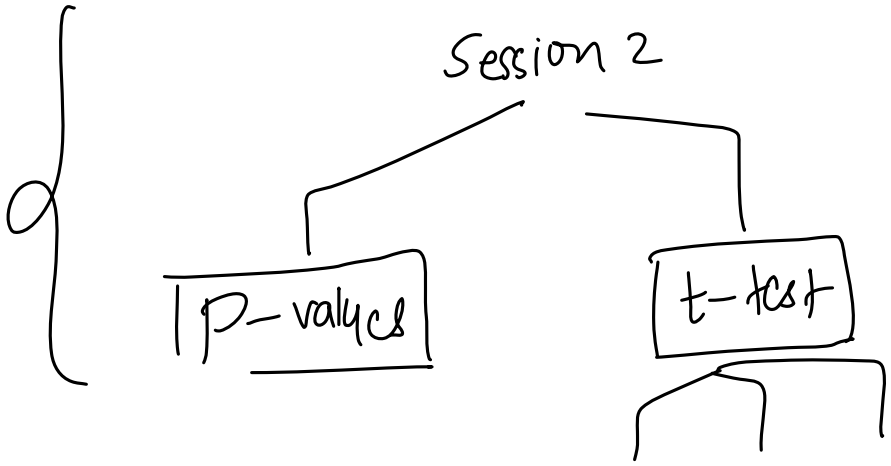
1. Model comparison: Hypothesis testing can be used to compare the performance of different machine learning models or algorithms on a given dataset. For example, you can use a paired t-test to compare the accuracy or error rate of two models on multiple cross-validation folds to determine if one model performs significantly better than the other.
2. Feature selection: Hypothesis testing can help identify which features are significantly related to the target variable or contribute meaningfully to the model's performance. For example, you can use a t-test, chi-square test, or ANOVA to test the relationship between individual features and the target variable. Features with significant relationships can be selected for building the model, while non-significant features may be excluded.
3. Hyperparameter tuning: Hypothesis testing can be used to evaluate the performance of a model trained with different hyperparameter settings. By comparing the performance of models with different hyperparameters, you can determine if one set of hyperparameters leads to significantly better performance.
4. Assessing model assumptions: In some cases, machine learning models rely on certain statistical assumptions, such as linearity or normality of residuals in linear regression. Hypothesis testing can help assess whether these assumptions are met, allowing you to determine if the model is appropriate for the data.

Recap

06 April 2023 19:43



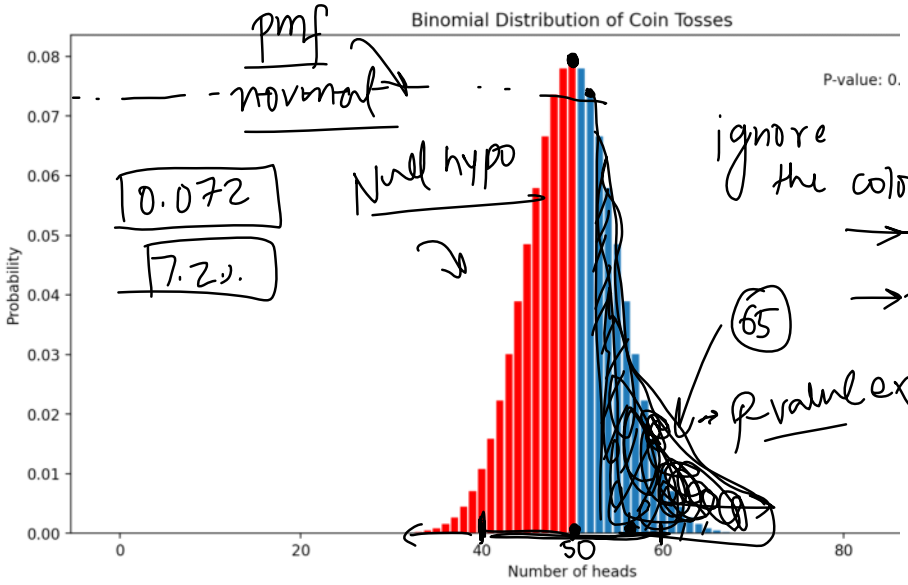
significance level (α)



P-value

06 April 2023 06:48

P-value is the probability of getting a sample as or more extreme(having more evidence against H0) than our own sample given the Null Hypothesis(H0) is true.



In simple words p-value is a measure of the strength of the evidence against the Null Hypothesis that is provided by our sample data.

Interpreting p-value

06 April 2023 08:25

With significance value

$\alpha = 0.05 / 0.01$ → p-value $\leq \alpha$ → reject your H_0

Without significance value → α → 0.05

1. Very small p-values (e.g., $p < 0.01$) indicate strong evidence against the null hypothesis, suggesting that the observed effect or difference is unlikely to have occurred by chance alone.
2. Small p-values (e.g., $0.01 \leq p < 0.05$) indicate moderate evidence against the null hypothesis, suggesting that the observed effect or difference is less likely to have occurred by chance alone.
3. Large p-values (e.g., $0.05 \leq p < 0.1$) indicate weak evidence against the null hypothesis, suggesting that the observed effect or difference might have occurred by chance alone, but there is still some level of uncertainty.
4. Very large p-values (e.g., $p \geq 0.1$) indicate weak or no evidence against the null hypothesis, suggesting that the observed effect or difference is likely to have occurred by chance alone.

P-value in context of Z-test

06 April 2023 07:08

rejection region approach p-value approach

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was 50 units per day. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day and the pop std is 4. The company wants to know if the new training program has significantly increased productivity.

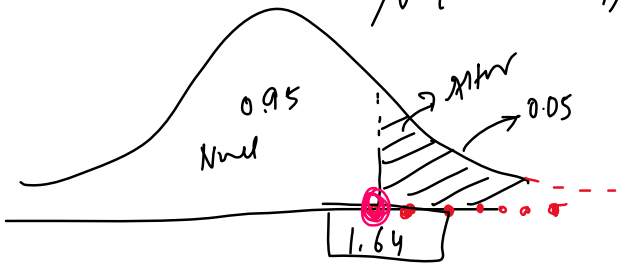
$\mu = 50$ $n = 30$ $\bar{x} = 53$

$\sigma = 4$ $\alpha = 0.05$

$H_0: \mu = 50$

$H_a: \mu > 50$

$$\text{Z-stat} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{53 - 50}{4/\sqrt{30}} = \frac{3 \times \sqrt{30}}{4} = 4.10 \quad / \quad 1.7$$



reject

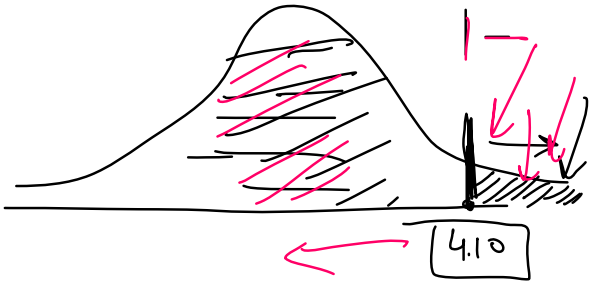
$1 - 0.95$

0.05

p-value

$Z = 4.10$

p-value → critical point



$0.999 = 0.0001$

p-value < 0.05

reject Null hypo

Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a pop standard deviation of 5 grams.

$\mu = 50$ $n = 40$ $\bar{x} = 49$

$\sigma = 5$ $\alpha = 0.05$

$H_0: \mu = 50$

$H_a: \mu \neq 50$

p-value ↙ 2 tailed $0.206 > 0.05$

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

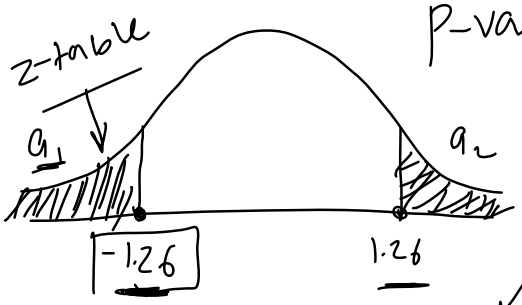
p-val 0.206

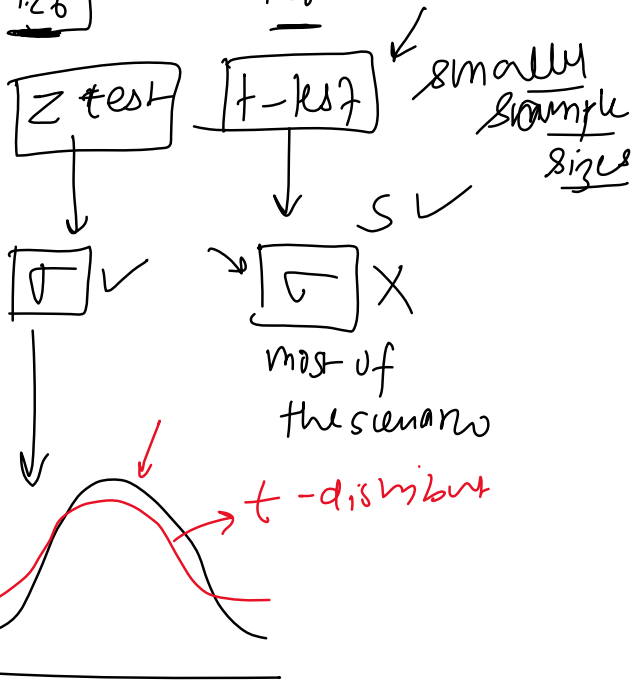
p-value $= a_1 + a_2$

$s/\sqrt{40}$ s

p-value = $a_1 + a_2$ | p-val | 0.206 |

$a_1 = 0.103 \times 2$





T-tests

06 April 2023 14:14

A t-test is a statistical test used in hypothesis testing to compare the means of two samples or to compare a sample mean to a known population mean. The t-test is based on the t-distribution, which is used when the population standard deviation is unknown and the sample size is small.

There are three main types of t-tests:

1 sample → $\bar{X}$ → μ

One-sample t-test: The one-sample t-test is used to compare the mean of a single sample to a known population mean. The null hypothesis states that there is no significant difference between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.

→ σ X

1 class → test A → popl

→ test B → pop

Independent two-sample t-test: The independent two-sample t-test is used to compare the means of two independent samples. The null hypothesis states that there is no significant difference between the means of the two samples, while the alternative hypothesis states that there is a significant difference.

Paired t-test (dependent two-sample t-test): The paired t-test is used to compare the means of two samples that are dependent or paired, such as pre-test and post-test scores for the same group of subjects or measurements taken on the same subjects under two different conditions. The null hypothesis states that there is no significant difference between the means of the paired differences, while the alternative hypothesis states that there is a significant difference.

Single Sample t-test

06 April 2023 14:14

A one-sample t-test checks whether a sample mean differs from the population mean.

Assumptions for a single sample t-test

1. Normality - Population from which the sample is drawn is normally distributed
2. Independence - The observations in the sample must be independent, which means that the value of one observation should not influence the value of another observation.
3. Random Sampling - The sample must be a random and representative subset of the population.
4. Unknown population std - The population std is not known.

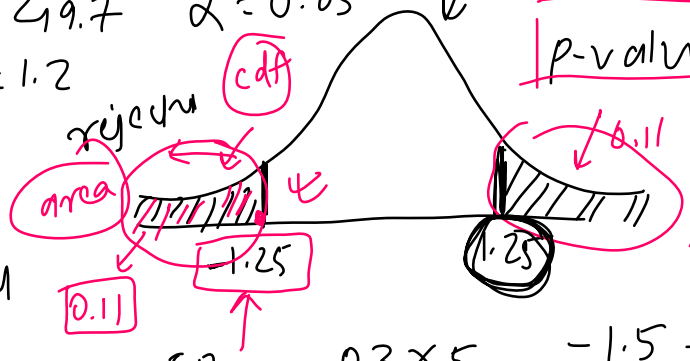
Suppose a manufacturer claims that the average weight of their new chocolate bars is 50 grams, we highly doubt that and want to check this so we drew out a sample of 25 chocolate bars and measured their weight, the sample mean came out to be 49.7 grams and the sample std deviation was 1.2 grams. Consider the significance level to be 0.05

$\mu = 50 \quad n = 25 \quad \bar{x} = 49.7 \quad \alpha = 0.05 \quad s = 1.2$

$H_0: \mu = 50$

$H_a: \mu \neq 50$

assuming it is normal



p-value = 0.22 > 0.05

Null reject ✗

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{49.7 - 50}{1.2/\sqrt{25}} = \frac{-0.3 \times 5}{1.2} = \frac{-1.5}{1.2} = -1.25$

$df = n - 1 = 24$

$H_0: \mu = 35$

$H_a: \mu < 35$

25 sample age → normal

Shapiro-wilk test

$\mu = 35$ ↓ σ ✗ → t-test

$\bar{x}, s, \alpha = 0.05$

p-value < 0.05 not normal

p-value > 0.05 normal

Python Case Study 1

06 April 2023 17:27

Independent 2 sample t-test

06 April 2023 14:15

An independent two-sample t-test, also known as an unpaired t-test, is a statistical method used to compare the means of two independent groups to determine if there is a significant difference between them.

Assumptions for the test:

1. Independence of observations: The two samples must be independent, meaning there is no relationship between the observations in one group and the observations in the other group. The subjects in the two groups should be selected randomly and independently.
2. Normality: The data in each of the two groups should be approximately normally distributed. The t-test is considered robust to mild violations of normality, especially when the sample sizes are large (typically $n \geq 30$) and the sample sizes of the two groups are similar. If the data is highly skewed or has substantial outliers, consider using a non-parametric test, such as the Mann-Whitney U test.
3. Equal variances (Homoscedasticity): The variances of the two populations should be approximately equal. This assumption can be checked using F-test for equality of variances. If this assumption is not met, you can use Welch's t-test, which does not require equal variances.
4. Random sampling: The data should be collected using a random sampling method from the respective populations. This ensures that the sample is representative of the population and reduces the risk of selection bias.

Suppose a website owner claims that there is no difference in the average time spent on their website between desktop and mobile users. To test this claim, we collect data from 30 desktop users and 30 mobile users regarding the time spent on the website in minutes. The sample statistics are as follows:

desktop users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14]

mobile_users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11]

Desktop users:

- Sample size (n1): 30
- Sample mean (mean1): 18.5 minutes
- Sample standard deviation (std_dev1): 3.5 minutes

Mobile users:

- Sample size (n2): 30
- Sample mean (mean2): 14.3 minutes
- Sample standard deviation (std_dev2): 2.7 minutes

We will use a significance level (α) of 0.05 for the hypothesis test.

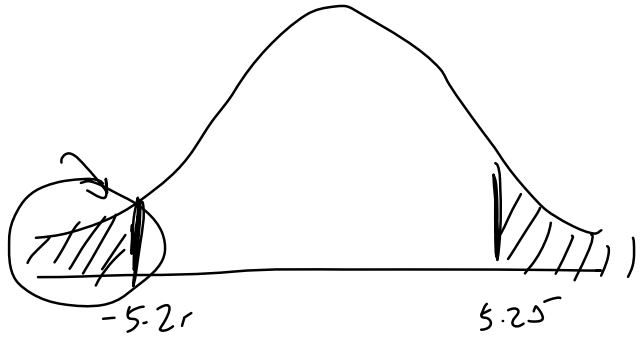
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

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

t-statistic

$$t = \frac{18.5 - 14.3}{} = \frac{4.2}{} = 5.25$$

$$t = \frac{\quad}{\sqrt{\frac{(3.5)^2}{30} + \frac{(2.7)^2}{30}}} = \frac{4.2}{\sqrt{\frac{19.54}{30}}} = \boxed{5.45}$$



$n-1$

$n_1 + n_2 - 2$

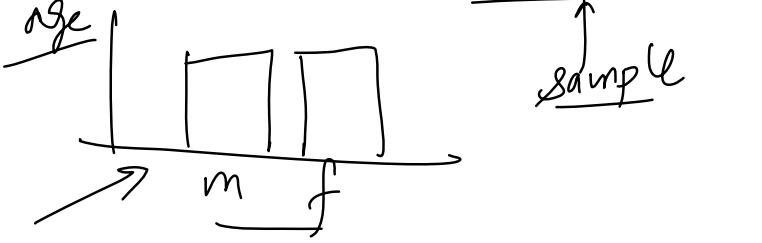
$30+30-2 = 58$

$\mu_m > \mu_f$

$\mu_m = \mu_f$

age and gender

pop



$H_0: \mu_m = \mu_f$

$H_1: \mu_m > \mu_f$

$\alpha = 0.05$

Python Case Study 2

06 April 2023 17:27

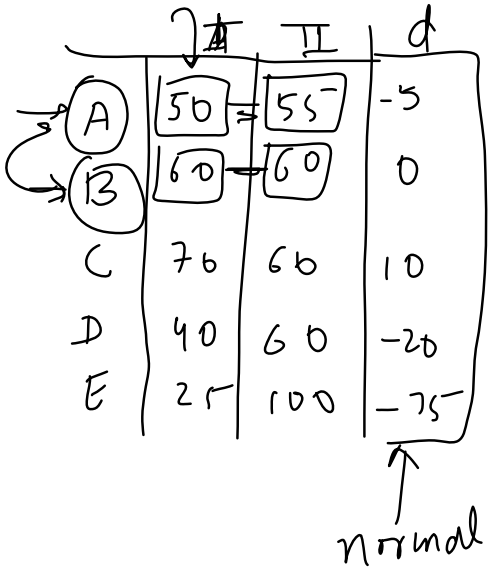
Paired 2 sample t-test

06 April 2023 14:21

A paired two-sample t-test, also known as a dependent or paired-samples t-test, is a statistical test used to compare the means of two related or dependent groups.

Common scenarios where a paired two-sample t-test is used include:

1. Before-and-after studies: Comparing the performance of a group before and after an intervention or treatment.
2. Matched or correlated groups: Comparing the performance of two groups that are matched or correlated in some way, such as siblings or pairs of individuals with similar characteristics.



Assumptions

1. Paired observations: The two sets of observations must be related or paired in some way, such as before-and-after measurements on the same subjects or observations from matched or correlated groups.
2. Normality: The differences between the paired observations should be approximately normally distributed. This assumption can be checked using graphical methods (e.g., histograms, Q-Q plots) or statistical tests for normality (e.g., Shapiro-Wilk test). Note that the t-test is generally robust to moderate violations of this assumption when the sample size is large.
3. Independence of pairs: Each pair of observations should be independent of other pairs. In other words, the outcome of one pair should not affect the outcome of another pair. This assumption is generally satisfied by appropriate study design and random sampling.

Let's assume that a fitness center is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants' weight.

$\mu_{dif} = \mu_{bef} - \mu_{aft} = 0$

$H_0: \mu_{before} = \mu_{after}$

$H_1: \mu_{before} > \mu_{after}$

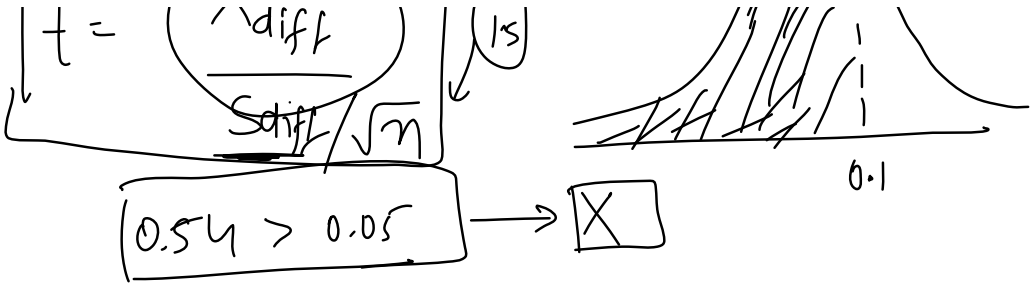
Before the program:
[80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91]

After the program:
[78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]

Significance level (α) = 0.05

> 0.05

$$t = \frac{\bar{x}_{diff}}{\quad} \quad (15)$$



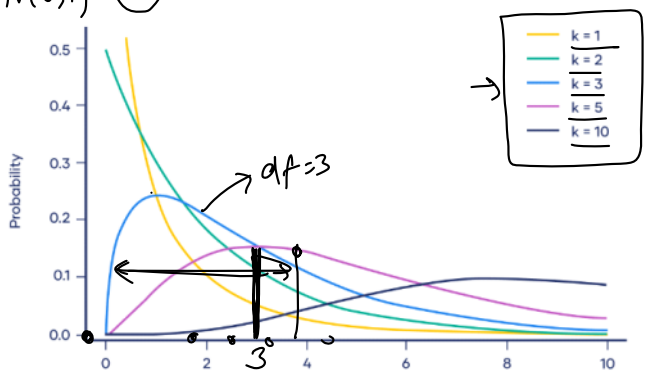
$0.54 > 0.05$ → X

Chi Square Distribution → distribution → continuous pd

07 April 2023 10:29

$\chi \sim N(0,1)$ (N)

The Chi-Square distribution, also written as χ^2 distribution, is a continuous probability distribution that is widely used in statistical hypothesis testing, particularly in the context of goodness-of-fit tests and tests for independence in contingency tables. It arises when the sum of the squares of independent standard normal random variables follows this distribution.

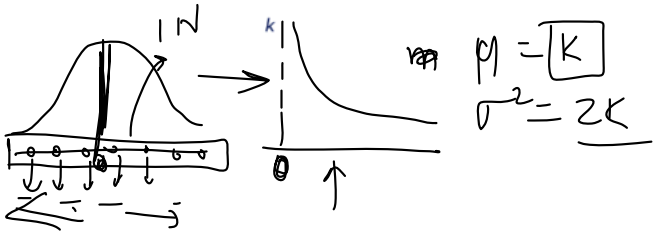


The Chi-Square distribution has a single parameter, the degrees of freedom (df), which influences the shape and spread of the distribution. The degrees of freedom are typically associated with the number of independent variables or constraints in a statistical problem.

Some key properties of the Chi-Square distribution are:

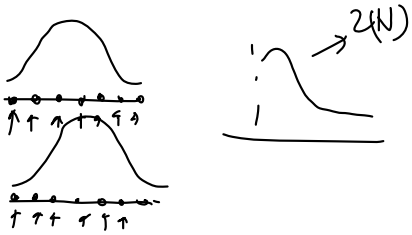
a. It is a continuous distribution, defined for non-negative values.
b. It is positively skewed, with the degree of skewness decreasing as the degrees of freedom increase.
c. The mean of the Chi-Square distribution is equal to its degrees of freedom, and its variance is equal to twice the degrees of freedom.
d. As the degrees of freedom increase, the Chi-Square distribution approaches the normal distribution in shape.

χ^2



The Chi-Square distribution is used in various statistical tests, such as the Chi-Square goodness-of-fit test which evaluates whether an observed frequency distribution fits an expected theoretical distribution, and the Chi-Square test for independence which checks the association between categorical variables in a contingency table.

chi-square



$\chi^2 = Z^2$ → degree of freedom

→ df=1

$\chi^2 = Z_1^2 + Z_2^2$ → df=2

$\chi^2 = Z_1^2 + Z_2^2 + Z_3^2$ → df=3

$\chi^2 = \sum_{i=1}^{k} Z_k^2$ df=K

df ↑

Chi Square Test

07 April 2023 15:03

→ Categorical ←

Z-test t-test

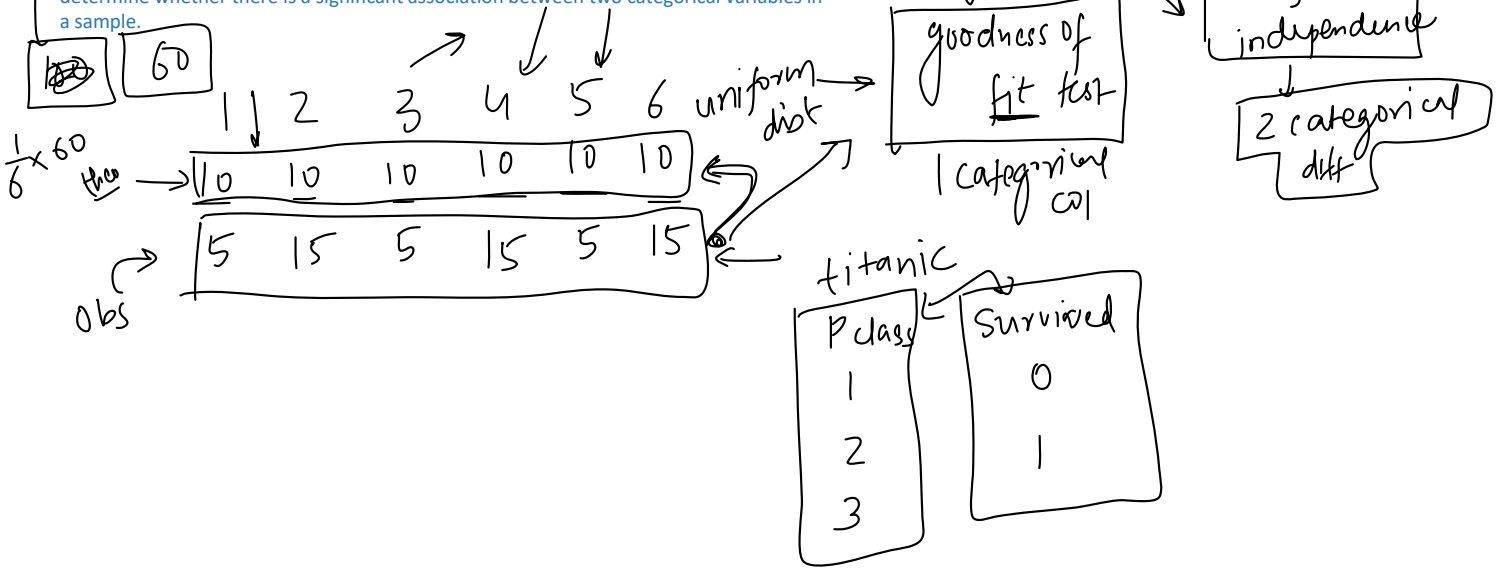
↓

continous

χ^2-square

The Chi-Square test is a statistical hypothesis test used to determine if there is a significant association between categorical variables or if an observed distribution of categorical data differs from an expected theoretical distribution. It is based on the Chi-Square (χ^2) distribution, and it is commonly applied in two main scenarios:

1. Chi-Square Goodness-of-Fit Test: This test is used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution. It is often applied to check if the data follows a specific probability distribution, such as the uniform or binomial distribution.
2. Chi-Square Test for Independence (Chi-Square Test for Association): This test is used to determine whether there is a significant association between two categorical variables in a sample.



titanic

Pclass	Survived
1	0
2	1
3	

Goodness of Fit Test

07 April 2023 10:29

The Chi-Square Goodness-of-Fit test is a statistical hypothesis test used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution. It helps to evaluate whether the data follows a specific probability distribution, such as uniform, binomial, or Poisson distribution, among others. This test is particularly useful when you want to assess if the sample data is consistent with an assumed distribution or if there are significant deviations from the expected pattern.

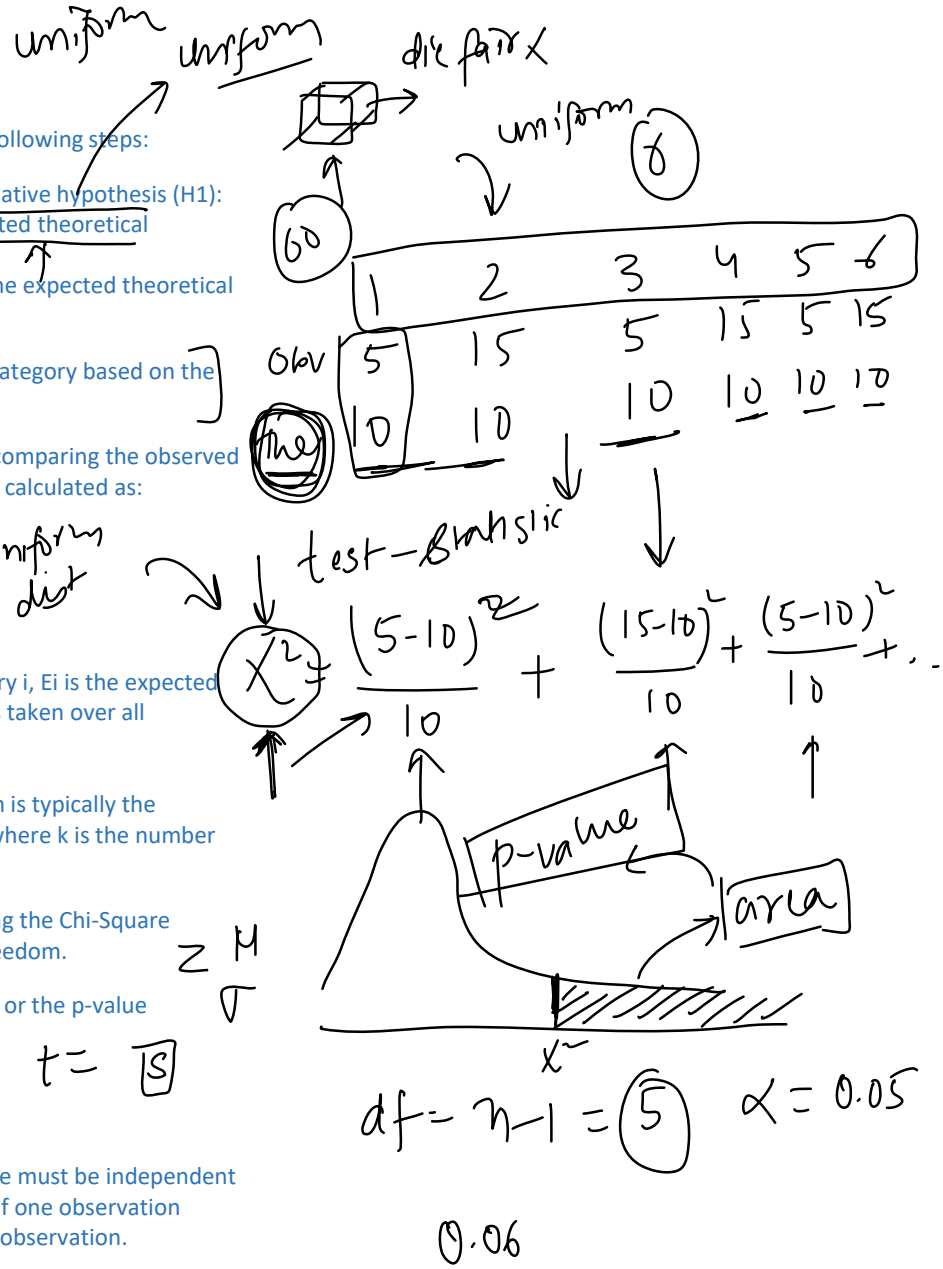
Steps

The Chi-Square Goodness-of-Fit test involves the following steps:

- Define the null hypothesis (H0) and the alternative hypothesis (H1):
  - H0: The observed data follows the expected theoretical distribution.
  - H1: The observed data does not follow the expected theoretical distribution.
- Calculate the expected frequencies for each category based on the theoretical distribution and the sample size.
- Compute the Chi-Square test statistic (χ^2) by comparing the observed and expected frequencies. The test statistic is calculated as:

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

- where Oi is the observed frequency in category i, Ei is the expected frequency in category i, and the summation is taken over all categories.
- Determine the degrees of freedom (df), which is typically the number of categories minus one (df = k - 1), where k is the number of categories.
- Calculate the p-value for the test statistic using the Chi-Square distribution with the calculated degrees of freedom.
- Compare the test statistic to the critical value or the p-value



Assumptions

1. Independence: The observations in the sample must be independent of each other. This means that the outcome of one observation should not influence the outcome of another observation.
2. Categorical data: The variable being analysed must be categorical, not continuous or ordinal. The data should be divided into mutually exclusive and exhaustive categories.
3. Expected frequency: Each category should have an expected frequency of at least 5. This guideline helps ensure that the Chi-Square distribution is a reasonable approximation for the distribution of the test statistic. Having small expected frequencies can lead to an inaccurate estimation of the Chi-Square distribution, potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).

potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).

4. Fixed distribution: The theoretical distribution being compared to the observed data should be specified before the test is conducted. It is essential to avoid choosing a distribution based on the observed data, as doing so can lead to biased results.

parametric or non-parametric

The Chi-Square Goodness-of-Fit test is a non-parametric test. Non-parametric tests do not assume that the data comes from a specific probability distribution or make any assumptions about population parameters like the mean or standard deviation.

In the Chi-Square Goodness-of-Fit test, we compare the observed frequencies of the categorical data to the expected frequencies based on a hypothesized distribution. The test doesn't rely on any assumptions about the underlying distribution's parameters. Instead, it focuses on comparing observed counts to expected counts, making it a non-parametric test.

Example 1

07 April 2023 16:26

Suppose we have a six-sided fair die, and we want to test if the die is indeed fair. We roll the die 60 times and record the number of times each side comes up. We'll use the Chi-Square Goodness-of-Fit test to determine if the observed frequencies are consistent with a fair die (i.e., a uniform distribution of the sides).

Observed frequencies:

- Side 1: 12 times
- Side 2: 8 times
- Side 3: 11 times
- Side 4: 9 times
- Side 5: 10 times
- Side 6: 10 times

uniform

H_0: die is fair → uniform

H_1: die is not fair

	1	2	3	4	5	6	
Obs	12	8	11	9	10	10	→ 60
expected	10	10	10	10	10	10	

$60 \cdot \frac{1}{6}$

$$\chi^2 = \frac{(12-10)^2 + (8-10)^2 + (11-10)^2 + (9-10)^2}{10} + \frac{4+4+1+1}{10} = \frac{10}{10} = 1$$

$\chi^2 = 1$ → chisquare

$df = n - 1 = 6 - 1 = 5$

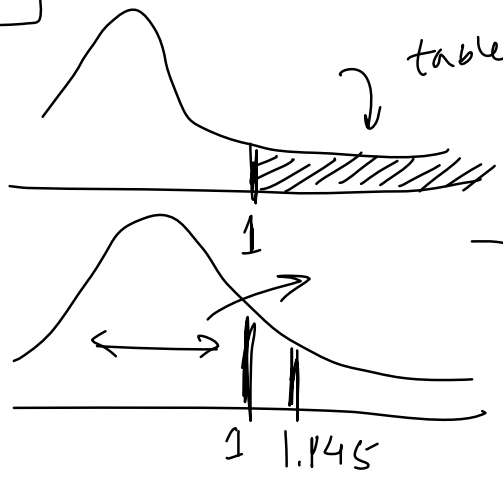


table → python pvalue 0.96

$\alpha = 0.05$

pvalue > 0.05

not reject — fair die

Example 2

07 April 2023 15:26

Suppose a marketing team at a retail company wants to understand the distribution of visits to their website by day of the week. They have a hypothesis that visits are uniformly distributed across all days of the week, meaning they expect an equal number of visits on each day. They collected data on website visits for four weeks and want to test if the observed distribution matches the expected uniform distribution.

Observed frequencies (number of website visits per day of the week for four weeks):

- Monday: 420
- Tuesday: 380
- Wednesday: 410
- Thursday: 400
- Friday: 410
- Saturday: 430
- Sunday: 390

uniform

	mon	tue	wed	thu	fri	sat	sun
Obs	420	380	410	400	410	430	390
EXP	405	405	405	405	405	405	405

H_0: uniform dis

H_a: Not uniform

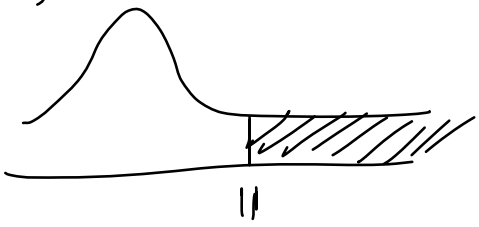
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(420-405)^2 + (}{405}$$

$\chi^2 =$ LB

0.08 — p-value

$df = n - 1 = 7 - 1 = 6$

$0.08 < \alpha \rightarrow$



Example 3

07 April 2023 15:27

survey → village → 800 families → 4 children

A survey of 800 families in a village with 4 children each revealed the following distribution:

# girls	4	3	2	1	0
# boys	0	1	2	3	4
# families (theoretical)	32	178	290	236	64

binomial

50%

$P(S) = P(d) = \frac{1}{2}$

$H_0: P(m) = P(f) = \frac{1}{2}$

$H_a: P(m) \neq P(f)$

→ binomial, $p = \frac{1}{2}$

Is this data consistent with the result that male and female births are equally probable?

$p, q = 1 - p$

→ $n = 4$, $p = \frac{1}{2}$

$nC_x \, p^x (1-p)^{n-x}$

	0	1	2	3	4
Obs	32	178	290	236	64
Theo	50	200	300	200	50

$P(0) = {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \frac{1}{16} \times 800 = 50$

$P(1) = {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \frac{4!}{3!} \times \frac{1}{16} = \frac{1}{4} \times 800 = 200$

$P(2) = {}^4C_2 \left(\frac{1}{2}\right)^4 = \frac{4!}{2!2!} \times \frac{1}{16} = \frac{4 \times 3}{2} \times \frac{1}{16} \times 800 = \frac{6}{16} \times 800$

$P(3) = {}^4C_3 \left(\frac{1}{2}\right)^4$

4×1

$$\chi^2 = \frac{(32-50)^2}{50} + \frac{(178-200)^2}{200} + \frac{(290-300)^2}{300} + \frac{(236-200)^2}{200} + \frac{(64-50)^2}{50}$$

$$= \frac{324}{50} + \frac{484}{200} + \frac{100}{300} + \frac{1296}{200} + \frac{196}{50}$$

$$= 6.2 + 2.3 + 0.33 + 6.2 + 3.9$$

$\chi^2 = 18.93$

$df = 5 - 1 = 4$

$0.00081 < \alpha\ (0.05)$

reject the Null hypothesis

Python Case Study

07 April 2023 15:27

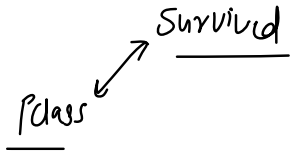
1	2	3	
216	184	491	obs
299	299	270	exp

$\chi^2 =$

Test for Independence

07 April 2023 10:29

The Chi-Square test for independence, also known as the Chi-Square test for association, is a statistical test used to determine whether there is a significant association between two categorical variables in a sample. It helps to identify if the occurrence of one variable is dependent on the occurrence of the other variable, or if they are independent of each other.



The test is based on comparing the observed frequencies in a contingency table (a table that displays the frequency distribution of the variables) with the frequencies that would be expected under the assumption of independence between the two variables.

Steps

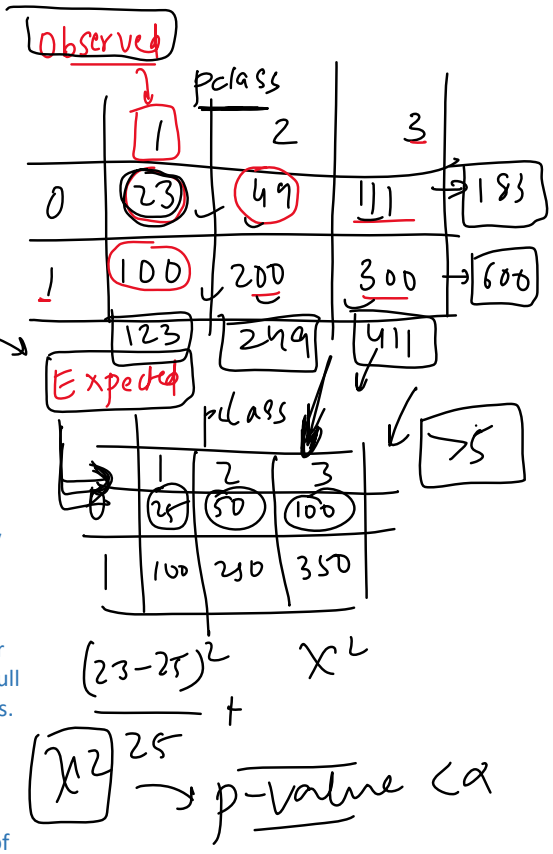
1. State the null hypothesis (H0) and alternative hypothesis (H1):
   - H0: There is no association between the two categorical variables (they are independent).
   - H1: There is an association between the two categorical variables (they are dependent).
2. Create a contingency table with the observed frequencies for each combination of the categories of the two variables.
3. Calculate the expected frequencies for each cell in the contingency table assuming that the null hypothesis is true (i.e., the variables are independent).
4. Compute the Chi-Square test statistic:

   $\chi^2 = \Sigma [(O_{ij} - E_{ij})^2 / E_{ij}]$

   where O_ij is the observed frequency in each cell and E_ij is the expected frequency.

   $(2-1)(3-1)$ df

   $1 \times 2 = 2$
5. Determine the degrees of freedom: df = (number of rows - 1) * (number of columns - 1)
6. Obtain the critical value or p-value using the Chi-Square distribution table or a statistical software/calculator with the given degrees of freedom and significance level (commonly $\alpha = 0.05$).
7. Compare the test statistic to the critical value or the p-value to the significance level to decide whether to reject or fail to reject the null hypothesis. If the test statistic is greater than the critical value, or if the p-value is less than the significance level, we reject the null hypothesis and conclude that there is a significant association between the two variables.



Assumptions

1. Independence of observations: The observations in the sample should be independent of each other. This means that the occurrence of one observation should not affect the occurrence of another observation. In practice, this usually implies that the data should be collected using a simple random sampling method.
2. Categorical variables: Both variables being tested must be categorical, either ordinal or nominal. The Chi-Square test for independence is not appropriate for continuous variables.
3. Adequate sample size: The sample size should be large enough to ensure that the expected frequency for each cell in the contingency table is sufficient. A common rule of thumb is that the expected frequency for each cell should be at least 5. If some cells have expected frequencies less than 5, the test may not be valid, and other methods like Fisher's exact test may be more appropriate.
4. Fixed marginal totals: The marginal totals (the row and column sums of the contingency table) should be fixed before the data is collected. This is because the Chi-Square test for independence assesses the association between the two variables under the assumption that the marginal totals are fixed and not influenced by the relationship between the variables.

Example 1

07 April 2023 17:50

A researcher wants to investigate if there is an association between the level of education (categorical variable) and the preference for a particular type of exercise (categorical variable) among a group of 150 individuals. The researcher collects data and creates the following contingency table

Observed

Education	Exercise Type: Yoga	Running	Swimming	Total
High School	15	20	10	45
Bachelor's	20	30	15	65
Master's or PhD	5	15	20	40
Total	40	65	45	150

edu ⟷ exercise (independent)

H_0: they are independent ✗

H_1: they are associated

need a confidence

expected

	Yoga	run	swim
High	12	19	13.5
Bach	17	28	20
PhD	10	17	12

$\frac{45 \times 40}{150}$ $\frac{65 \times 45}{150}$ $\times$ $\frac{45 \times 45}{150}$

$\frac{(15-12)^2}{12} + \frac{(20-19)^2}{20} + \frac{(10-13.5)^2}{10}$

$+ \_\_ \_\_ \_\_$

$\chi^2 = 9.95$

p-value 0.04 (α

reject Null $df = (3-1)(3-1) = 4$

Python Case Study

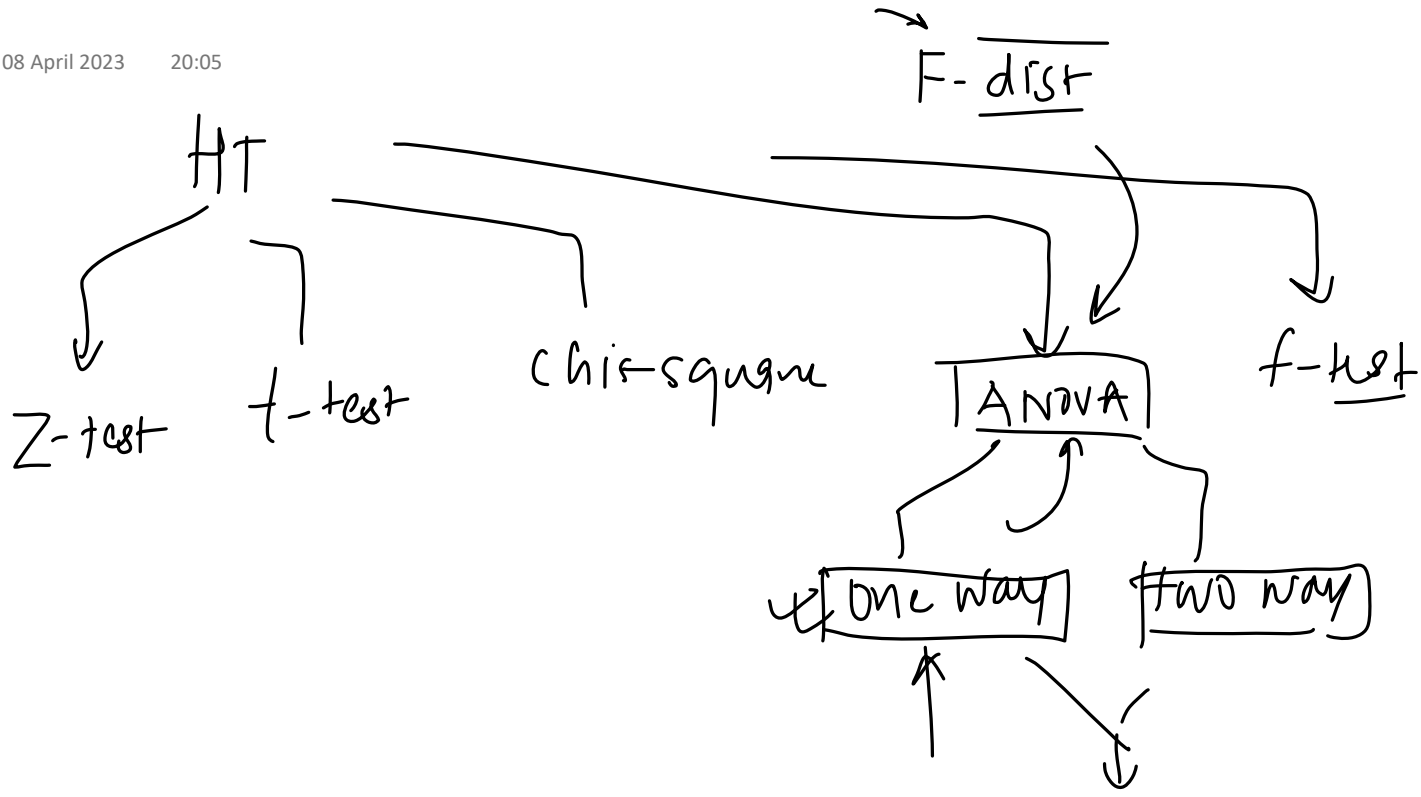
07 April 2023 15:06

Applications in Machine Learning

07 April 2023 17:30

1. Feature selection: Chi-Square test can be used as a filter-based feature selection method to rank and select the most relevant categorical features in a dataset. By measuring the association between each categorical feature and the target variable, you can eliminate irrelevant or redundant features, which can help improve the performance and efficiency of machine learning models.
2. Evaluation of classification models: For multi-class classification problems, the Chi-Square test can be used to compare the observed and expected class frequencies in the confusion matrix. This can help assess the goodness of fit of the classification model, indicating how well the model's predictions align with the actual class distributions.
3. Analysing relationships between categorical features: In exploratory data analysis, the Chi-Square test for independence can be applied to identify relationships between pairs of categorical features. Understanding these relationships can help inform feature engineering and provide insights into the underlying structure of the data.
4. Discretization of continuous variables: When converting continuous variables into categorical variables (binning), the Chi-Square test can be used to determine the optimal number of bins or intervals that best represent the relationship between the continuous variable and the target variable.
5. Variable selection in decision trees: Some decision tree algorithms, such as the CHAID (Chi-squared Automatic Interaction Detection) algorithm, use the Chi-Square test to determine the most significant splitting variables at each node in the tree. This helps construct more effective and interpretable decision trees.

08 April 2023 20:05

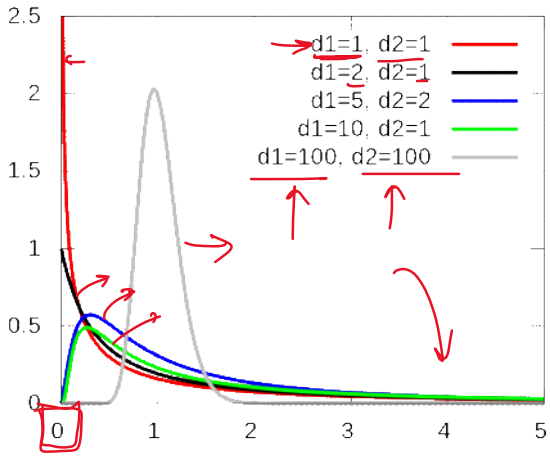


F - Distribution

08 April 2023 13:04

→ chi-square → (normal)²

1. Continuous probability distribution: The F-distribution is a continuous probability distribution used in statistical hypothesis testing and analysis of variance (ANOVA).
2. Fisher-Snedecor distribution: It is also known as the Fisher-Snedecor distribution, named after Ronald Fisher and George Snedecor, two prominent statisticians.
3. Degrees of freedom: The F-distribution is defined by two parameters - the degrees of freedom for the numerator (df1) and the degrees of freedom for the denominator (df2).
4. Positively skewed and bounded: The shape of the F-distribution is positively skewed, with its left bound at zero. The distribution's shape depends on the values of the degrees of freedom.
5. Testing equality of variances: The F-distribution is commonly used to test hypotheses about the equality of two variances in different samples or populations.
6. Comparing statistical models: The F-distribution is also used to compare the fit of different statistical models, particularly in the context of ANOVA.
7. F-statistic: The F-statistic is calculated by dividing the ratio of two sample variances or mean squares from an ANOVA table. This value is then compared to critical values from the F-distribution to determine statistical significance.
8. Applications: The F-distribution is widely used in various fields of research, including psychology, education, economics, and the natural and social sciences, for hypothesis testing and model comparison.



SSB ← χ_1^2 → df_1

SSW ← χ_2^2 → df_2

$F = \frac{\chi_1^2/df_1}{\chi_2^2/df_2}$ → f-distribution

df_1 df_2

One way ANOVA test

08 April 2023 13:12

One-way ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more independent groups to determine if there are any significant differences between them. It is an extension of the t-test, which is used for comparing the means of two independent groups. The term "one-way" refers to the fact that there is only one independent variable (factor) with multiple levels (groups) in this analysis.

The primary purpose of one-way ANOVA is to test the null hypothesis that all the group means are equal. The alternative hypothesis is that at least one group mean is significantly different from the others.

Steps

- Define the null and alternative hypotheses.
- Calculate the overall mean (grand mean) of all the groups combined and mean of all the groups individually.
- Calculate the "between-group" and "within-group" sum of squares (SS).
- Find the between group and within group degree of freedoms
- Calculate the "between-group" and "within-group" mean squares (MS) by dividing their respective sum of squares by their degrees of freedom.
- Calculate the F-statistic by dividing the "between-group" mean square by the "within-group" mean square.

ANOVA table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS) (This is SS Divided by d.f.) and is an Estimation of Variance to be Used in F-ratio	F-ratio
Between samples or categories	$n_1(\bar{X}_1-\bar{\bar{X}})^2+\cdots+n_k(\bar{X}_k-\bar{\bar{X}})^2$	$(k-1)$	$\frac{SS\text{ between}}{(k-1)}$	$\frac{MS\text{ between}}{MS\text{ within}}$
Within samples or categories	$\sum(X_{1i}-\bar{X}_1)^2+\cdots+\sum(X_{ki}-\bar{X}_k)^2$ $i=1,2,3,\ldots$	$(n-k)$	$\frac{SS\text{ within}}{(n-k)}$	
Total	$\sum(X_{ij}-\bar{\bar{X}})^2$ $i,j=1,2,3,\ldots$	$(n-1)$		

- Calculate the p-value associated with the calculated F-statistic using the F-distribution and the appropriate degrees of freedom. The p-value represents the probability of obtaining an F-statistic as extreme or more extreme than the calculated value, assuming the null hypothesis is true.
- Choose a significance level (alpha), typically 0.05.
- Compare the calculated p-value with the chosen significance level (alpha).

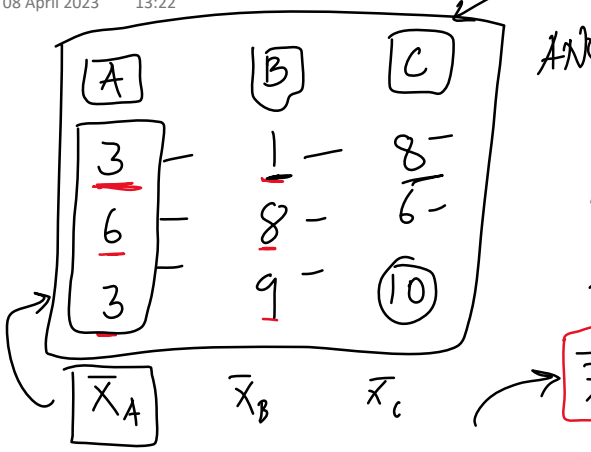
a. If the p-value is less than or equal to alpha, reject the null hypothesis in favour of the alternative hypothesis, concluding that there is a significant difference between at least one pair of group means.
b. If the p-value is greater than alpha, fail to reject the null hypothesis, concluding that there is not enough evidence to suggest a significant difference between the group means.

It's important to note that one-way ANOVA only determines if there is a significant difference between the group means; it does not identify which specific groups have significant differences. To determine which pairs of groups are significantly different, post-hoc tests, such as Tukey's HSD or Bonferroni, are conducted after a significant ANOVA result.

Example 1

08 April 2023 13:22

number of



ANOVA

H_0: $\mu_A = \mu_B = \mu_C$

H_1: atleast 1 of them is signifi diff

$n = 9$ $K = 3$

$\bar{\bar{x}} = 6$ $\bar{X}_A = 4$ $\bar{X}_B = 6$ $\bar{X}_C = 8$

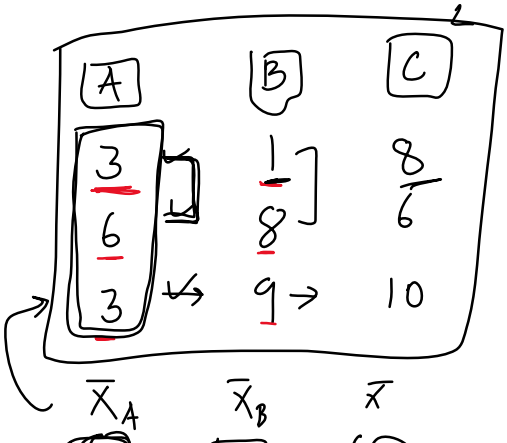
SST → Sum of square Total → $df = n-1 = 9-1 = 8$

$(6-3)^2 + (6-6)^2 + (6-3)^2 + (6-1)^2 + (6-8)^2 + (6-9)^2 +$
$(6-1)^2 + (6-6)^2 + (6-10)^2$

$9 + 0 + 9 + 25 + 4 + 9 + 4 + 0 + 16 = 76$

SSW → Sum of squares within

$df = 6$ $n-k = 9-3 = 6$



$\bar{\bar{x}} = 6$

$(4-3)^2 + (4-6)^2 + (4-3)^2 +$
$(6-1)^2 + (6-8)^2 + (6-9)^2 +$
$(8-8)^2 + (8-1)^2 + (8-10)^2$ SSW = 52

$1 + 9 + 1 + 25 + 4 + 9 + 0 + 4 + 4$

SSB → $df = 2$

$3 \times (6-4)^2 + 3 \times (6-6)^2 + 3 \times (6-8)^2$

$3 \times 4 + 0 + 3 \times 4 = 24 = SSB$

quantity	value	df	
SSW	24	12	$F = \frac{24}{2}$ 12 x 6

quantity

→ SSB

→ SSW

→ SST

24

52

76

df

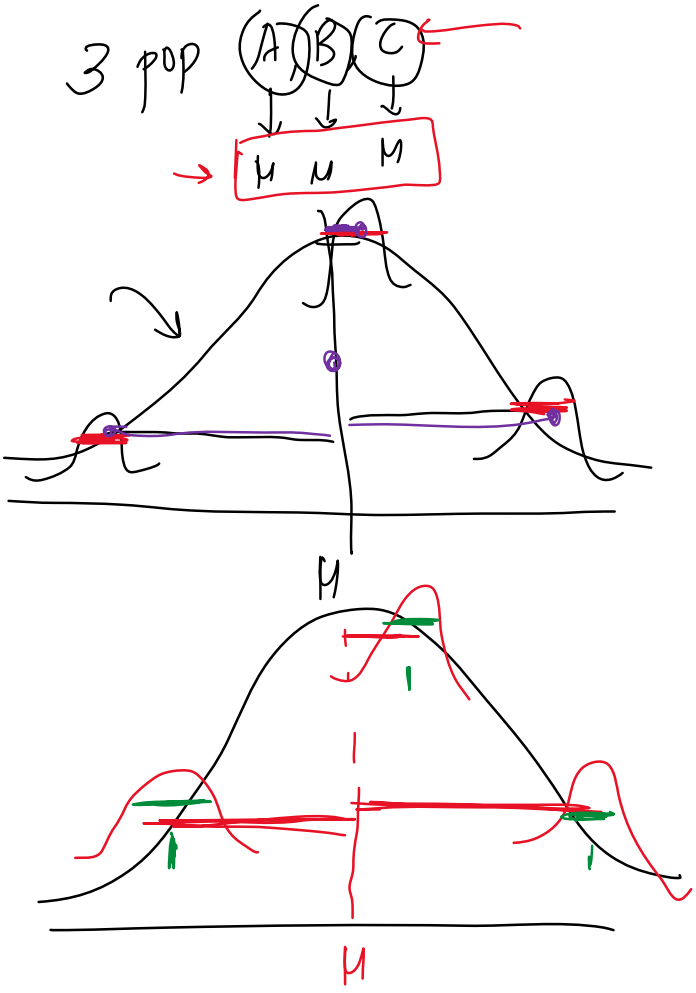
2

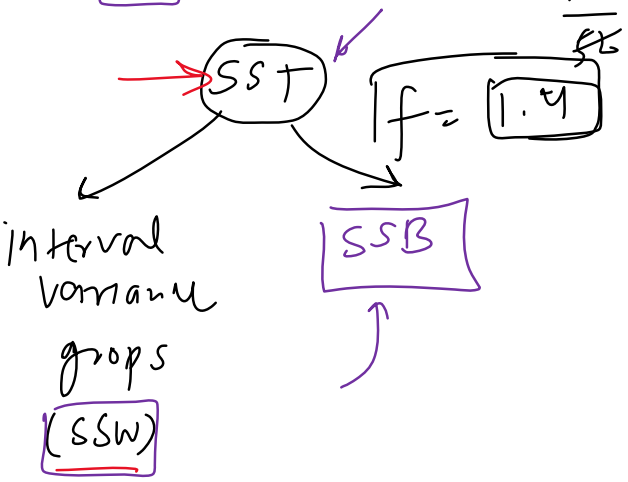
6

8

$F = \frac{\frac{24}{2}}{\frac{52}{6}} = \frac{12 \times 6}{52} = \frac{72}{52}$

SST = SSW + SSB





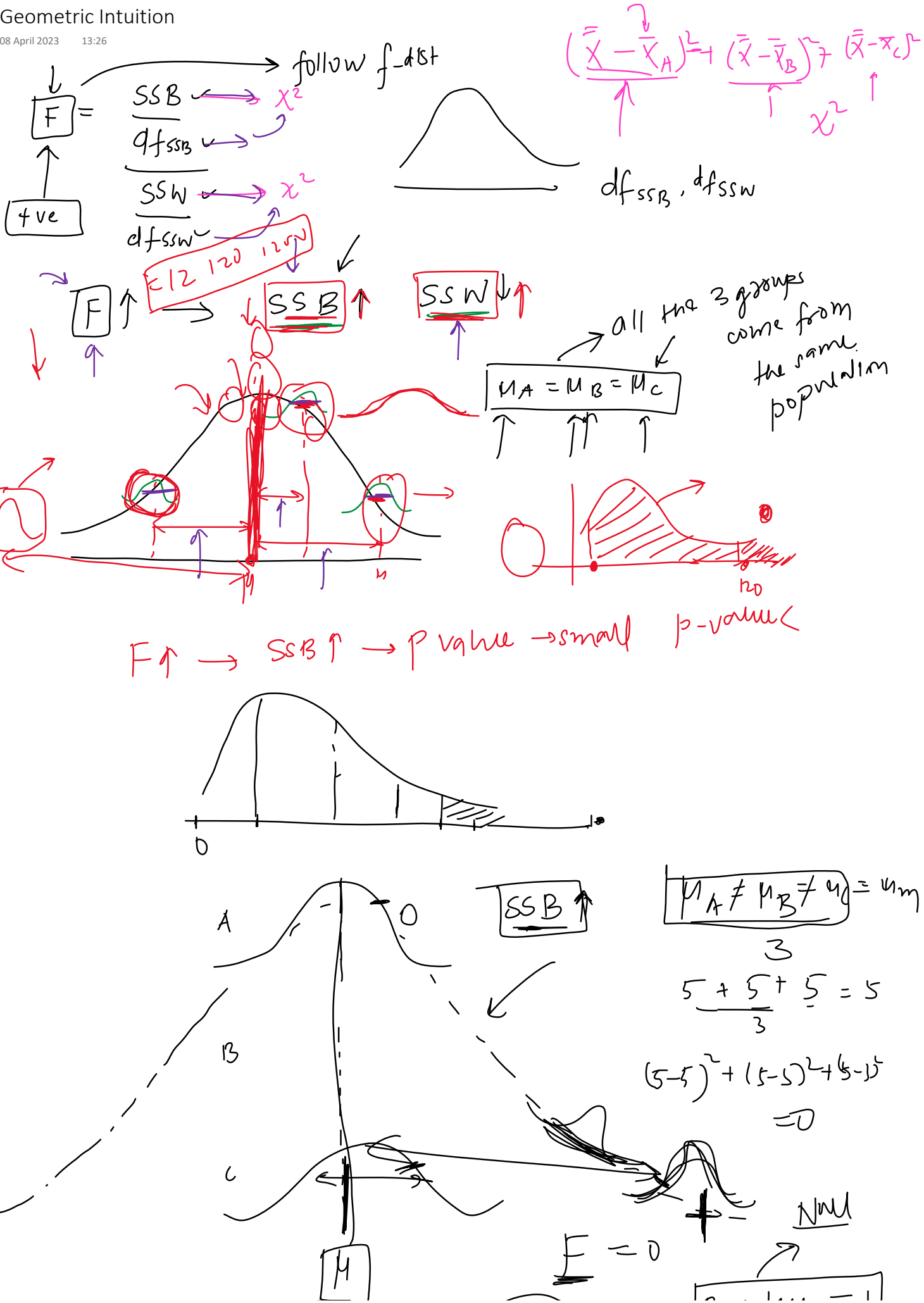
p-value = 0.31 > α (0.05)

Null hy can't reject

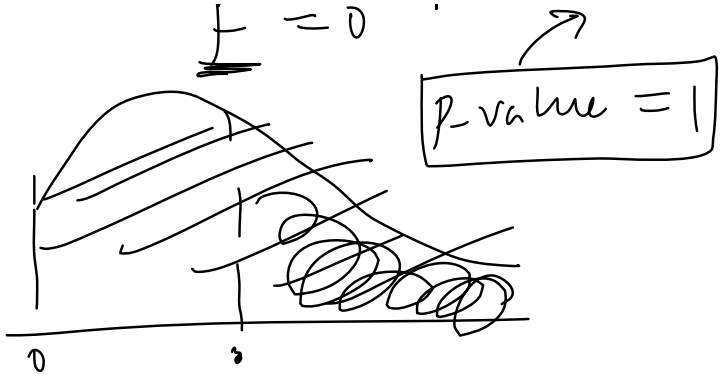
$\mu_A = \mu_B = \mu_C$

Geometric Intuition

08 April 2023 13:26



H_1



Assumptions

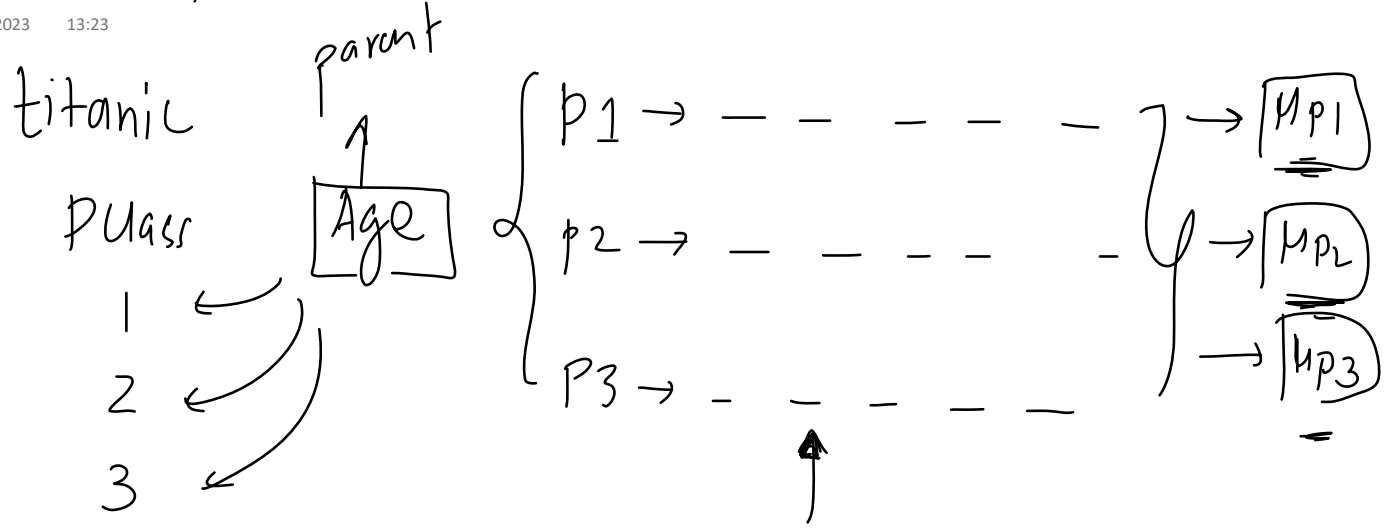
08 April 2023 16:48

Assumptions

1. Independence: The observations within and between groups should be independent of each other. This means that the outcome of one observation should not influence the outcome of another. Independence is typically achieved through random sampling or random assignment of subjects to groups.
2. Normality: The data within each group should be approximately normally distributed. While one-way ANOVA is considered to be robust to moderate violations of normality, severe deviations may affect the accuracy of the test results. If normality is in doubt, non-parametric alternatives like the Shapiro-wilk test can be considered.
3. Homogeneity of variances: The variances of the populations from which the samples are drawn should be equal, or at least approximately so. This assumption is known as homoscedasticity. If the variances are substantially different, the accuracy of the test results may be compromised. Levene's test or Bartlett's test can be used to assess the homogeneity of variances. If this assumption is violated, alternative tests such as Welch's ANOVA can be used.

Python Case Study

08 April 2023 13:23



Post-hoc Test

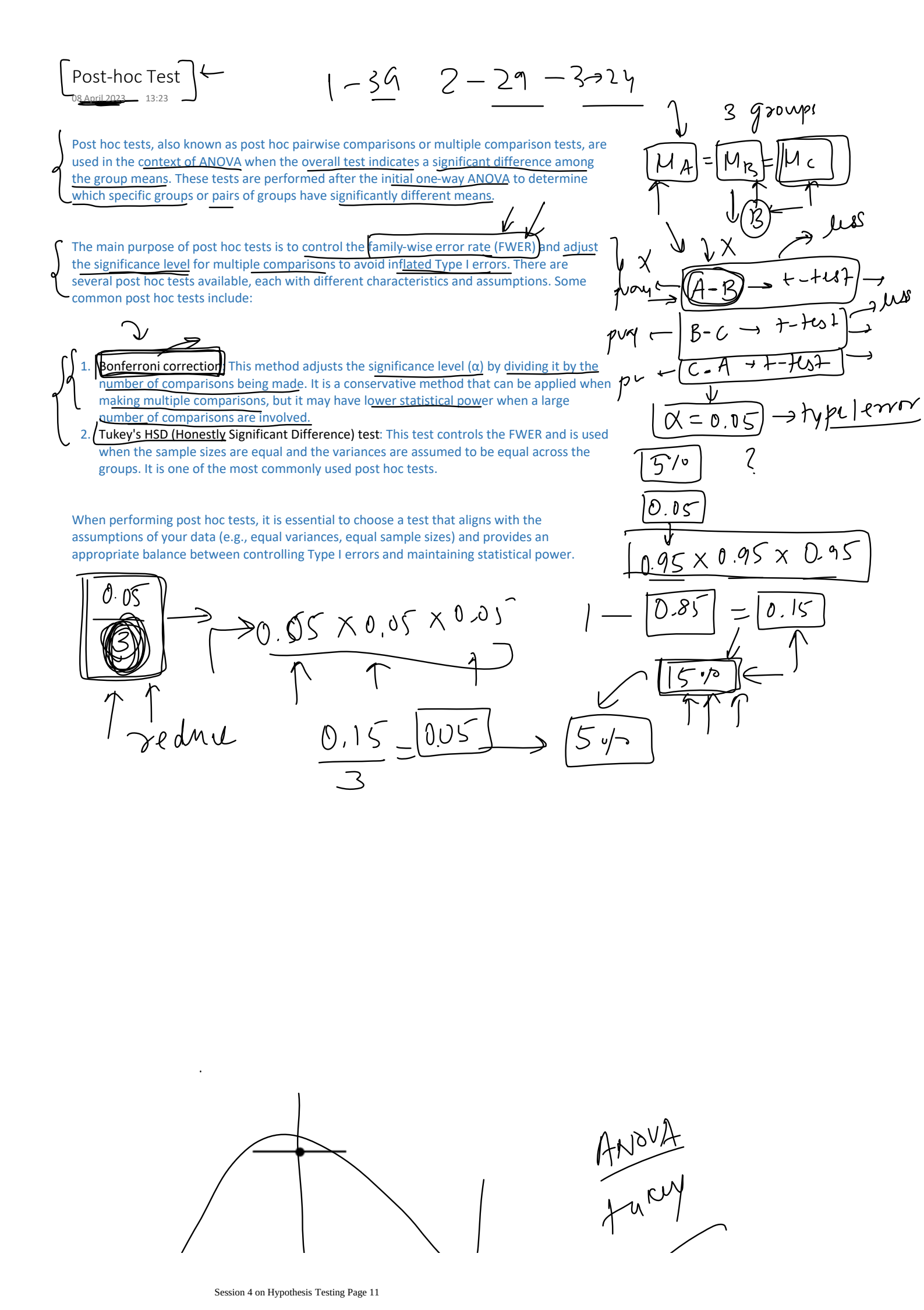
08 April 2023 13:23

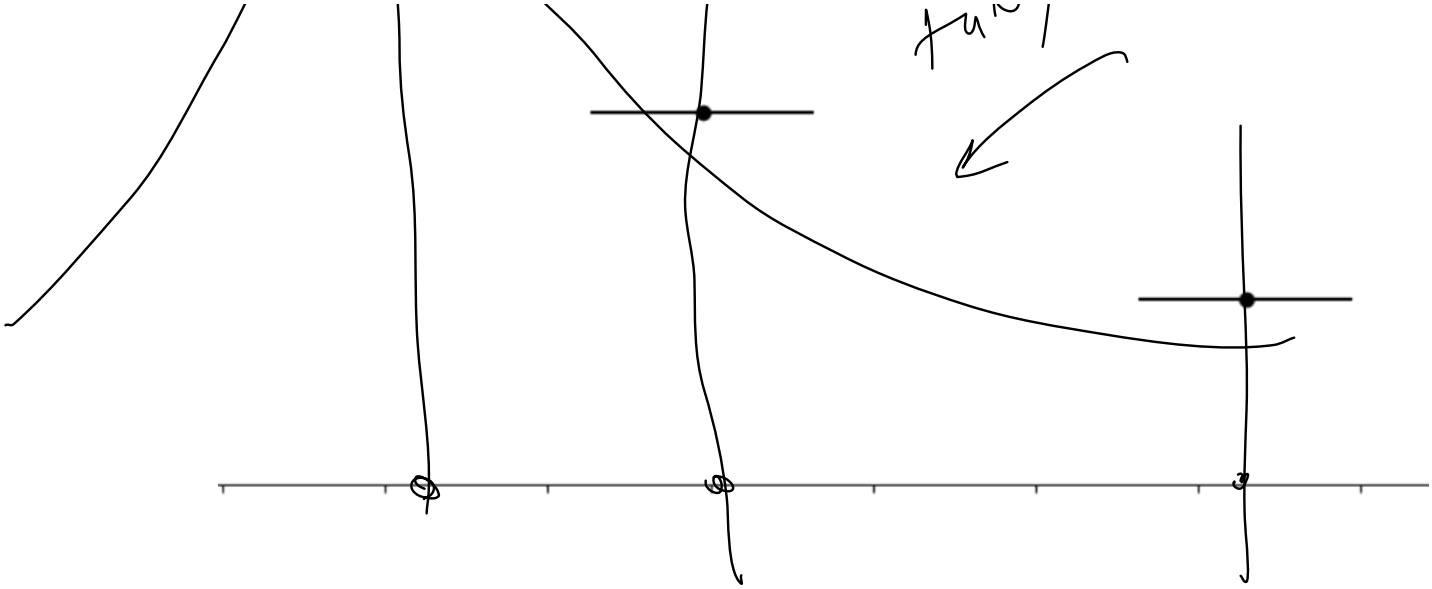
Post hoc tests, also known as post hoc pairwise comparisons or multiple comparison tests, are used in the context of ANOVA when the overall test indicates a significant difference among the group means. These tests are performed after the initial one-way ANOVA to determine which specific groups or pairs of groups have significantly different means.

The main purpose of post hoc tests is to control the family-wise error rate (FWER) and adjust the significance level for multiple comparisons to avoid inflated Type I errors. There are several post hoc tests available, each with different characteristics and assumptions. Some common post hoc tests include:

1. Bonferroni correction: This method adjusts the significance level (α) by dividing it by the number of comparisons being made. It is a conservative method that can be applied when making multiple comparisons, but it may have lower statistical power when a large number of comparisons are involved.
2. Tukey's HSD (Honestly Significant Difference) test: This test controls the FWER and is used when the sample sizes are equal and the variances are assumed to be equal across the groups. It is one of the most commonly used post hoc tests.

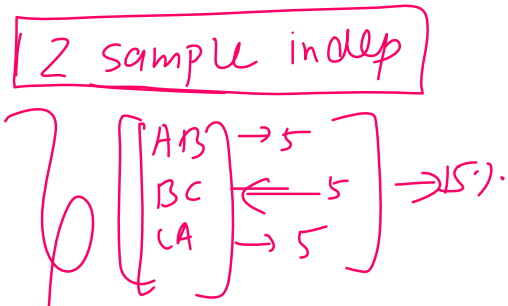
When performing post hoc tests, it is essential to choose a test that aligns with the assumptions of your data (e.g., equal variances, equal sample sizes) and provides an appropriate balance between controlling Type I errors and maintaining statistical power.





Why t-test is not used for more than 3 categories?

08 April 2023 13:23



1. Increased Type I error: When you perform multiple comparisons using individual t-tests, the probability of making a Type I error (false positive) increases. The more tests you perform, the higher the chance that you will incorrectly reject the null hypothesis in at least one of the tests, even if the null hypothesis is true for all groups.
2. Difficulty in interpreting results: When comparing multiple groups using multiple t-tests, the interpretation of the results can become complicated. For example, if you have 4 groups and you perform 6 pairwise t-tests, it can be challenging to interpret and summarize the overall pattern of differences among the groups.
3. Inefficiency: Using multiple t-tests is less efficient than using a single test that accounts for all groups, such as one-way ANOVA. One-way ANOVA uses the information from all the groups simultaneously to estimate the variability within and between the groups, which can lead to more accurate conclusions.

Applications in Machine Learning

08 April 2023 13:27

1. Hyperparameter tuning: When selecting the best hyperparameters for a machine learning model, one-way ANOVA can be used to compare the performance of models with different hyperparameter settings. By treating each hyperparameter setting as a group, you can perform one-way ANOVA to determine if there are any significant differences in performance across the various settings.

2. Feature selection: One-way ANOVA can be used as a univariate feature selection method to identify features that are significantly associated with the target variable, especially when the target variable is categorical with more than two levels. In this context, the one-way ANOVA is performed for each feature, and features with low p-values are considered to be more relevant for prediction.

3. Algorithm comparison: When comparing the performance of different machine learning algorithms, one-way ANOVA can be used to determine if there are any significant differences in their performance metrics (e.g., accuracy, F1 score, etc.) across multiple runs or cross-validation folds. This can help you decide which algorithm is the most suitable for a specific problem.

4. Model stability assessment: One-way ANOVA can be used to assess the stability of a machine learning model by comparing its performance across different random seeds or initializations. If the model's performance varies significantly between different initializations, it may indicate that the model is unstable or highly sensitive to the choice of initial conditions.