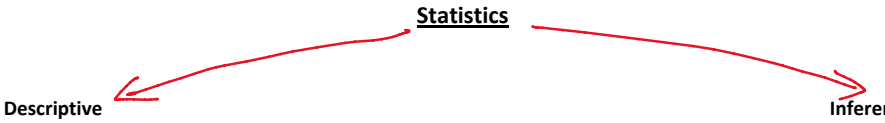
Descriptive Statistics

Wednesday, October 11, 2023 1:47 PM

What is Statistics ?
Statistics is a branch of mathematics that involves collecting, analyzing, 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, analyze 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



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

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

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.

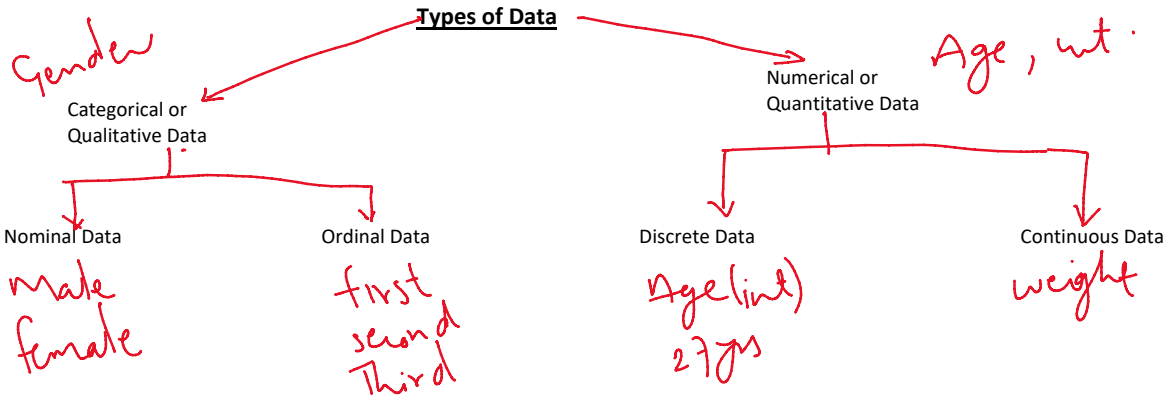
4. ***Regression analysis:*** This involves modelling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a product based on advertising expenditure.

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?



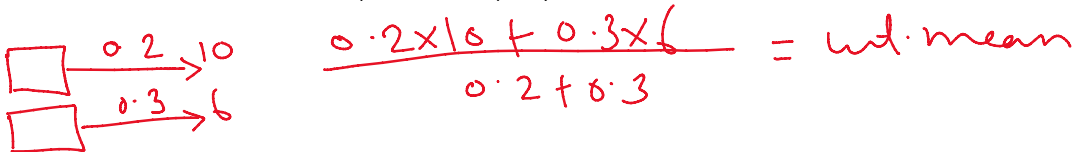
Measure of Central Tendency

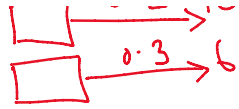
A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.

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

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

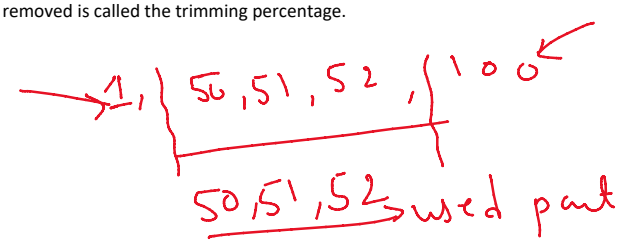
2. ***Median***: The median is the middle value in the dataset when the data is arranged in order.
3. ***Mode***: The mode is the value that appears most frequently in the dataset.
4. ***Weighted Mean:*** The weighted mean is the sum of the products of each value and its weight, divided by the sum of the weights. It is used to calculate a mean when the values in the dataset have different importance or frequency.





0.2 + 0.3

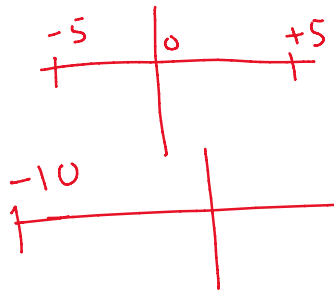
5. ***Trimmed mean*** is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called the trimming percentage.



1, 100 are trimmed
10% trim left & right.

Measure of Dispersion

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.



$\bar{x} = 0$

$= \bar{x} = 0$

Though, $\bar{x} = 0$ on both but dispersion of data is different.

1. ***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.

$R = L - S$

2. ***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.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

Mean Absolute Deviation

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

→ MAD not work properly in inferential statistics.

3. ***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.

Population Standard Deviation

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

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

SD unit same of data.

© uedufy.com

4. ***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

Categorical **Numerical**

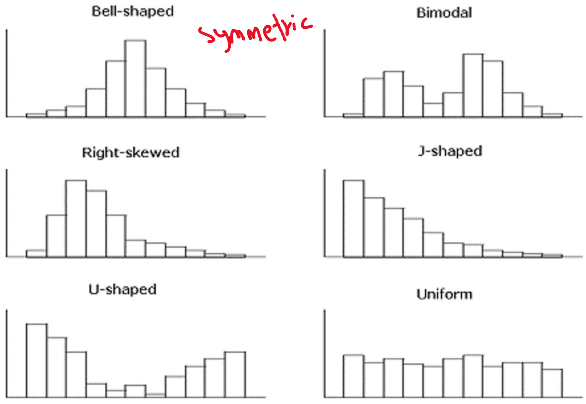
1. Categorical - Frequency Distribution Table & Cumulative Frequency

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

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.

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.

2. Numerical - Frequency Distribution Table & Histogram

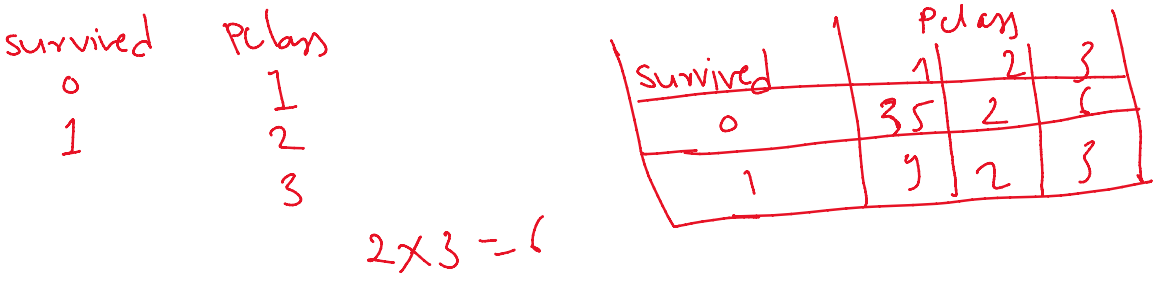


Graphs for Bivariate Analysis

Categorical-Categorical Categorical-Numerical Numerical-Numerical

1. Categorical - Categorical

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. Numerical - Numerical

Scatter Plot

2. Numerical - Numerical

Scatter Plot

3. Categorical - Numerical

	0-10	20-30	30-40	40-50	
M	60	20	25	15	15
f	10	30	15	6	32

Even we can convert integer into bin.

Quantiles and Percentiles

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

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

Things to remember while calculating these measures:

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

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

$PL = \frac{P}{100}(N+1) = \frac{75}{100}(10+1) = \frac{33}{4} = 8.25$ → 96 + 0.25 × (98 − 96) = 96.5

Percentile of Value :

Percentile of a value

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

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

percentile rank = $\frac{3 + 0.5 \times 1}{10} = \frac{3.5}{10} = 0.35$ → 35th.

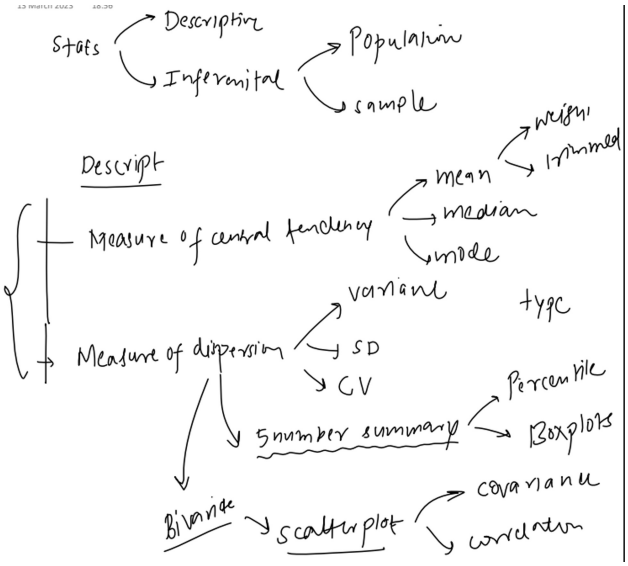
3 ... and the number 88.

10

3 no. are beyond the number 88.

Descriptive Statistics II

Monday, October 16, 2023 7:34 AM

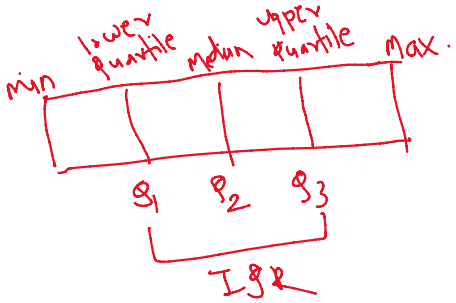


5 number summary

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.



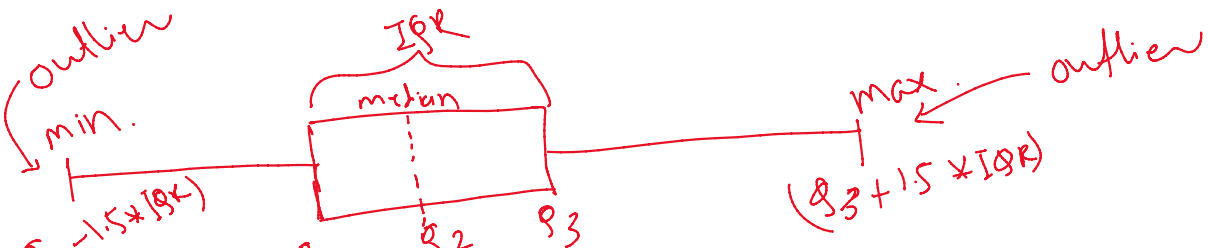
Interquartile Range (IQR)

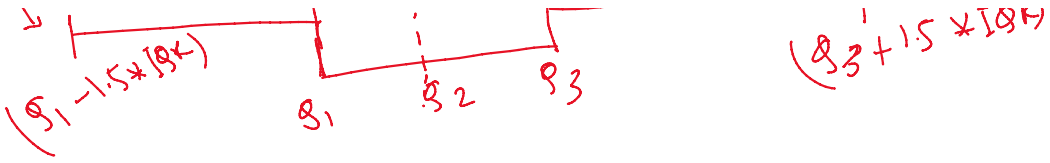
The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.

Boxplots

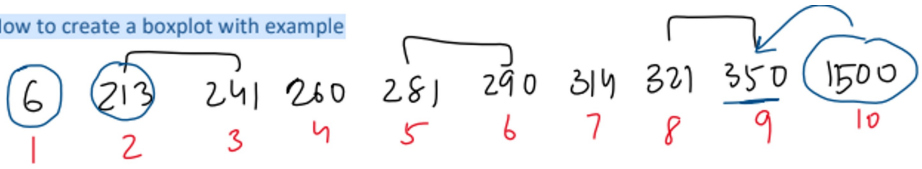
1. What is a boxplot

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



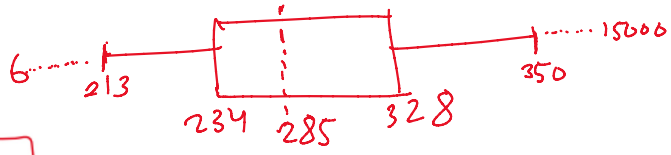
$Q_2 = \frac{50}{100}(11) = 5.5 = 285.5$

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

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

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

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



$IQR = 94$

MIN & MAX

$MIN = Q_1 - 1.5 * IQR$

$= 93$

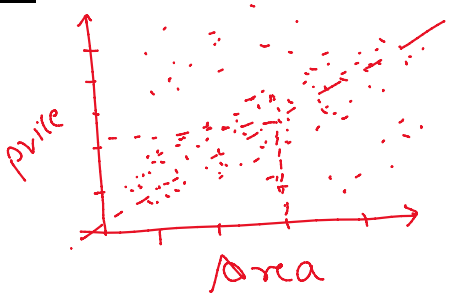
$MAX = Q_3 + 1.5 \times IQR$

$= 469$

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

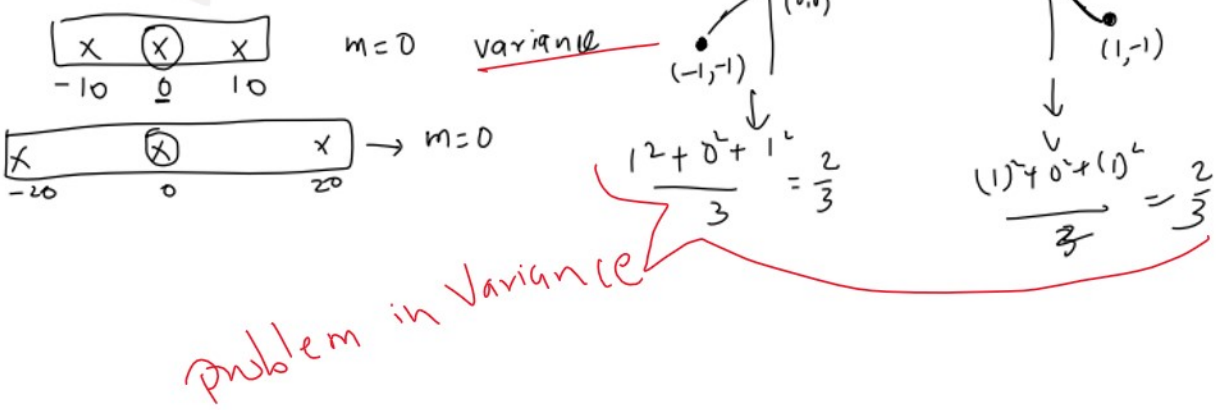
Scatterplot



Area	price
100	1100k
200	2200k
300	3300k

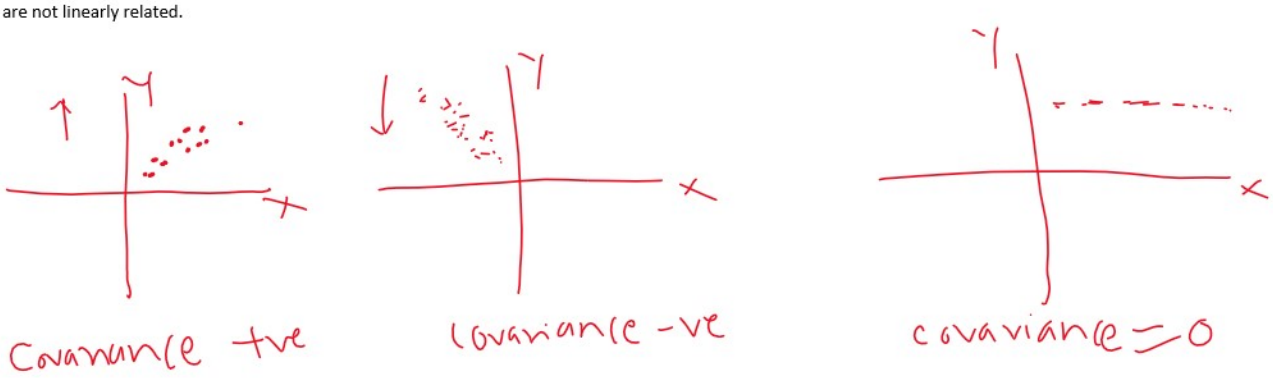
Covariance

- 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?

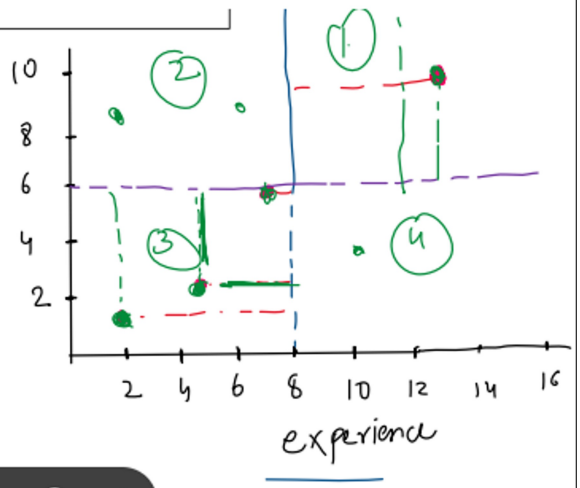
Covariance Formula	
Population	**Sample**
$\sigma_{xy} = \frac{\sum(X-\mu_x)(Y-\mu_y)}{N}$	$s_{xy} = \frac{\sum(X-\bar{x})(Y-\bar{y})}{n-1}$
X, Y – The Value of X and Y in the Population *μ_x, μ_y – The population Mean of X and Y* *N – Total Number of Observations*	*X, Y – The Value of X and Y in the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

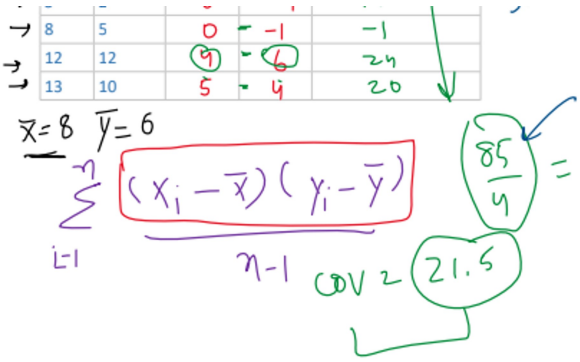
Exp(x)	Salary(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	-1
12	12	4	6	24
13	10	5	4	20

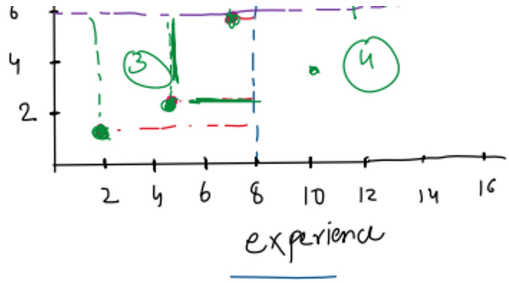
$\bar{X} = 8 \quad \bar{Y} = 6$

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

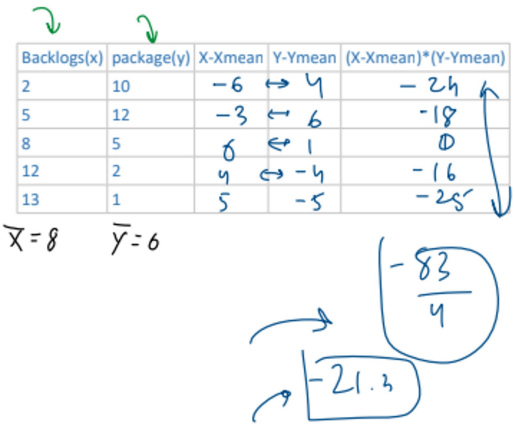
$\frac{85}{4} =$ 21.5

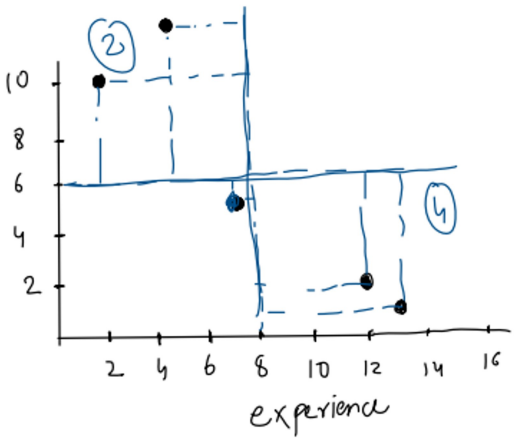






Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10	-6	4	-24
5	12	-3	6	-18
8	5	0	1	0
12	2	4	-4	-16
13	1	5	-5	-25





• 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

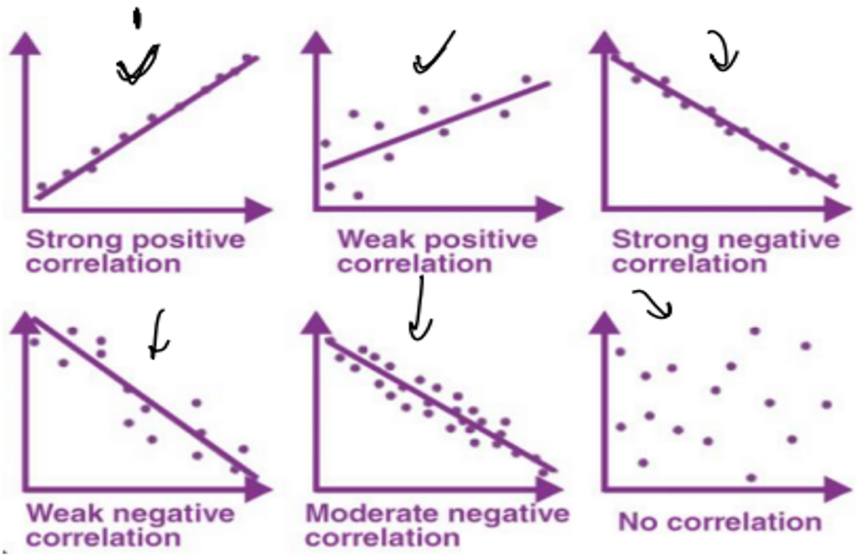
$$\sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

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

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

Correlation

1. What problem does Correlation solve?



Can we quantify this weak and strong relationship?

2. What is correlation?

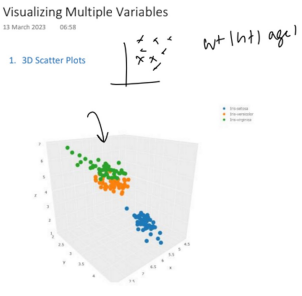
Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together. Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 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.

$$\text{Correlation} = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$

Correlation and Causation

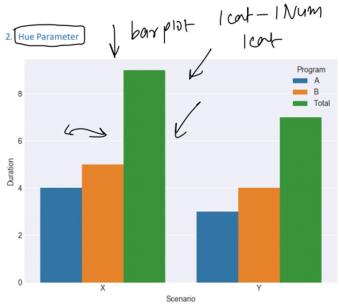
The phrase "correlation does not imply causation" means that just because two variables are associated with each other, it does not necessarily mean that one causes the other. In other words, a correlation between two variables does not necessarily imply that one variable is the reason for the other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

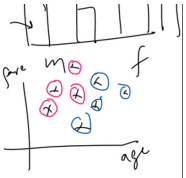
Visualizing Multiple Variables



2. Hue Parameter

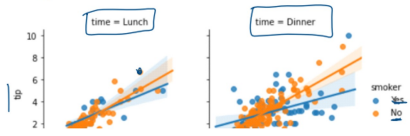
barplot 1cat – 1Num 1cat

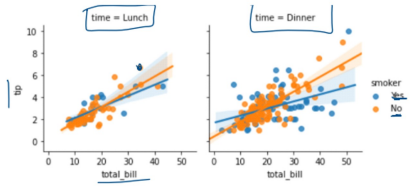




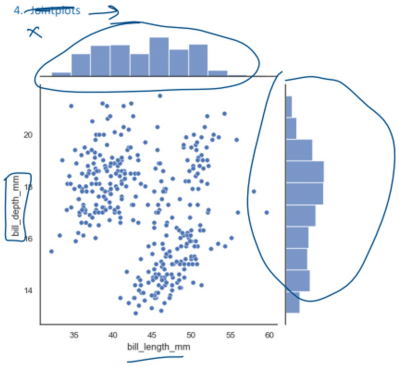
4 cols
4 rows

3. Facetgrids





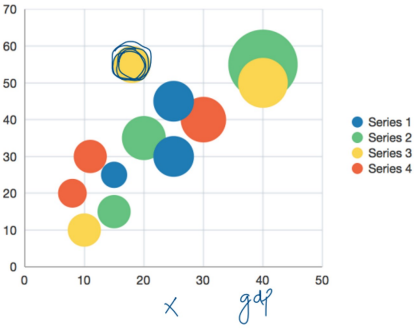
4. Jointplots



6. Bubble Plots

Bubble Chart (pairwise here)

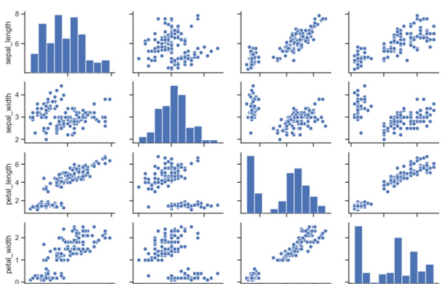
pop

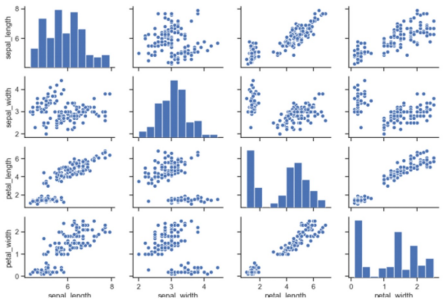


Y

X gdp

Pairplot




sepal_length
sepal_width
petal_length
petal_width
sepal_length
sepal_width
petal_length
petal_width

PDF-CDF-PMF_Descriptive

Friday, October 20, 2023 5:01 PM

Random Variables

15 March 2023 11:43

- What are Algebraic Variables?

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

  $x + 5 = 10 \Rightarrow x = 5$

  Die: 1 – 4, 2 – 5, 3 – 6

- What are Random Variables in Stats and Probability?

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

  Coin toss: H, T

$X = \{1, 0\}$ $Y = \{1, 2, 3, 4, 5, 6\}$

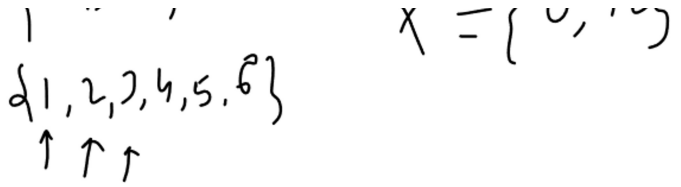
H = 1 T = 0 randomly → sample space

X Y Z x, y, z

- Types of Random Variables?

Discrete RV $\{H, T\}$ $\{1, 2, 3, 4, 5, 6\}$

Continuous RV $X = \{0, 10\}$

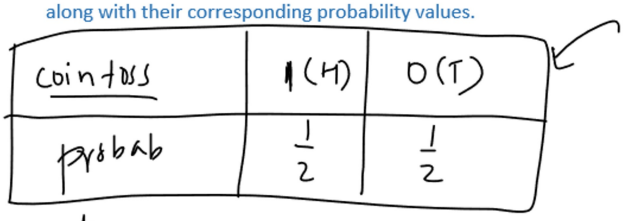


Probability Distributions

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.



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 dice →

2 3 4 5 6 7 8 9

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

Problem with Distribution?

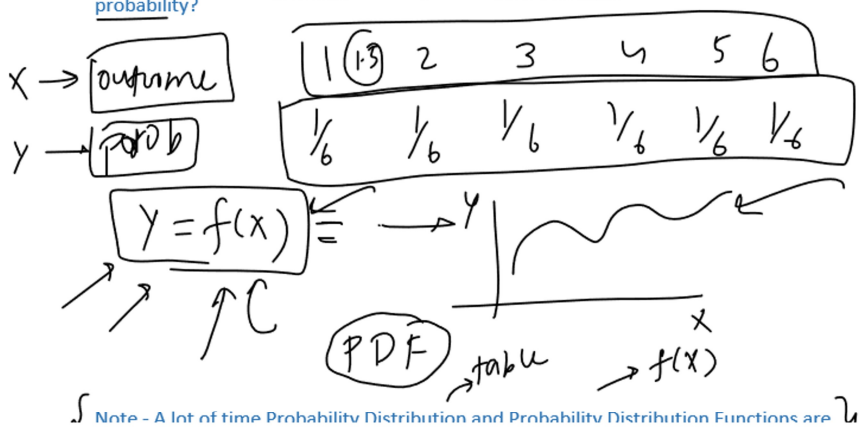
In many scenarios, the number of outcomes can be much larger and hence a table would be tedious to write down. Worse still, the number of possible outcomes could be infinite, in which case, good luck writing a table for that.

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

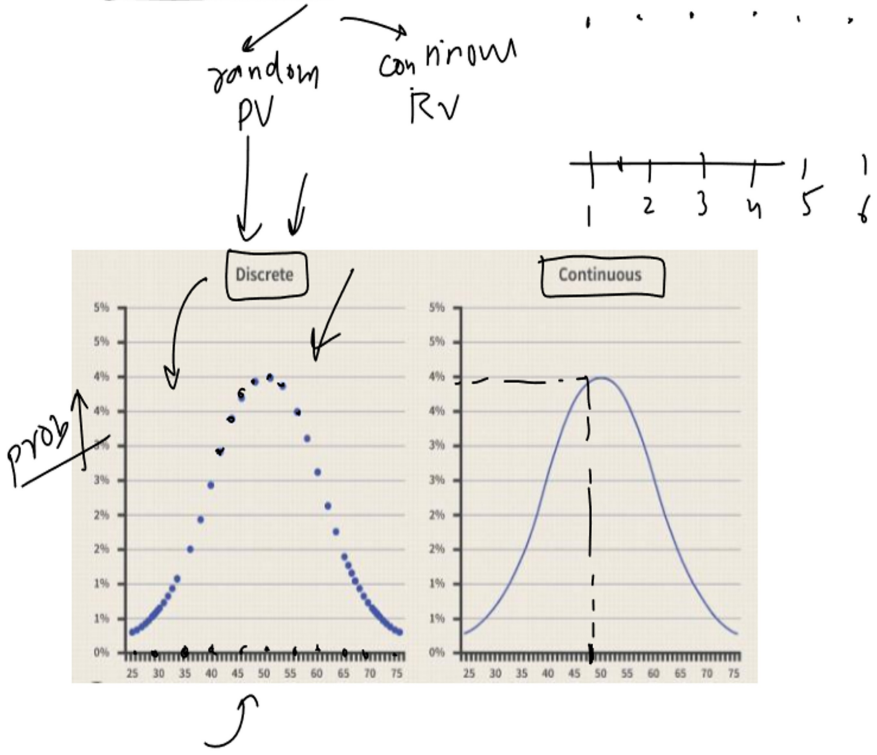
→ Solution - Function?

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

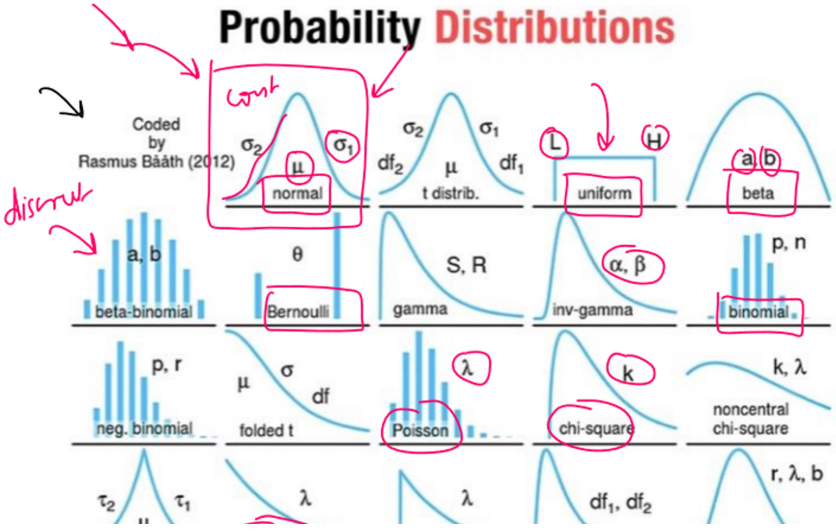


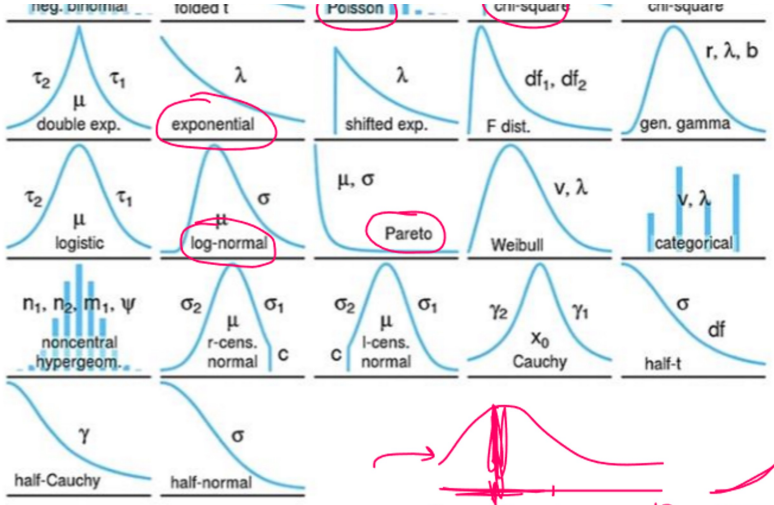
Note - A lot of time Probability Distribution and Probability Distribution Functions are used interchangeably.

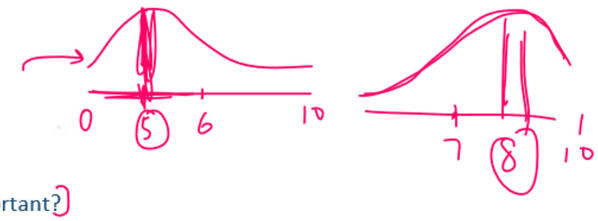
1. Types of Probability Distributions (PDF)



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.

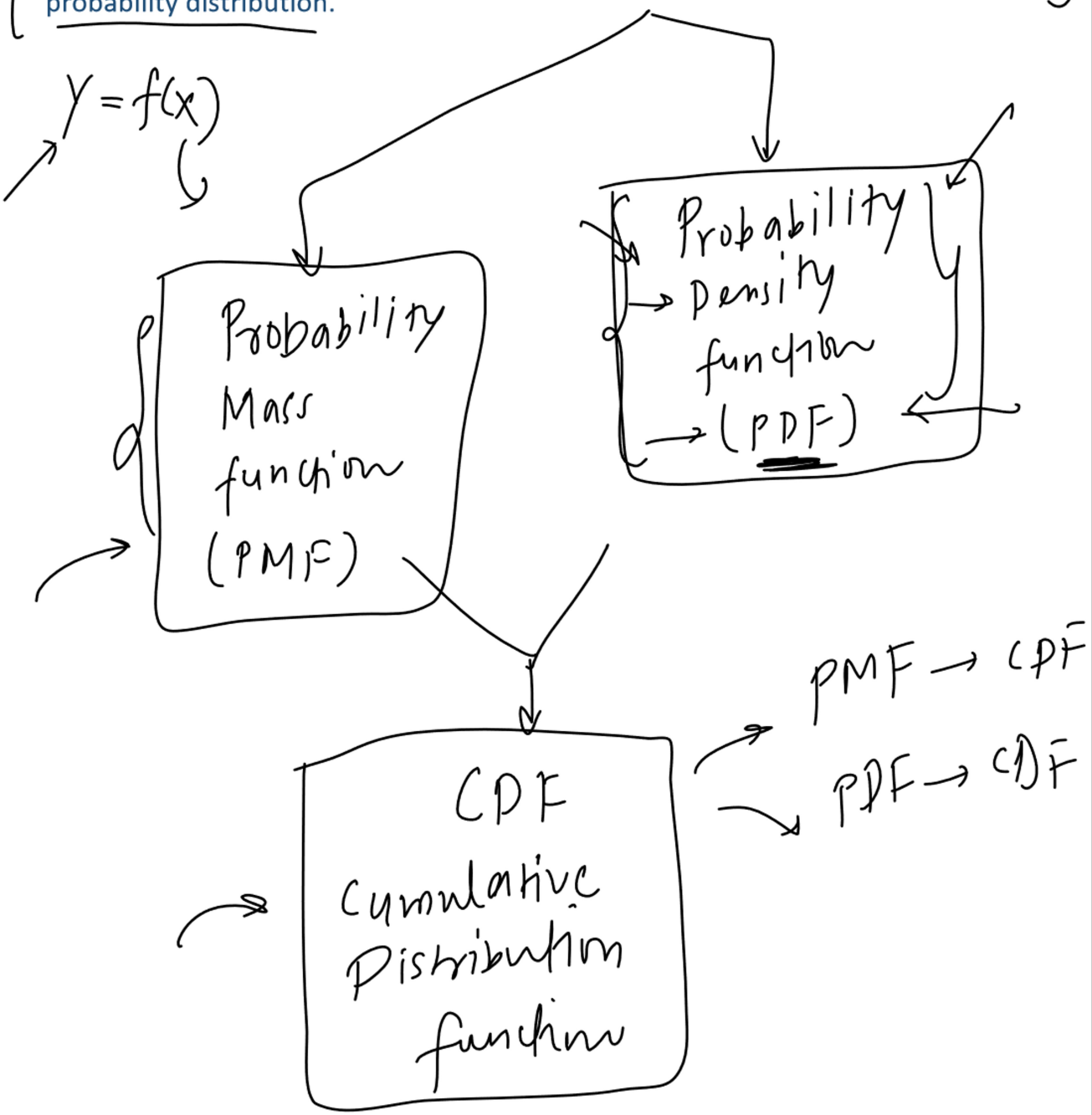
Different probability distributions have different sets of parameters that determine their shape and characteristics, and understanding these parameters is essential in statistical analysis and inference.

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.

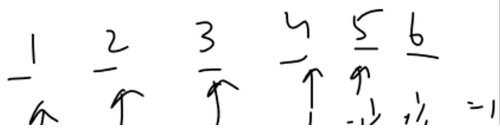
probability distribution.



Probability Mass Function (PMF)

15 March 2023 15:25

PMF stands for Probability Mass Function. It is a mathematical function that

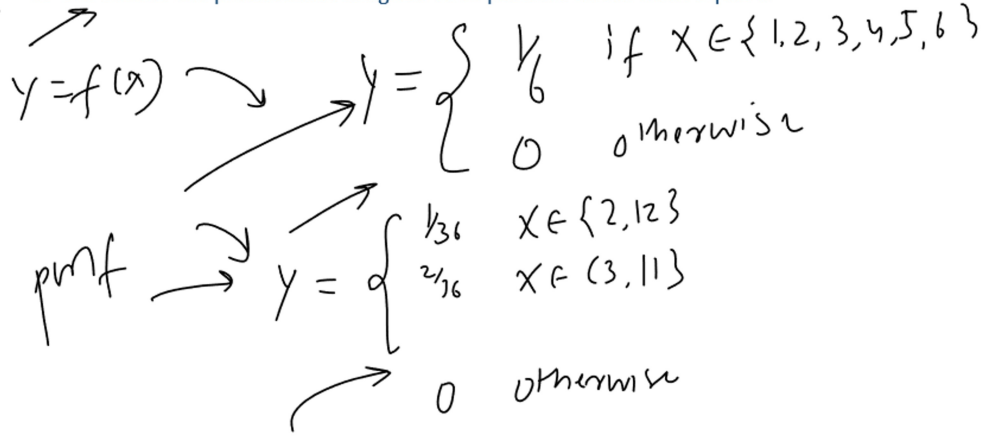


15 March 2023 15:25

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)]

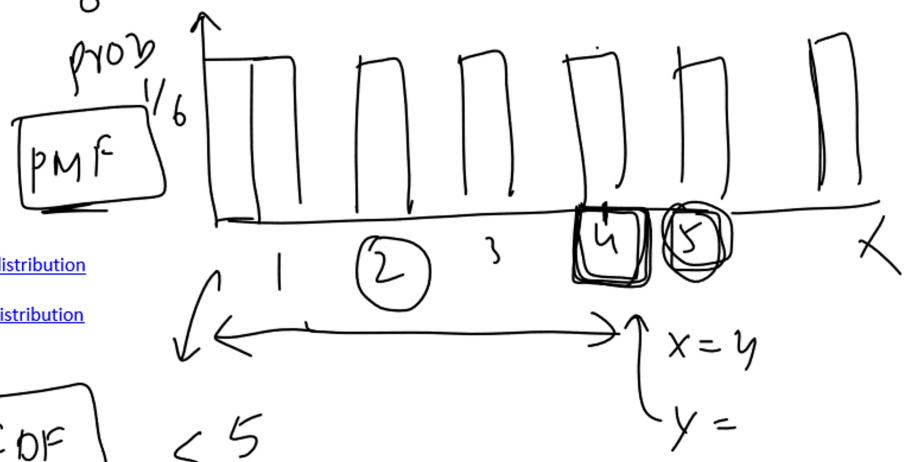
$f(x) =$ PMF $f(x)$

$f(X=6) = 1/6$

$f(X=5) = \frac{1}{6}$

$f(X \le 4) = f(X=4) + f(X=3) + f(X=2) + f(X=1)$

$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$



Examples:

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

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

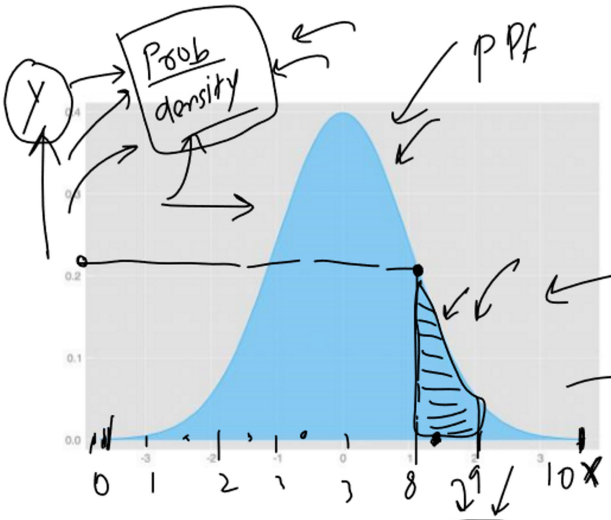
CDF ≤ 5

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

7.912345...9

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

area

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

8 → 9

probability between

8 → 8.1

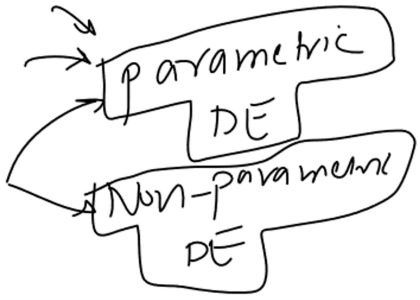
8 → 8.01

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.

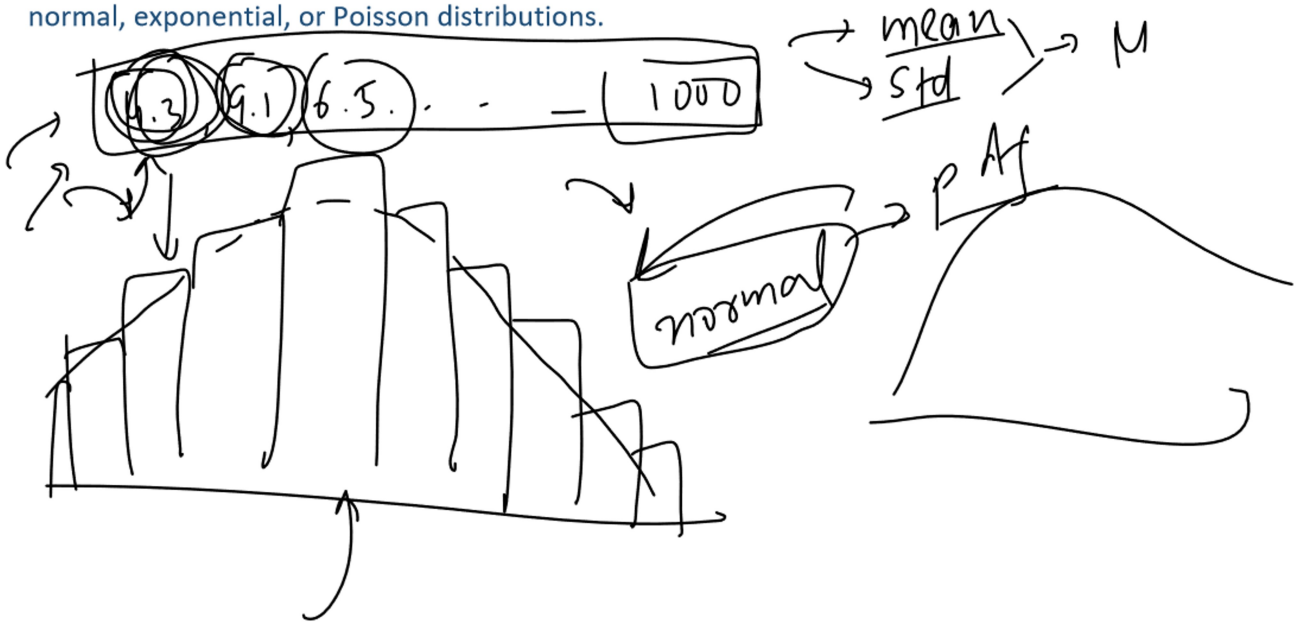
There are various methods for density estimation, including **parametric** and **non-parametric approaches**. Parametric methods assume that the data follows a specific probability distribution (such as a normal distribution), while non-parametric methods do not make any assumptions about the distribution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

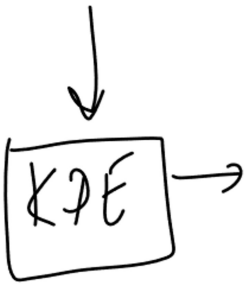
But sometimes the distribution is not clear or it's not one of the famous distributions.

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.

estimation can be **computationally intensive and may require more data to achieve accurate estimates** compared to parametric methods.



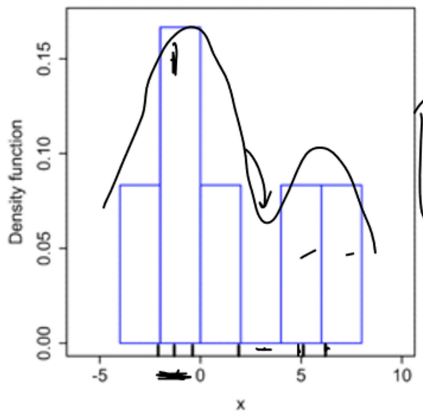
Kernel Density Estimate(KDE)

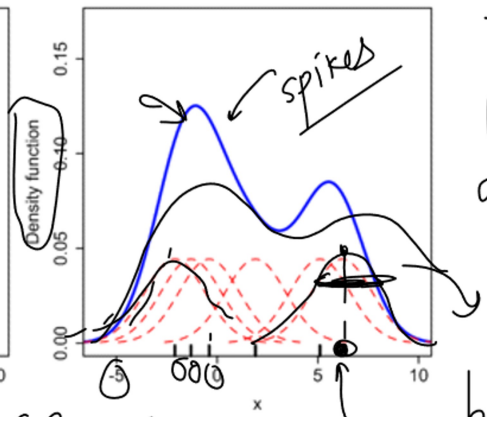
16 March 2023 16:08

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

6 datapoints

KDE





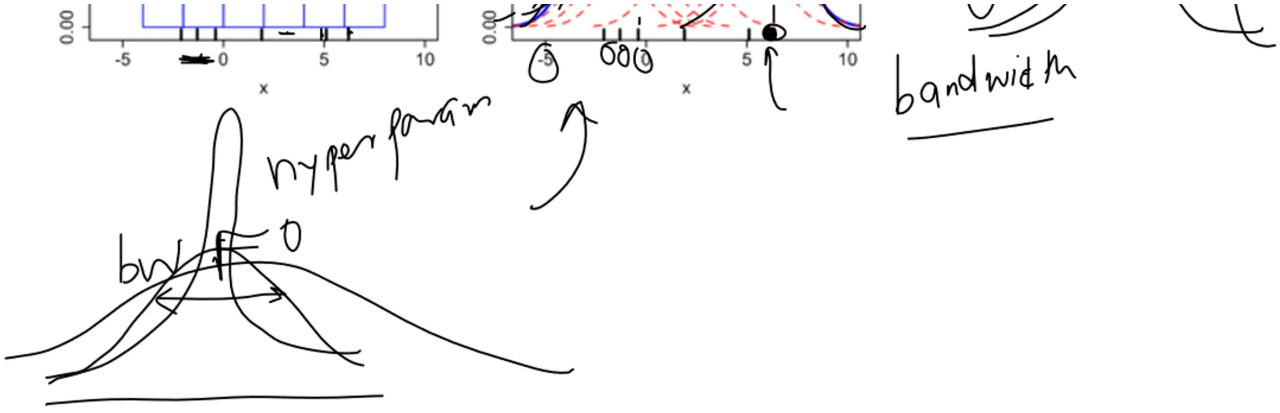
Kernel → prob

gaussian

μ = mean

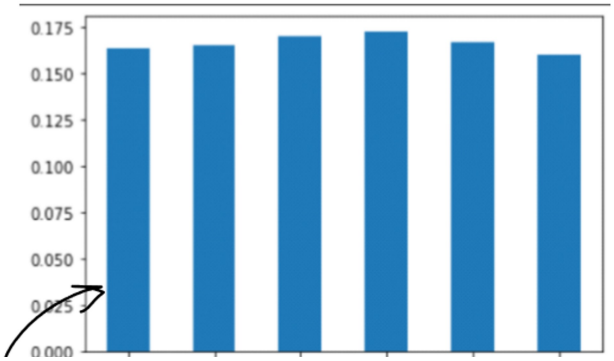
σ = std

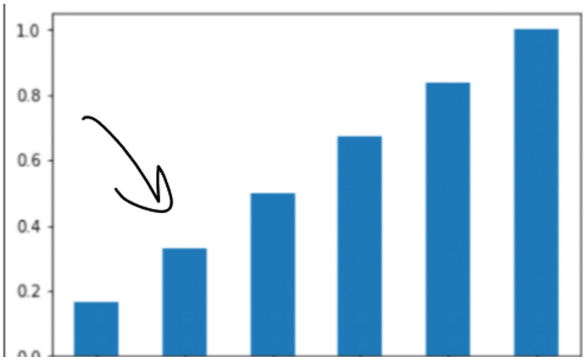
Bandwidth

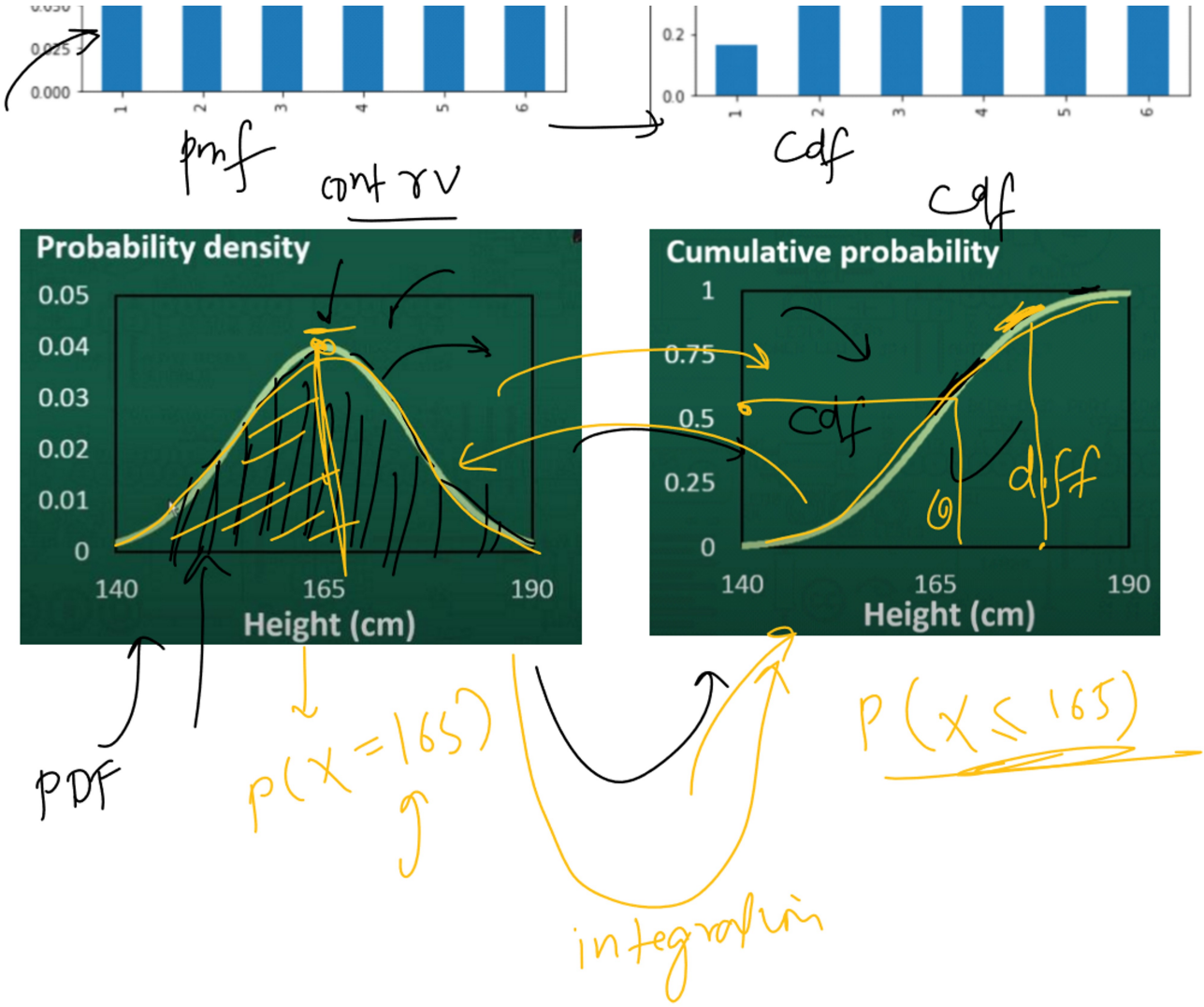


Cumulative Distribution Function(CDF) of PDF

15 March 2023 15:25






pmf
cont rv
cdf
cdf
Probability density
0.05
0.04
0.03
0.02
0.01
0
140
165
190
Height (cm)
Cumulative probability
1
0.75
0.5
0.25
0
cdf
diff
140
165
190
Height (cm)
PDF
p(X=165)
integration
P(X≤165)