Mastering the Fundamentals of Statistics for Data Science -Basic to Advance Level- Part 1

Ritesh Gupta · Follow

11 min read · Jan 28

Statistics:

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. In data science, statistics play a crucial role in understanding and making sense of large sets of data. Statistics helps data scientists to understand patterns and trends in data, and make predictions based on that data. It allows data scientists to make inferences about a population based on a sample of data, and to test hypotheses about relationships between variables.

In data science, statistics is used in a wide range of tasks, such as data exploration, data cleaning, feature selection, model selection and evaluation, and hypothesis testing.

For example, when exploring a dataset, a data scientist might use descriptive statistics, such as mean, median, and standard deviation, to get a sense of the overall distribution of the data. In the process of cleaning and preparing the data, statistical techniques such as outliers detection, missing value imputation, and normalization may be used.

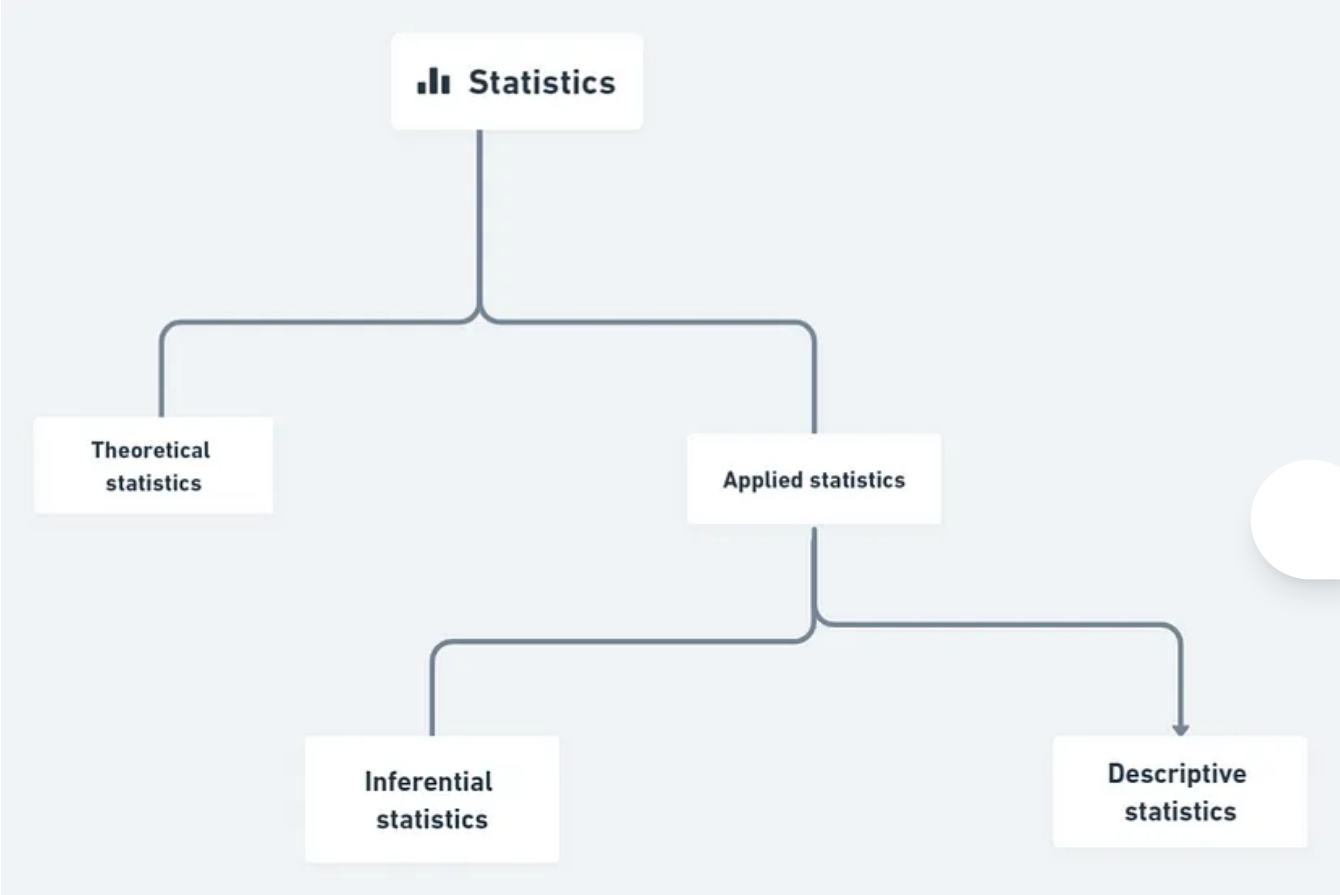
In the feature selection process, a data scientist can use statistical methods such as correlation and regression analysis to identify which variables are most important for the problem at hand.

When it comes to model selection, data scientists use statistical techniques such as cross-validation and A/B testing to ensure that their model is generalizable and performs well on unseen data.

In short, statistics is an essential tool for data scientists, as it provides the means to extract insights and make predictions from data.

Types of Statistics:

- Theoretical Statistics
- Applied Statistics



Inferential statistics:

Inferential statistics is a branch of statistics that deals with making inferences about a population based on a sample of data. It involves using probability and statistical models to make predictions or estimates about a population from sample data. This is done by using statistical methods to draw conclusions about the population from the sample data. Some common inferential statistical techniques include hypothesis testing, estimation, and correlation analysis. These techniques are used to make predictions, estimate population parameters, and draw conclusions about the population based on the sample data.

Here is an example of inferential statistics in action:

Imagine a researcher wants to know the average income of all adults in a certain city. It would be impractical to survey every adult in the city, so instead they take a sample of 1000 adults and record their income. Using inferential statistics, the researcher can use this sample data to make inferences about the entire population of adults in the city.

The researcher calculates the mean income of the sample, which is $50,000. The researcher also calculates the standard deviation of the sample, which is $10,000. With this information, the researcher can use inferential statistical methods to

estimate the mean income of all adults in the city and also calculate the margin of error.

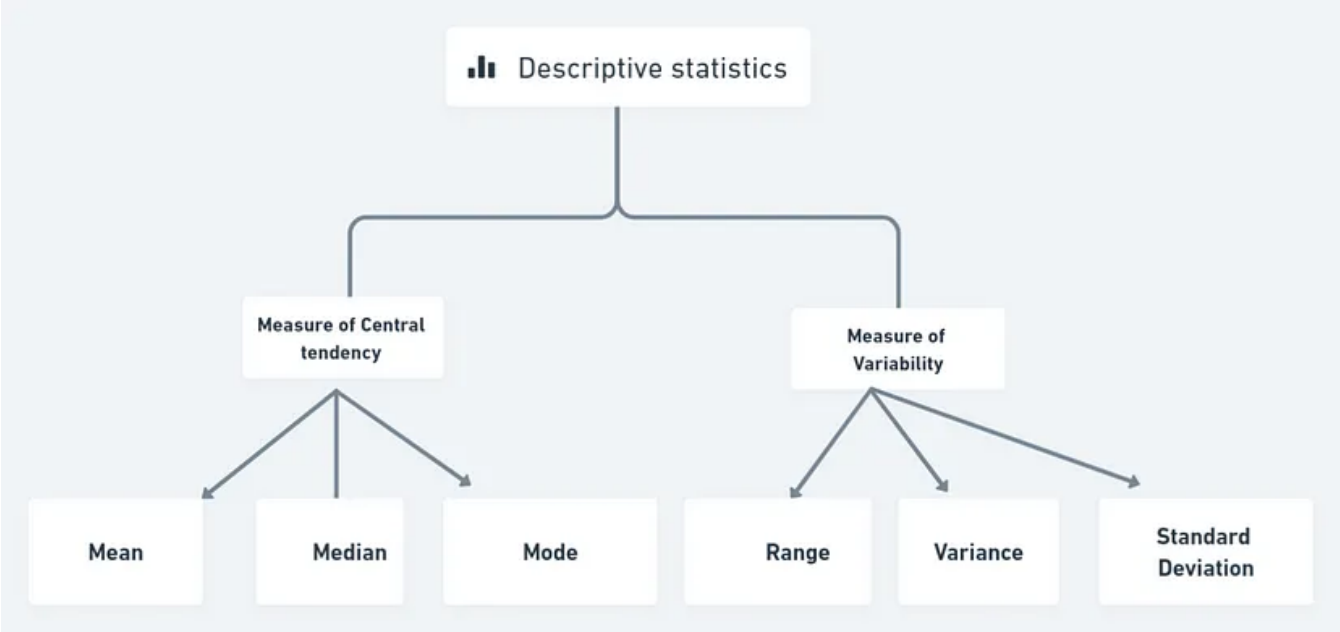
For example, using a z-test and assuming a normal distribution, the researcher can estimate that the average income of all adults in the city is likely between $49,500 and $50,500, with a 95% level of confidence.

Descriptive statistics:

Descriptive statistics is a branch of statistics that deals with describing and summarizing the characteristics of a sample of data. It is used to describe and summarize the sample data, and does not involve making inferences about a population.

Types of Descriptive Statistics:

- Measure of Central Tendency
- Measure of Variability

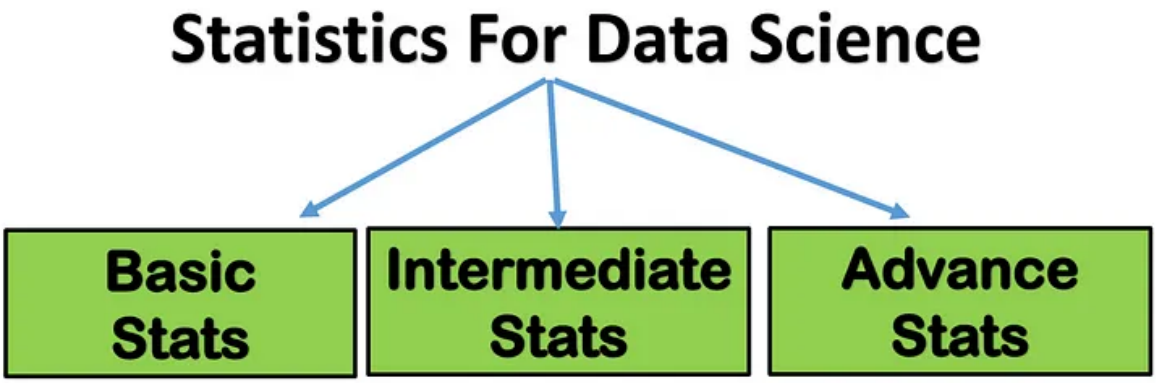


An example of descriptive statistics in action is:

A researcher wants to describe the height of students in a certain school. They measure the height of a sample of 100 students. Using descriptive statistics, the researcher can summarize the height data by calculating measures such as the mean, median, and mode of the sample. They can also create a frequency distribution table and a histogram to show the distribution of the height data.

The researcher finds that the mean height is 5.6 feet, the median height is 5.5 feet, and the mode height is 5.6 feet. They also find that most of the students have a height between 5.4 and 5.8 feet.

This is just one example of how descriptive statistics can be used to describe and summarize data.



Basic Statistics :

1. Probability Introduction
2. Addition Rule in Probability
3. Multiplication Rule in Probability
4. Population And Sample
5. Measure Of Central Tendency(Mean, Median, Mode)
6. Measure Of Dispersion(Variance, Standard Deviation)
7. What is Sampling Method And Its Types
8. Population Mean And Sample Mean
9. What is Variables And Its Types?
10. Frequency Distribution And Cumulative Frequency

11. Histograms

Intermediate Statistics:

1. Percentiles And Quantiles
2. Five Number Summary
3. Inter Quartile Range(IQR)
4. Boxplots
5. Effect Of Outliers And It’s Removal
6. Probability Density Function
7. Normal Distribution or Gaussian Distribution And Emperical Formula
8. Z score
9. Standardization Vs Normalization
10. Standard Normal Distribution
11. Central Limit Theorem
12. Chebyshevs Inequality
13. Covariance
14. Pearson Correlation Coefficient

Advance Statistics:

1. QQ plot
2. Bernoulli Distribution And Binomial Distribution
3. Log Normal Distribution
4. Power Law Distribution
5. Boxcox Tranform
6. All Transformation Techniques

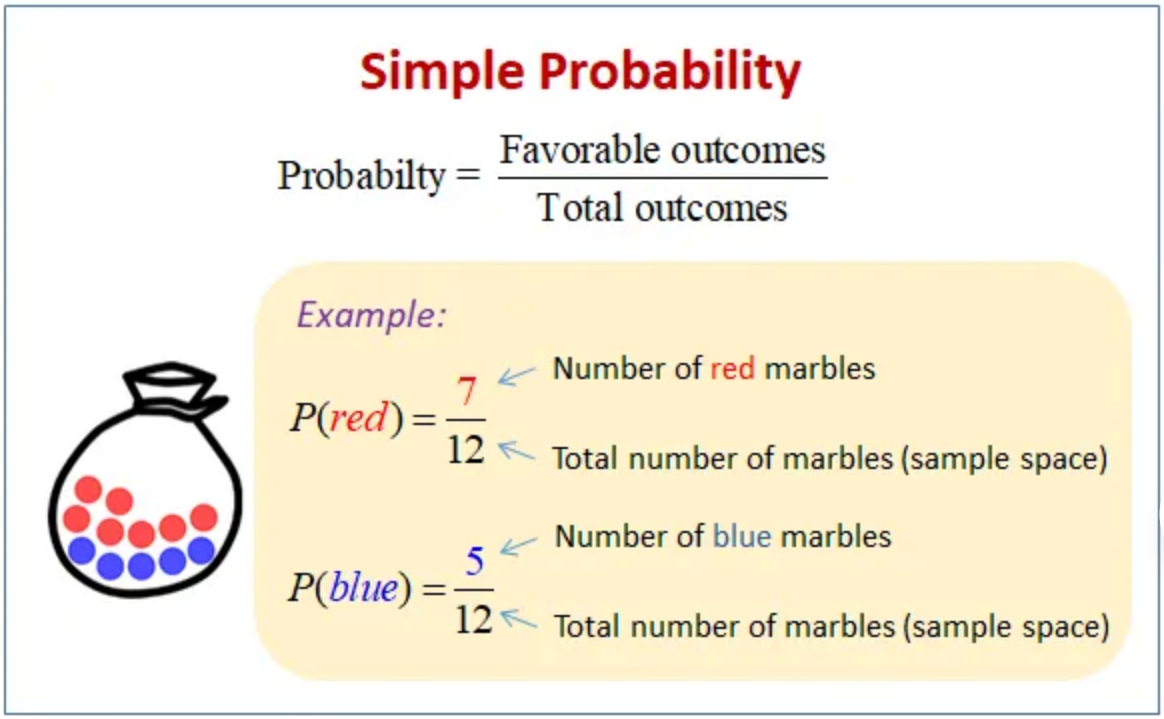
7. Confidence Interval In statistics
8. Type 1 And Type 2 error
9. One-Tailed And 2 Tailed Tests
10. Hypothesis Testing
11. p-value
12. Steps For Hypothesis Testing
13. T-test
14. Z- test
15. Annova test
16. Chi-square test

> ***In this article, we will cover all the basic statistics, and the remaining will be covered in Part 2 & Part 3.***

Let's start

Probability:

Probability is a measure of the likelihood of an event occurring. It is a number between 0 and 1, with 0 indicating that an event is impossible and 1 indicating that an event is certain to happen. For example, the probability of flipping a fair coin and it landing on heads is 0.5, or 50%. This is because there are two possible outcomes (heads or tails) and they are equally likely to occur.



Credit : onlinemathlearning

Addition Rule in Probability:

The addition rule of probability states that the probability of the union of two mutually exclusive events (events that cannot occur at the same time) is the sum of the probabilities of each individual event. The formula for the addition rule is:

P(A or B) = P(A) + P(B)

For example, if we flip a coin, the probability of getting heads (H) is 0.5 and the probability of getting tails (T) is also 0.5. These are mutually exclusive events, since the coin cannot land on both heads and tails at the same time. So, using the addition rule, we can say that the probability of getting heads or tails is:

P(H or T) = P(H) + P(T) = 0.5 + 0.5 = 1

which means that the event of getting heads or tails is certain to happen.

Multiplication Rule in Probability:

The multiplication rule in probability, also known as the conditional probability formula, states that the probability of two events occurring together (A and B) is equal to the probability of event A occurring multiplied by the probability of event B occurring, given that event A has already occurred.

Formula: P(A and B) = P(A) * P(B|A)

Example: Let's say we have a bag of 6 red marbles and 4 blue marbles.

- The probability of drawing a red marble is 6/10 or 3/5
- The probability of drawing a blue marble is 4/10 or 2/5

If we want to calculate the probability of drawing a red marble and then a blue marble, we would use the multiplication rule.

- P(red and blue) = P(red) * P(blue|red)
- P(red and blue) = 3/5 * 4/9 (since we have 6 red marbles and 4 blue marbles left in the bag)
- P(red and blue) = 8/45

Population And Sample:

Population and sample are two key terms used in statistics and research. A population is the entire group of individuals or objects that a researcher is interested in studying. For example, a population could be all adults living in a certain city or all students in a particular school district.

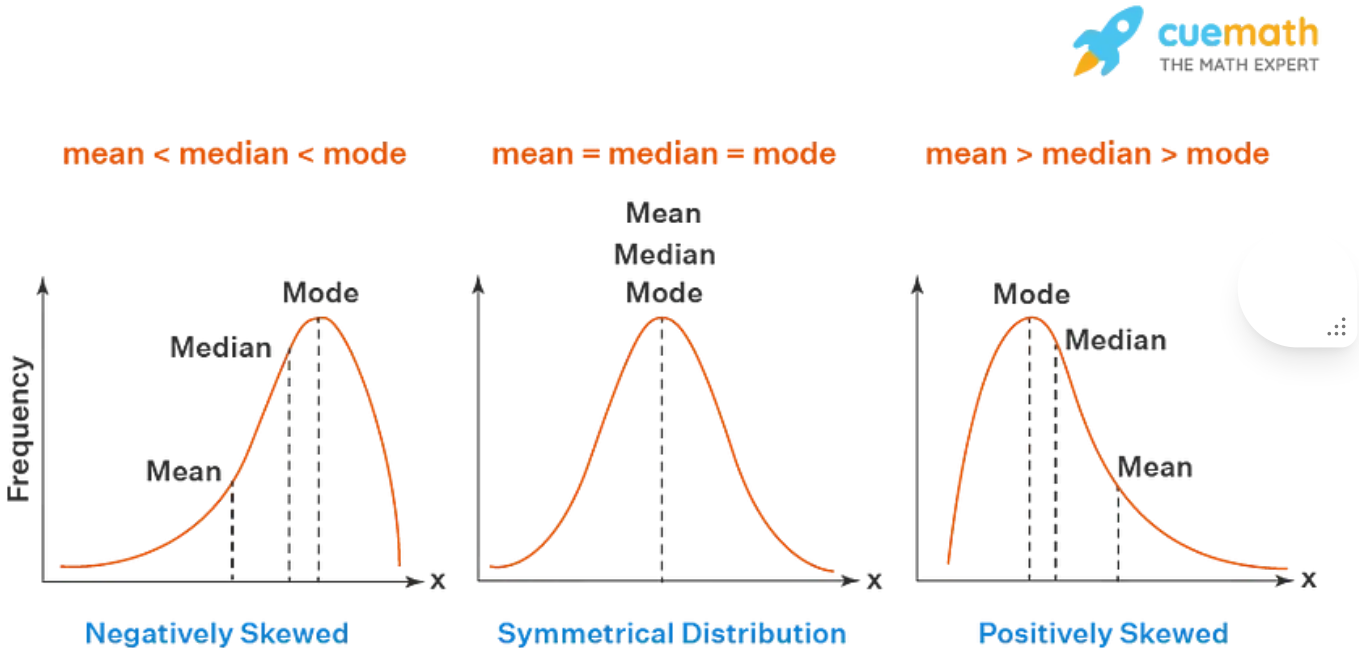
A sample, on the other hand, is a smaller group of individuals or objects selected from the population. The sample is used to represent the population and to make inferences about the population based on the sample's characteristics. For example, a researcher might select a sample of 100 adults living in a certain city to represent the population of all adults living in that city.

There are several different ways to select a sample from a population. One of the most common is simple random sampling, in which individuals or objects are chosen at random from the population. Another method is stratified sampling, in which the population is divided into subgroups (or strata) and a sample is selected from each subgroup.

It is important to note that the sample must be representative of the population in order for the research results to be valid.

Measure Of Central Tendency(Mean, Median, Mode):

Measure of central tendency refers to a single value that represents the center or typical value of a dataset. There are three main measures of central tendency: mean, median, and mode.



Credit: cuemath

The mean is the arithmetic average of a dataset. It is calculated by adding up all the values in a dataset and dividing by the number of values. The mean is a commonly used measure of central tendency, but it can be affected by outliers or extreme values.

The median is the middle value of a dataset when it is arranged in order. It is a useful measure of central tendency when a dataset has outliers or extreme values, as it is not affected by these values. To find the median, the data must be ordered, if there is an even number of observations the median is the average of the two middle values.

The mode is the value that appears most often in a dataset. A dataset can have one mode, more than one mode (multimodal) or no mode at all (unimodal). The mode is useful when the data is categorical, for example, when counting the different types of cars in a parking lot.

It is important to note that different datasets may have different measures of central tendency. For example, a dataset with a normal distribution will typically have a mean and median that are similar, while a dataset with a skewed distribution may have a mean and median that are quite different.

Measure Of Dispersion(Variance, Standard Deviation):

Measure of dispersion is a statistical term used to describe the degree to which a set of data is spread out. Two common measures of dispersion are variance and standard deviation.

Variance is a measure of how far each data point in a set is from the mean (average) of the set. It is calculated by taking the average of the squared differences from the mean. A high variance indicates that the data is spread out over a large range, while a low variance indicates that the data is clustered closely around the mean.

Standard deviation is a measure of the amount of variation or dispersion of a set of values. It is calculated as the square root of the variance. Standard deviation is used as a way to measure the volatility of a stock or other investments. A low standard deviation indicates that the data points tend to be close to the mean, while a high standard deviation indicates that the data points are spread out over a wider range.

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}}$

Sampling Method And Its Types:

Sampling is the process of selecting a representative group from a larger population. There are several different types of sampling methods, including:

1. Simple random sampling: This is a method where each member of the population has an equal chance of being selected.

2. Systematic sampling: This method involves selecting every nth member of the population, where n is determined by the size of the sample desired.

3. Stratified random sampling: This method involves dividing the population into groups (or strata) based on certain characteristics, and then randomly selecting members from each stratum.

4. Cluster sampling: This method involves dividing the population into groups (or clusters) and then randomly selecting a number of clusters to be included in the sample.

5. Convenience sampling: This method involves selecting participants based on their availability or willingness to participate.

6. Quota sampling: This method involves selecting a certain number of participants from specific subgroups within the population.

It is important to choose the appropriate sampling method based on the research question and the characteristics of the population being studied.

Population Mean And Sample Mean:

Population mean refers to the average value of a variable in a population, while sample mean refers to the average value of a variable in a sample drawn from that population. The population mean is denoted by the symbol "μ" (mu) and is calculated by adding all the values of the variable and dividing by the total number of observations in the population. The sample mean is denoted by the symbol "$\bar{x}$" (x-bar) and is calculated by adding all the values of the variable in the sample and dividing by the total number of observations in the sample. The sample mean is an estimate of the population mean.

What is Variables And Its Types?

A variable is a characteristic or attribute that can take on different values. In statistics, variables are used to represent the data that is being collected and analyzed. There are several types of variables, including:

1. Categorical variables: These are variables that can be divided into categories or groups. Examples include gender, race, and political party affiliation.
2. Numerical variables: These are variables that can take on numerical values and can be measured on a scale. Examples include age, income, and weight.

3. Ordinal variables: These are variables that can be put in a specific order or ranking, but the difference between the values is not known. Examples include education level and satisfaction.
4. Continuous variables: These are variables that can take on any value within a certain range, such as weight or height.
5. Discrete variables: These are variables that can only take on certain values, such as the number of children in a family.

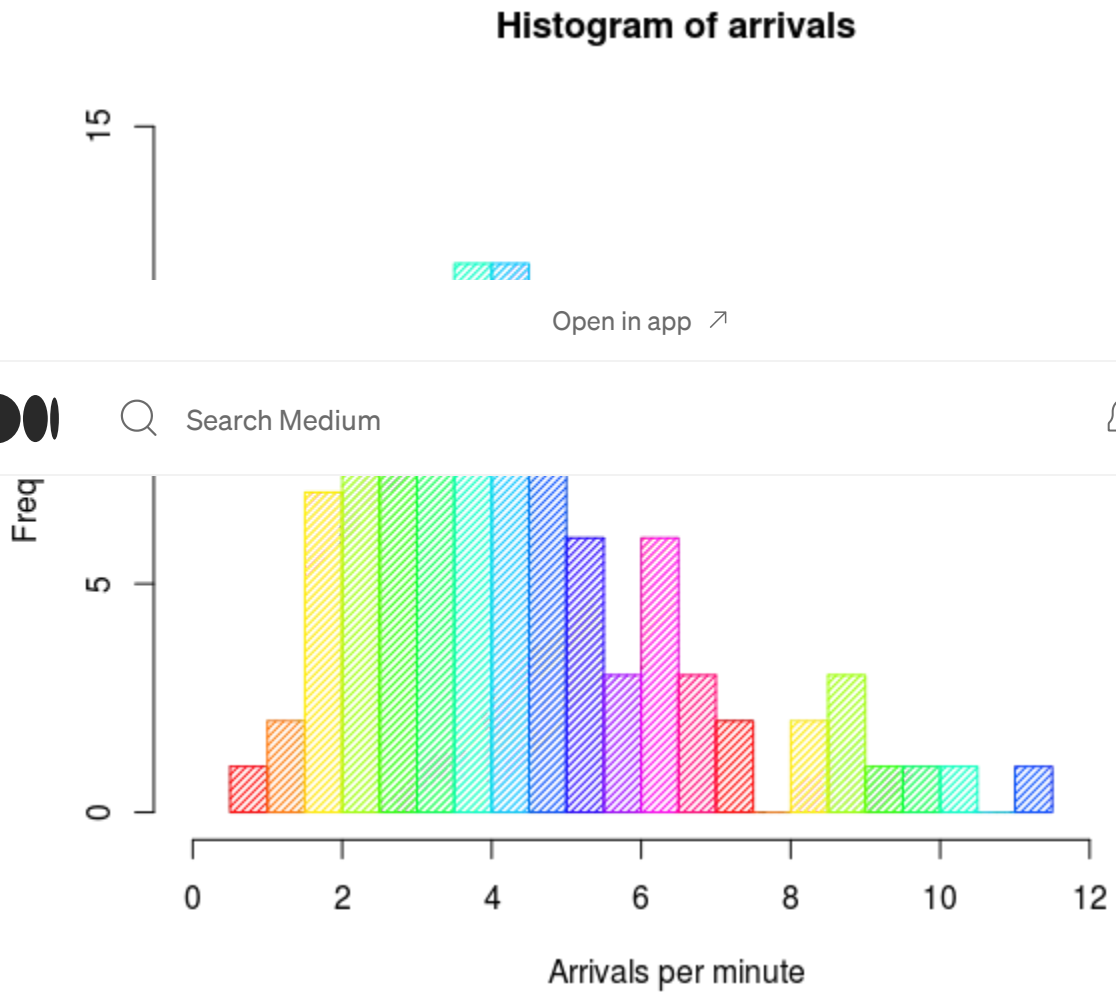
Frequency Distribution And Cumulative Frequency:

Frequency distribution is a table or graph that shows the number of occurrences (frequency) of each value or range of values of a variable. It is a way to organize and summarize large sets of data.

Cumulative frequency is the running total of frequencies. It is calculated by adding each frequency to the total of the previous frequencies. A cumulative frequency distribution is a table or graph that shows the cumulative frequency of each value or range of values.

Histograms:

A histogram is a graphical representation of a frequency distribution. It is a way to visually display the distribution of a continuous numerical variable. The x-axis represents the range of values for the variable, and the y-axis represents the frequency of those values. The range of values is divided into “bins” (or “intervals”), and each bin represents the frequency of observations that fall within that range of values. The height of each bar represents the frequency of observations in that bin.



To be Continued.....!

> **Link:-** Mastering the Fundamentals of Statistics for Data Science -Basic to Advance Level- Part 2

Follow me for Part 2 & Part 3 of Statistics for Data Science -Basic to Advance

Thanks for Reading!

If you enjoyed this, follow me to never miss another article on data science guides, tricks and tips, life lessons, and more!

Statistics | Data Science | Machine Learning | Artificial Intelligence | Mathematics