Population vs samples

Let's have a look at the things we're going to discuss in this article

- Basic Definitions
- Types of Variables
- Data Visualisation Basics
- Descriptive Statistics

Basic Definitions

- ### What is population?

  It refers to the total amount of things

- ### What is a sample?

  Small part of the population that is used for study.

- ### What is a sample size?

  Total amount of things in a sample.

- ### What is a variable?

  How a particular person can have different heights, weights and hair colour. These are variables. It refers to characteristics of what we are studying. Can vary among different individuals.

Types of Variables

There are mainly 2 types of variables:

- ### Quantitative data

  Data that is measured in numbers. It deals with numbers that make sense to perform arithmetic calculations with.

Ex:Height,Weight,Midterm Score

Quantitative data is further divided into two categories:

- **Discrete**
  Refer to variables that can only be measured in certain numbers.
  Ex: Number of pets you own.
  0 1 2 30
- **Continuous**
  Refer to variables that can take on any numerical value.
  Ex:Weight
  105 185 170.683

Quantitative data is also further divided into two categories:

- **Interval**
  An interval scale is one where there is order and the difference between two values is meaningful.
  Ex: Temperature (Fahrenheit), Temperature (Celsius), pH,
- **Ratio**
  A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.
  Ex: Weight, length, temperature in Kelvin

- The difference between interval and ratio scales comes from their ability to dip below zero. Interval scales hold no true zero and can represent values below zero. For example, you can measure temperature below 0 degrees Celsius, such as -10 degrees.
- Ratio variables, on the other hand, never fall below zero. Height and weight measure from 0 and above, but never fall below it.

Categorical data

Refers to the values that place “things" into different groups or categories.
Ex:Hair Colour,Type Of Cat,Letter Grade.

Categorical data is further divided into two categories:

- **Categorical and ordinal**
  Logical ordering to the values of a categorical variable.
  Ex: Letter Grade

- **Categorical and Nominal**
  No logical ordering to the values of a categorical variable.
  Ex: Hair Colour

Data Visualisation Basics

- ## What is data Visualization?

  Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen.

- ## What is its importance?

  We need data visualization because a visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows on a spreadsheet. It's the way the human brain works.

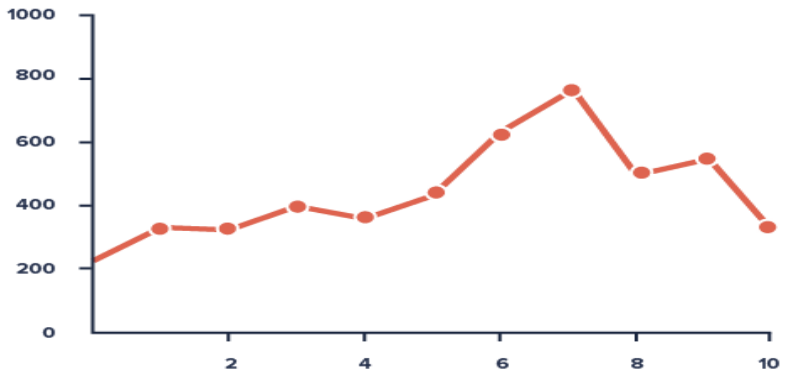
  Since the purpose of data analysis is to gain insights, data is much more valuable when it is visualized.

  Even if a data analyst can pull insights from data without visualization, it will be more difficult to communicate the meaning without visualization.

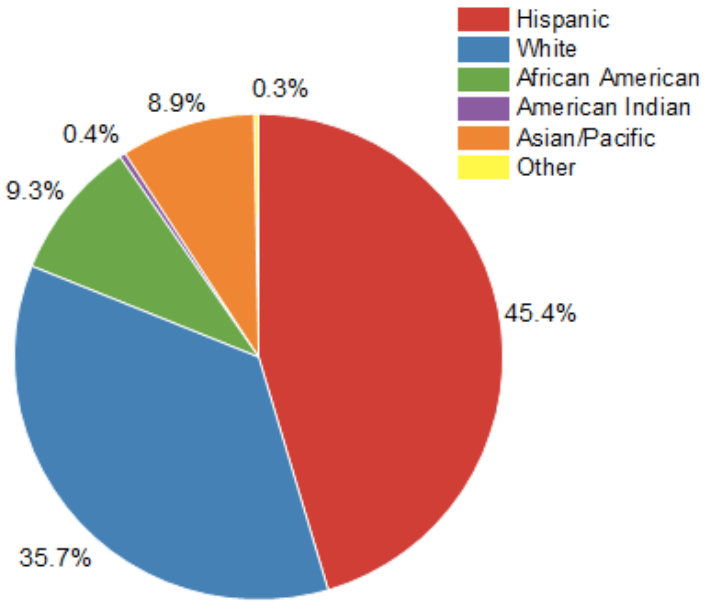
  Charts and graphs make communicating data findings easier even if you can identify the patterns without them.

- ## Different types of data visualisation

  - **Line Chart**
    A line chart is, as one can imagine, a line or multiple lines showing how single, or multiple variables develop over time.

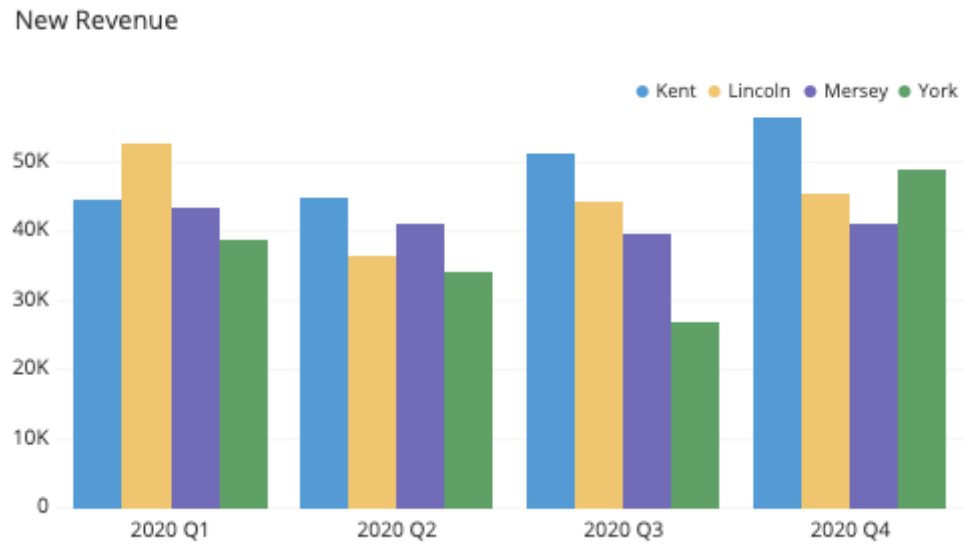


- **Pie Chart**
  A pie chart is a circular graph divided into slices. The larger a slice is, the bigger portion of the total quantity it represents.

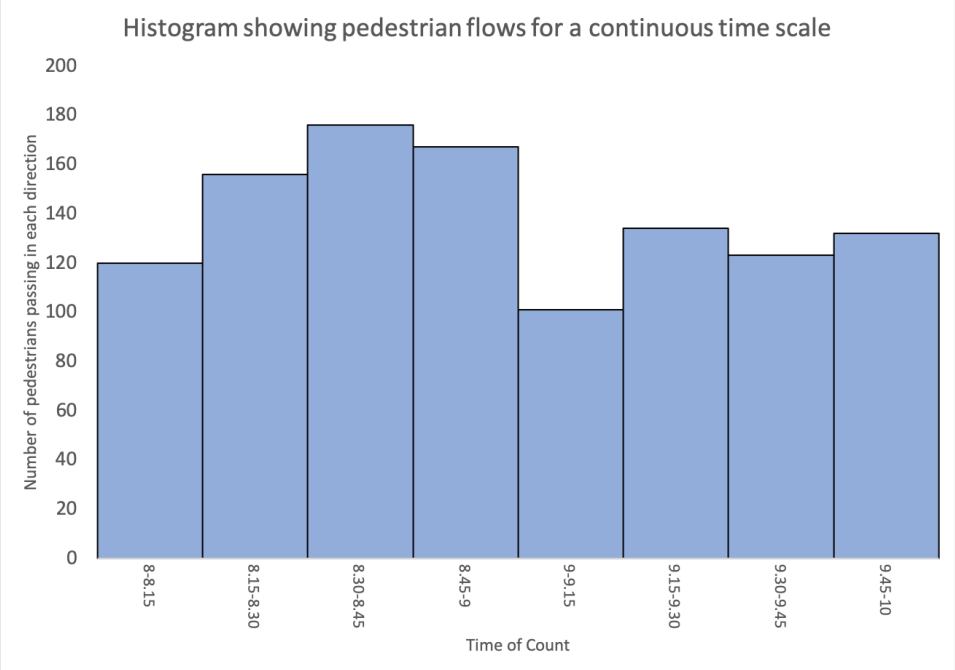


- **Bar Graph**
  A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that

they represent. The bars can be plotted vertically or horizontally. Can be of one variable or many variables.

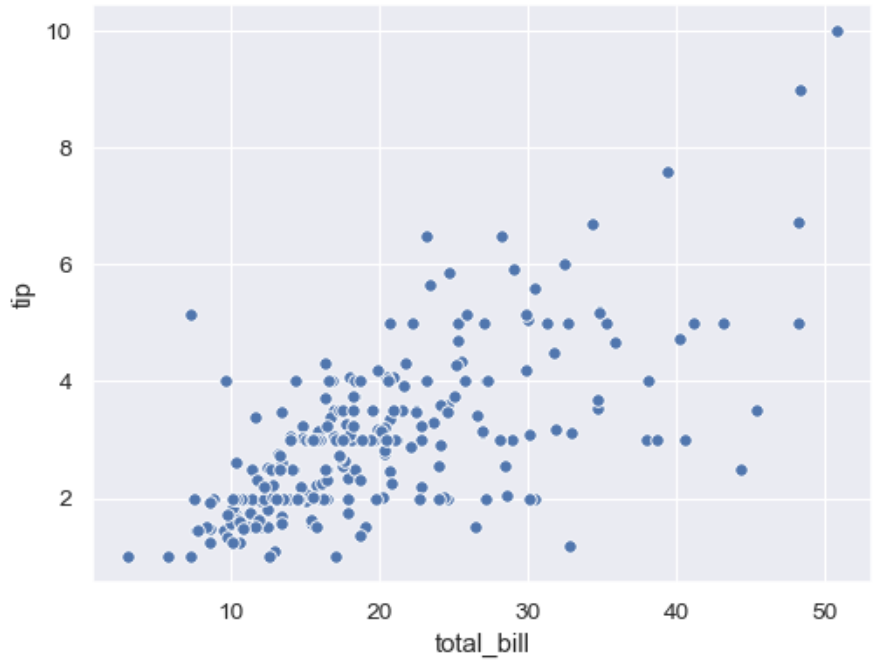


- **Histogram**
  A series of bins showing us the frequency of observations of a given variable.



- **Scatter Plots**

A scatter plot is a great indicator that allows us to see whether there is a pattern to be found between two variables. E.g. : Positive or negative relationship.



Descriptive Statistics

It deals with the processing of data without attempting to draw any inferences from it. The characteristics of the data are described in simple terms. Events that are dealt with include everyday happenings such as accidents, prices of goods, business, incomes, epidemics, sports data, population data.

When we give description of data, there can be 3 kinds:

1. Measures of Central Tendency – Mean, Median and Mode
2. Measures of Dispersion – Standard Deviation, Variance, Range, IQR (Inter Quartile Range)
3. Measure of Symmetricity – Skewness and Kurtosis

1. Measure of Central Tendency

A measure of central tendency is a summary statistic that represents the centre point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution.

- **Mean**
  Average value of the set of Numbers. Mean is a number around which a whole data is spread out. Denoted by population mean and for sample mean.
  Ex: Find the mean of 5,5,2,6,3,8,9?
  Mean is (5+5+2+6+3+8+9)/7 = 38/7 = 5.43
- **Median**
  Median is the value which divides the data in 2 equal parts i.e. number of terms on the right side of it is the same as number of terms on the left side of it when data is arranged in either ascending or descending order.

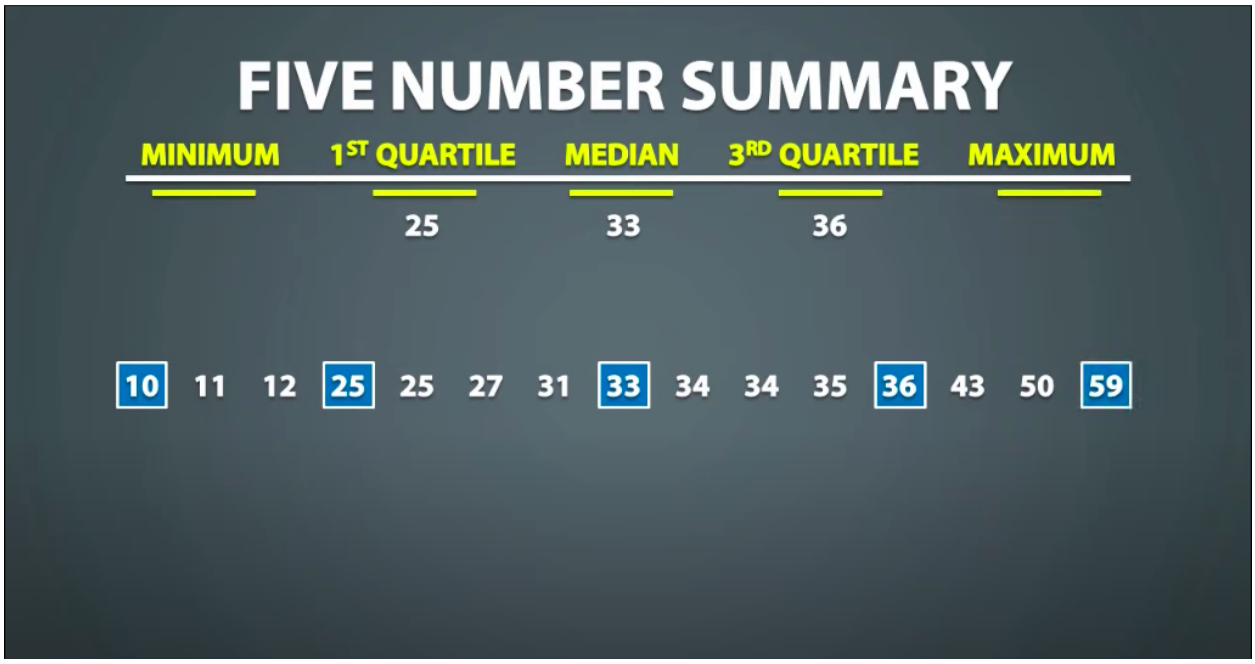
  (Note: If you sort data in descending order, it won't affect median but IQR will be negative.)
  Ex: Find the Median of 5,5,2,6,3,8,9?
  Putting it in ascending order = 2,3,5,5,6,8,9. Hence, Median = Mid Number = 5.

  (Note: Median of an even set of numbers can be found by taking the average of the 2 middle numbers.
- **Mode**
  Mode is the term appearing maximum time in a data set i.e. the term that has the highest frequency.
  Example: Find the Median of 5,5,2,6,3,8,9?
  Mode = Maximum number of repetition in dataset = 5. Hence, Mode = 5.
  (Note: If there is no repetition of data then mode is not present.

Let’s build some concepts before going ahead.

1. What is Minimum and Maximum value?
   It is the minimum and Maximum values of the dataset respectively.

2. What is the 1st and 3rd Quartile?
   Also called the lower and upper quartile respectively.When we divide the dataset into two groups while calculating median (sorted in ascending order), then the median of first half is 1st Quartile and median of second half is 3rd Quartile.



Learnvista Pvt Ltd.
2nd Floor, 147, 5th Main Rd, Rajiv Gandhi Nagar HSR Sector 7,Near Salarpuria Serenity, Bengaluru, Karnataka 560102
Mob:- +91 779568798, Email:- contacts@learnbay.co