**[“Statistics” by Saikiran Dasari for Data Science](https://medium.com/@saikirandasari41/statistics-by-saikiran-dasari-for-data-science-f995cebd2fd1)**

**Connecting the Dots of Stats (Layman Terms 1)**

A Quick Intro and Deliverables you get through this entire Stats Blog:

Hey guys, my name is Saikiran Dasari (Aspiring Data Scientist and AI enthusiast), and a Computer Engineer. I know many of you are struggling to revise the concepts of Statistics which you guys went through many of your courses and need to give it a shot for final preparation. I have carefully designed with aka Terms of each keyword of Stats you have never come across! and

**NOTE: This Blog will be Focusing on QUESTIONS and their Crisp Answers (where a Layman can understand easily!)**

I learned it all the hard way, and carefully analyzed all the courses and different trainer's explanations on topics, in my DS journey and designed this blog, the least I can do for you is give you a rock-solid foundation that you can use to upgrade your career, become a Statistically Best Data Scientist, and be part of the most lucrative career path of the future!

I have Segregated Stats in Parts wise!

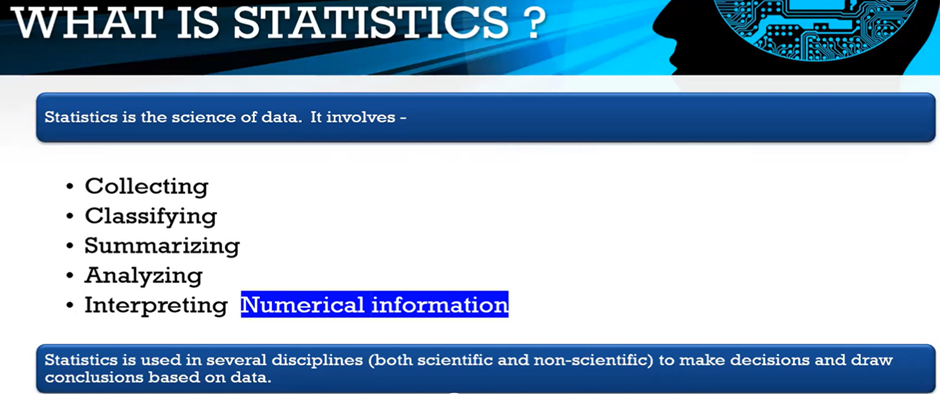
***Part 1: Understanding the Statistics, Difference between Data and Information, Its Data Types, Levels of Measurements, Population & Sample, Data Sources,* Good Questions that should meet the characteristics*, and Good to know terms***

— — — — — — — — — — — — — — — — — — — — — — — — — — —

Understanding the Statistics

**What is Statistics?**

The Definition of statistics given by R.A. Fisher is considered to be the best and most exact: which says: “Statistics is the branch of Applied Mathematics that specializes in data”

Overview of Stats!

**Collecting:** Based on launching Survey/ Rolling Google forms/Mailed Questionnaires/Info from local agents, etc

**Classifying:** Based on the Dataset we will Prepare a Subset of the Variable like Location1,2,3 or Experience levels into (0 to 1)(1–2) etc.., We will break the variables into different buckets to create subsets of the criteria

**Summarizing:** The data is summarised and explained in **Descriptive statistics.** A simple way to summarize data is to generate a table representing counts of various types of observations. One reason that we summarize data is that it provides us with a way to generalize that is, to make general statements that extend beyond specific observations.

**Analyzing: Analyse** data to find patterns, relationships, and trends.

**Interpreting:** It involves taking the result of data analysis, making inferences on the relations studied or it’s giving meaning to the collected ‘cleaned’ raw data.

**Limitation of Statistics?**

-> Stats deals with Quantitative data only, Even Qualitative info is converted into numerical data by methods like ranking, scoring, or scaling. Deals with masses, not an individual.

-> Stats results are correct in a general sense. They are always subject to a certain amount of error.

* > Stats is only meant to draw conclusions about masses or population but not to sort all problems.
* > We only infer from the Sample Data we do not conclude

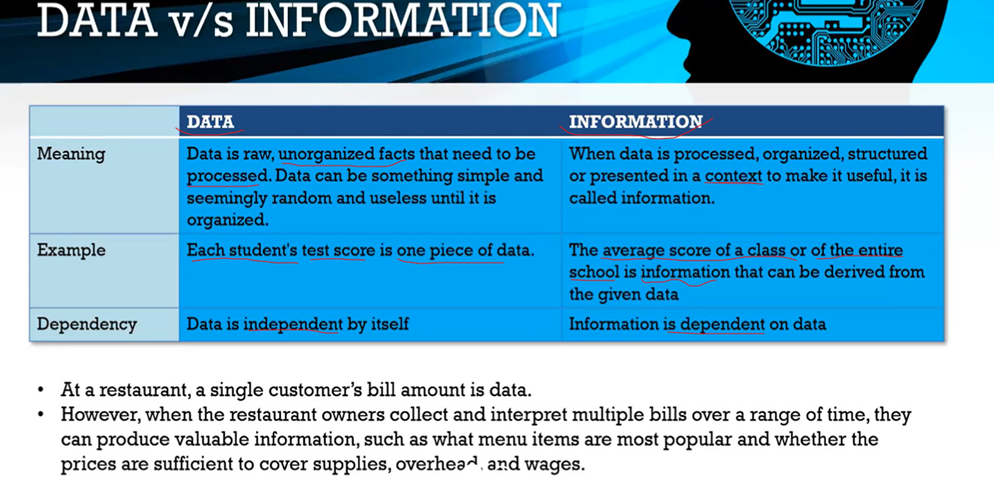
***(Terms like Quantitative, Qualitative, Sample, and Population used above will be answered in detail below)***

**NOTE:** **One DataPoint may not see Valuable. Every DP is put together and you will get a lot of information that you can build a lot of Analysis, which we typically called as Analytics or Data Science.**

— — — — — — — — — — — — — — — — — — — — — — — — — — —

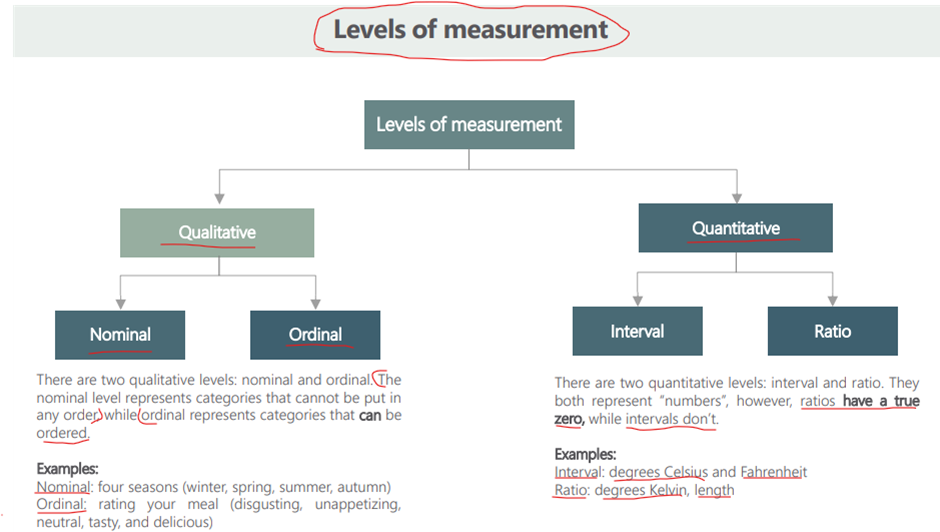
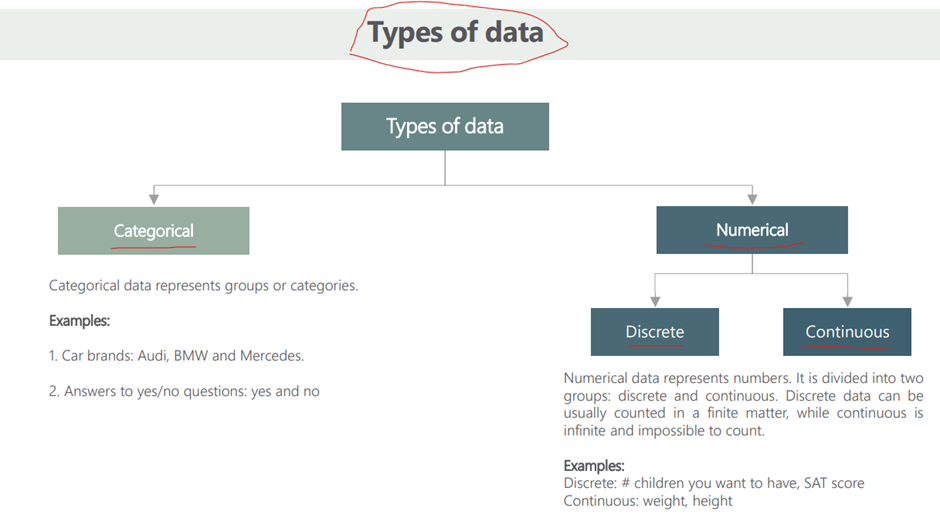
**Difference between Data and Information**

**Que) What is the difference between Data and Information?**

High-level differencing Data and Info!

 — — — — — — — — — — — — — — — — — — — — — — — — — — —

**Overview of “Types of Data” && “Levels of Measurement”:**



— — — — — — — — — — — — — — — — — — — — — — — — — — —

**1. Types of Data:**

**What are the Data Types in Statistics?**

NOTE: YOUR Data Can be OBSERVED(Qualitative) or MEASURED(Quantitative) that’s all!

**There are Predominently 2 Types of Data:**

**1. Categorical (Qualitative)/Observed**

**-> [Nominal, Ordinal]**

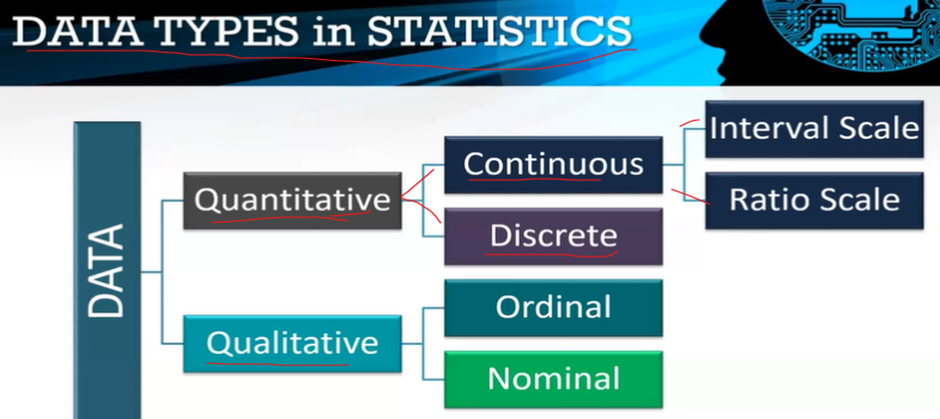
**Ordinal (Because It can be Ordered), Nominal (because it's just a Name)**

**2. Numerical (Quantitative)/Measured**

**-> [Continuous, Discrete]**

**Further, in Continuous we have [Interval scale, Ratio scale]**

**2. Levels of Measurements:**

Overview of Data Types Classification!

**A) Quantitative (Numerical)/Measured?**

**Information about something that is described by numbers.**

**Examples: Income, Age, Height**

With **Numerical** data, we can calculate statistics like the **average** income in a country, or the **range** of heights of players in a football team.

**Que) Difference between Continuous Data & Discrete Data?**

**Continuous data?**

Any data which can be expressed in Decimal Value then can be called a Continuous DataType. (questions like HOW Many?)

Eg: 10.5, 13.2 / Time and Money are continuous, Length, Weight, Height, etc

**Discrete data?**

Any data which can only take a WHOLE number can be called a Discrete DataType (Which does not take any Decimal Data Type).

You cannot get 2.5 on a Die, nor can you have a shoe size of 3.49

Eg: Number of Students, Patients, Cars, etc are examples of Discrete data.

***Further, in ‘CONTINUOUS’, we have { Interval Scale and Ratio Scale }***

**INTERVAL?**

**Data that can be ordered and the differences can be quantified/measured**

**No Absolute Zero**

**Eg**: Temperature measured in Fahrenheit/ Celcius, Negative Values are meaningful, Years in a calendar (Temperature may have 0 but they are not absolutely meaningful)

**RATIO?**

**Data that can be ordered and there is a consistent and meaningful distance between them.**

**And it also has an Absolute Zero**

Negative values basically don’t make sense

**Eg**: Money, Age, Time, Weight, Height, Length, liters, Sales of a new product

We have Meaningful Zeros for 0 Sales/Profit means No Sales/Profit! Etc) So, it’s a **RATIO**

— — — — — — — — — — — — — — — — — — — — — — — — — — —

**B) Qualitative (Categorical)/Observed**

**Information about something that can be sorted into different categories that can’t be described directly by numbers.**

**Examples: Brands, Nationality, Professions, Gender**

With **categorical** data, we can calculate **statistics** like **proportions**. For example, the proportion of Indian people in the world, or the percent of people who prefer one brand to another.

**NOMINAL? refers to names/Just labels give to something**

The NOMINAL level represents if there is no ORDER to the Categorical data/ categories that cannot be put in any order

Cannot be Ordered!

Eg: Seasons (Winter, Spring, Summer, Autumn), Colours, Countries

**ORDINAL? refers to order but cannot be measured!**

The ORDINAL level represents if you have a Preference for one thing over the other / categories that can be ordered

Can be Ordered!

Eg: Ratings, Rankings, CustomerLevelSatisfaction

— — — — — — — — — — — — — — — — — — — — — — — — — — —

**POPULATION and SAMPLE**

**What is the Difference between Population and Sample?**

A population is the collection of all items of interest to our study and is usually denoted with an uppercase N. The numbers we’ve obtained when using a population are called parameters.

A sample is a subset of the population and is denoted with a lowercase n, and the numbers we’ve obtained when working with a sample are called statistics.

That’s more or less what you are expected to say!

**Why do we draw Samples from the population?**

The population will include each & every bit of data point/observation which is relevant to one's study but it is not reachable/possible to get that observation and that’s why we draw a Sample (a part of the population) and we try to generalize it on the population.

**NOTE: The sample should have 2 properties: RANDOMness and REPRESENTATIVEness**

**Eg:**

In your office, say we have 4 departments: IT, Marketing, HR and Sales. There are 1000 people in each department and a total of 4000 people in all.

You want to evaluate to move to a new office and you decide you don’t want to ask all 4000 people but 100 is a nice example.

Now from 4 departments, we expect that out of 100 people 25 from each department!

**1st Case:**

We pick 100 people out of 4000 at Random and realize we have 30 IT, 30 Marketing, 30 HR, and 10 from Sales (Here we have a Sample which is Random but NOT Representative)

**2nd Case:**

Let’s say I’ve been working in that office for quite a while now. And you have many friends in each department. So, you pick 25 people from each department. (Here the Sample is representative but not Random! because you are considering a specific group of people and not the general public as a whole!)

**Solution:**

If we want to be Random and Representative, we will pick 25 people from IT at Random, then 25 people from Marketing, HR, and Sales at Random, and not based on the specific circle of people (like friends) only the general ‘public’.

— — — — — — — — — — — — — — — — — — — — — — — — — — —

**Data Sources**

**Name 2 Kinds of Statistical data and describe them in brief.**

**1)PRIMARY data**: PD which are collected from units/individuals directly and these data have never been used for any purpose earlier (eg: surveys, census data, etc).

**2)Secondary data:** SD, which had been collected by some individual or agency & statistically tested to draw certain conclusions. again the same data are used and analysed to Extract some other info (eg: Published thesis, Research papers, project reports, etc).

— — — — — — — — — — — — — — — — — — — — — — — — — — —

**Good Questions that should meet the characteristics:**

**Characteristics of a Good Questionnaire or a schedule?**

1. A number of questions should be such that it extracts all information required for reporting.
2. Each question should have almost all alternative answers.
3. The question should be clear and *without* any ambiguity (i.e ambiguity means containing more than 1 meaning which leads to confusion).
4. Should not be very lengthy and time-consuming.

— — — — — — — — — — — — — — — — — — — — — — — — — — —

**Additional Good to Know Terms:**

**Mention Main Divisions of Statistics?**

1. Mathematical or Theoretical Statistics
2. Statistical Methods or functions
3. Descriptive Stats
4. Inferential Stats
5. Applied Stats

**Types of Investigations?**

1. Investigation through Census Method
2. Investigation through Sample Methods

**Requisites of a Reliable Data?**

It should be Complete, Consistent, Accurate & should be Homogeneous (Homogeneous means the same kind) wrt a unit of information.

**Summary**

**What have we learned? We have learned —**

* Understanding the Statistics,
* Difference between Data and Information & Its Data Types,
* Levels of Measurements:(Categorical {Qualitative}, Numerical{Quantitative})
* Population & Sample,
* Data Sources,
* Good Questions that should meet the characteristics, & Good to-know terms

— — — — — — — — — — — — — — — — — — — — — — — —

**Next Part/Connecting Piece: Measures of Central Tendency, Measures of Dispersion, Measures of Position, Measure of Asymmetry, 4th Moment Business decision & much more.**

[**https://medium.com/@saikirandasari41/how-descriptive-stats-can-be-this-structured-6697b4f22202**](https://medium.com/@saikirandasari41/how-descriptive-stats-can-be-this-structured-6697b4f22202)

I have tried to write a detailed article and I hope that I am successful in doing so. I’ll try to keep on Updating this Blog with new insights & DOTs!