Road to Data Science in 50 Days - Day 2

Statistics for Data Science

Statistics is one of the key fundamental skills required for data science. Any expert in data science would surely recommend learning / upskilling yourself in statistics.

A lot of people skip the statistics part and directly jump to the modelling/ ML considering statistics to be tough. I was no different when i started my career as a Data Analyst. But being in a Market Research company, helped me understand the process of how important it is to understand the data (which is also a very crucial skill in Data Science) and learn to describe the data using the statistical techniques before going in to the modelling..

There are 3 types of Statistics:

1. **Descriptive** - What are the characteristics of the data?
2. **Inferential / Predictive** - What are the characteristics of the population? / Can we predict something given we have something?
3. **Prescriptive** - What can we do to make something happen?

Summarizing the above 3 in simple terms would be, Given the full stats for a player, analyzing that Kohli has XX century is *Decriptive Statistics*, How much centuries will kohli make in the next world cup given his stats and form is *Predictive Statistics*, And how can we make kohli make centuries in the world cup given the data we have in hand is *Prescriptive Statistcis*.

Having said that, let’s understand the basics of data before diving deeper into descriptive and inferential statistics.

Types of Data

The different types of data you will usually come across can be classified into 2 groups:

1. Quantitative (Numerical)

- **Continuous (Scale)**: Continuous Data represents measurements and therefore their values can't be counted but they can be measured. An example would be the height of a person, which you can describe by using intervals on the real number line. It can be further classified into:
  - Interval
  - Ratio
- **Discrete (Finite)**: These are data that can take only certain specific values rather than a range of values. For example, data on the blood group of a certain population or on their genders is termed as discrete data.

2. Qualitative (Categorical):

- **Nominal**: Nominal data are used to label variables without any quantitative value. Common examples include male/female, hair color, nationalities, names of people, and so on.

- **Ordinal**: The key with ordinal data is to remember that ordinal sounds like order - and it's the order of the variables which matters. Not so much the differences between those values. Ordinal scales are often used for measures of satisfaction, happiness, and so on. Have you ever taken one of those surveys, like this? "How likely are you to recommend our services to your friends?" * Very likely * Likely * Neutral * Unlikely * Very unlikely

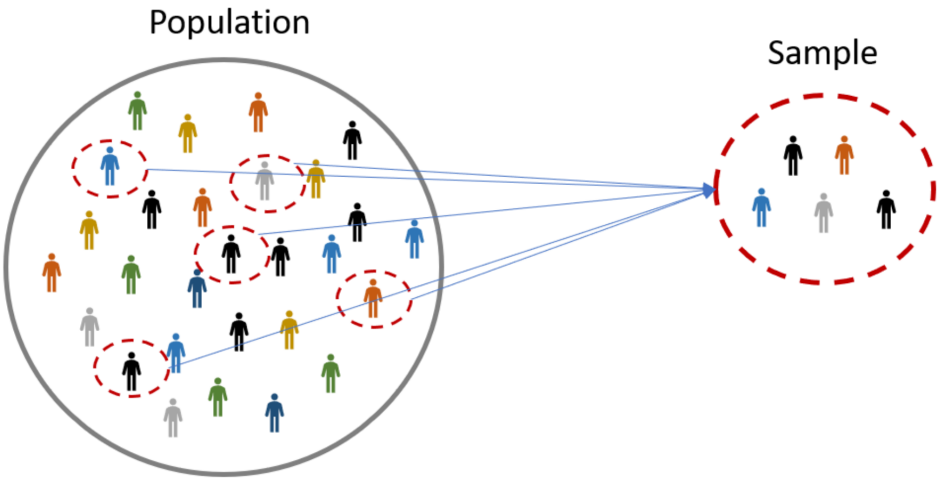
Why Data Types are important?

Datatypes are an important concept because statistical methods can only be used with certain data types. You have to analyze continuous data differently than categorical data otherwise it would result in a wrong analysis. Therefore knowing the types of data you are dealing with, enables you to choose the correct method of analysis. P.S. These are different from the datatypes used in programming language, which we will see ahead in some

Research

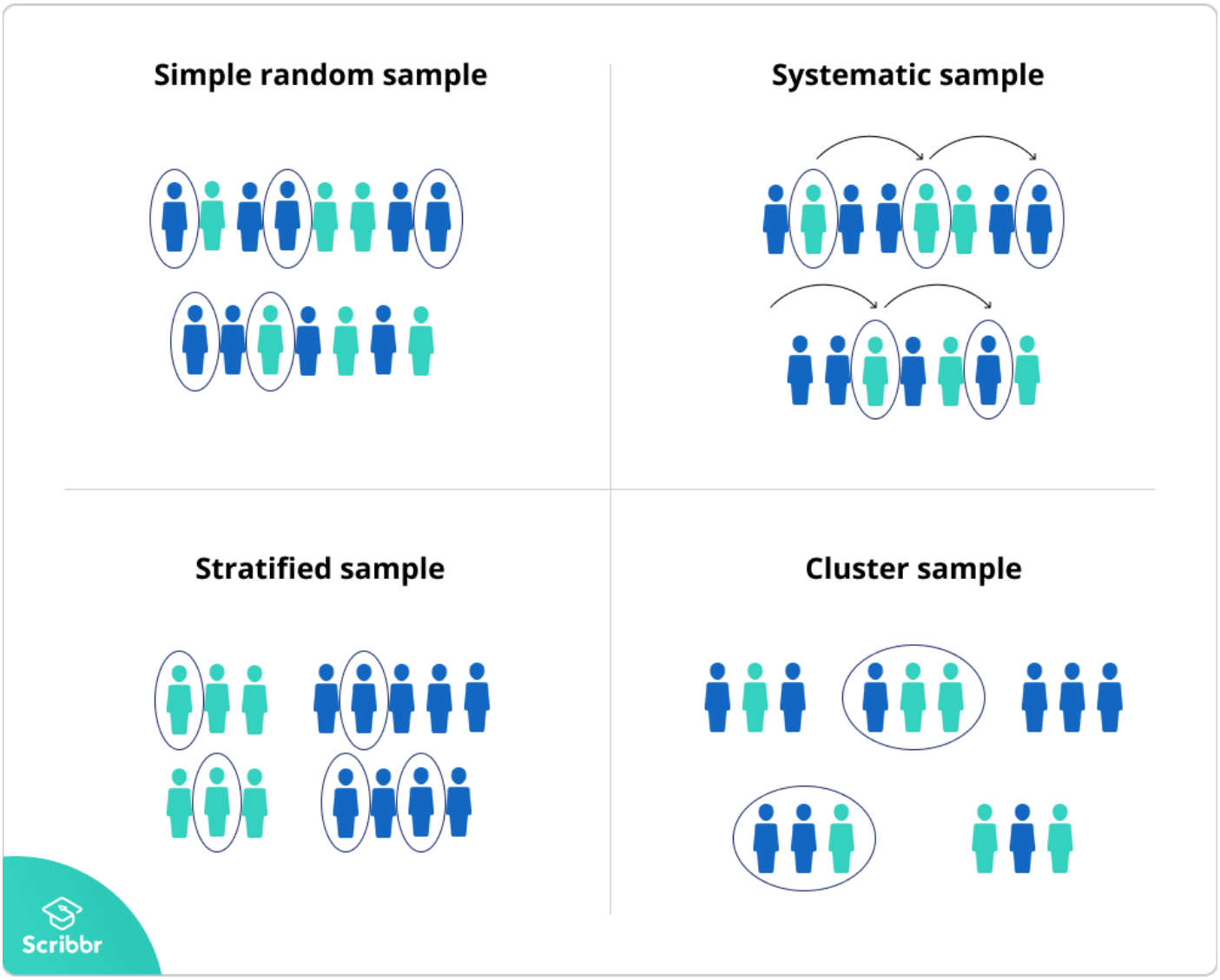
In any research objective, there are few crucial terminologies one should be aware of being a data scientist. Mentioned below are a few of them:

- **Population**: Population is the set of sources from which data has to be collected.
- **Sample**: A Sample is a subset of the Population.
- **Sampling**: The process of selection or the drawing of the accurate representation of a unit, group or sample from a population of interest is called as sampling.
- **Sampling Error**: The variation between the means of sample groups as well as population mean is called sampling error.



It is usually said that we almost never know the characteristics of Population or collecting the population data can be very costly, hence we sample the data which would represent the population and we perform the research objective on the sample data. The process of conducting a survey to collect data from the entire population is called a census.

Sampling Techniques:



(https://www.scribbr.com/methodology/sampling-methods/)

1. Non- Probability Sampling

- **Convenience Sampling** : Sampling When the researcher selects sample for the study at his own convenience is called as convenience sampling
- **Quota Sampling** : In this sampling the investigator initially sets some relevant categories of people and decides the number of units should be selected for the study as a sample. Such as male= 10, female=10
- **Purposive Sampling** : It is also known as judgment sampling. It is valuable in special circumstances. In purposive sampling the researcher never knows whether the cases, selected represent the population
- **Snowball Sampling** : It is a sociometric sampling method and also known as network, chain referral or reputation sampling method. In this method the researcher starts collection of data from the person who known to the researcher

2. Probability Sampling

- **Simple random sampling**: It is the simplest form of random sampling. In this sampling technique each elements of population might have given equal chance to be selected for the study

- **Stratified random sampling technique**: When the researcher needs stratification of population based on single characteristics or attributes such as male and female, urban and rural, married and unmarried and so forth he/ she warranted the stratified random sampling technique
- **Cluster Sampling**: Cluster sampling is a variation of simple random sampling. It is used when the population of the study is infinite and the population units are scattered across the wide geographical area.
- **Systematic Sampling**: Systematic sampling can be defined as selecting or drawing of every nth item or person from a pre determined list. Such as selection of every 10th person from a telephone directory or every 6th person from a college admission register.

To learn more about sampling in general, refer: https://tophat.com/marketplace/social-science/education/course-notes/oer-research-population-and-sample-dr-rafeedalie/1196/ (https://tophat.com/marketplace/social-

Descriptive and Inferential Statistics

Descriptive statistics, as the name suggests, is used to display and describe data by using tables, graphs and summary measures.

Inferential Statistics, on the other hand, pertains to studying a sample and use the results to make decisions or predictions about a population.

Descriptive Statistics and exploratory data analysis should be the first steps while building predictive or inference models. Descriptive statistics help understand large amounts of data by providing methods to summarise the data and retrieve information about the underlying structure of the data.

When we try to represent data in the form of graphs, like histograms, line plots, etc. the data is represented based on some kind of central tendency. Central tendency measures like, mean, median, or measures of the spread, etc are used for statistical analysis. To better understand Statistics lets discuss the different measures in Statistics with the help of an example.

Cars	mpg	cyl	disp	hp	drat
A	21	6	160	110	3.9
B	21	6	160	110	3.9
C	22.8	4	108	93	3.85
D	21.3	6	108	96	3
E	23	4	150	90	4
F	23	6	108	110	3.9
G	23	4	160	110	3.9
H	23	6	160	110	3.9

Here is a sample data set of cars containing the variables:

Cars Mileage per Gallon (mpg) Cylinder Type (cyl) Displacement (disp) Horse Power (hp) Real Axle Ratio (drat). Before we move any further, let’s define the main Measures of the Center or Measures of Central tendency.

Measures Of The Center

Mean: Measure of average of all the values in a sample is called Mean.

Median: Measure of the central value of the sample set is called Median.

Mode: The value most recurrent in the sample set is known as Mode.

Using descriptive Analysis, you can analyse each of the variables in the sample data set for mean, standard deviation, minimum and maximum.

- If we want to find out the mean or average horsepower of the cars among the population of cars, we will check and calculate the average of all values. In this case, we’ll take the sum of the Horse Power of each car, divided by the total number of cars:

Mean = (110+110+93+96+90+110+110+110)/8 = 103.625

- If we want to find out the center value of mpg among the population of cars, we will arrange the mpg values in ascending or descending order and choose the middle value. In this case, we have 8 values which is an even entry. Hence we must take the average of the two middle values.

The mpg for 8 cars: 21,21,21.3,22.8,23,23,23,23

Median = (22.8+23)/2 = 22.9

- If we want to find out the most common type of cylinder among the population of cars, we will check the value which is repeated most number of times. Here we can see that the cylinders come in two values, 4 and 6. Take a look at the data set, you can see that the most recurring value is 6. Hence 6 is our Mode.

Some of the other measures are Geometric mean and harmonic mean

- Geometric mean: The Geometric Mean is a special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc.
- Harmonic mean: The harmonic mean is the reciprocal of the average of the reciprocals

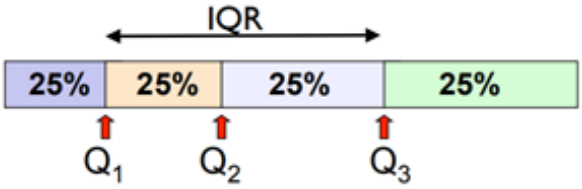
$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \dots}$$

Measures Of The Spread

Just like the measure of center, we also have measures of the spread, which comprises of the following measures:

Range: It is the given measure of how spread apart the values in a data set are.

Inter Quartile Range (IQR): It is the measure of variability, based on dividing a data set into quartiles. The formula for IQR is Q3 - Q1. This is the range where your 50% of the data lies.



Variance: It describes how much a random variable differs from its expected value. It entails computing squares of deviations.

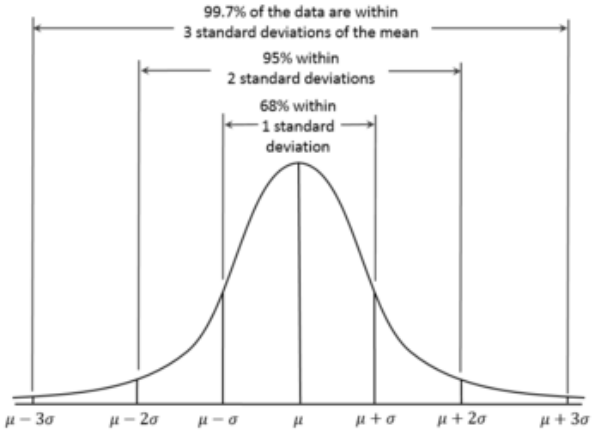
Deviation is the difference between each element from the mean.

Population Variance is the average of squared deviations

Sample Variance is the average of squared differences from the mean

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

Standard Deviation: It is the measure of the dispersion of a set of data from its mean.



Referece: https://www.edureka.co/blog/math-and-statistics-for-data-science/ (https://www.edureka.co/blog/math-and-statistics-for-data-science/)

Skewness

Skewness is a key statistics concept you must know in the data science and analytics fields

The concept of skewness is baked into our way of thinking. When we look at a visualization, our minds intuitively discern the pattern in that chart.

As you might already know, India has more than 50% of its population below the age of 25 and more than 65% below the age of 35. If you'll plot the distribution of the age of the population of India, you will find that there is a hump on the left side of distribution and the right side is comparatively planar. In other words, we can say that there's a skew towards the end, right?

So even if you haven't read up on skewness as a data science or analytics professional, you have definitely interacted with the concept on an informal note. And it's actually a pretty easy topic in statistics – and yet a lot of folks skim through it in their haste of learning other seemingly complex data science concepts. To me, that's a mistake.

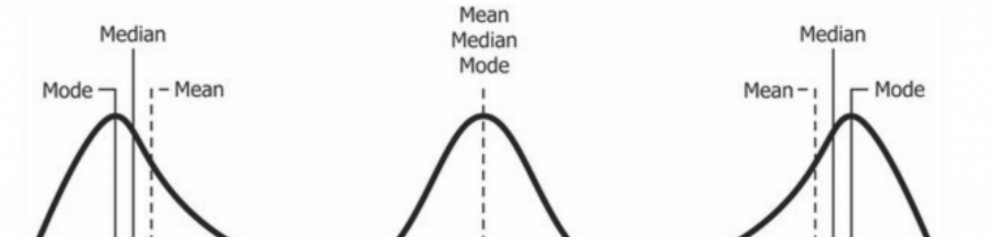
What is Skewness?

Skewness is the measure of the asymmetry of an ideally symmetric probability distribution and is given by the third standardized moment. If that sounds way too complex, don't worry! Let me break it down for you.

In simple words, skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution. Now, you might be thinking – why am I talking about normal distribution here?

Well, the normal distribution is the probability distribution without any skewness. You can look at the image below which shows symmetrical distribution that's basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. Apart from this, there are two types of skewness:

- Positive Skewness
- Negative Skewness



And that's a wrap up for Day 2, we will learn little bit more about Descriptive statistics tomorrow and then we will move towards Inferential Statistics and Probability.

References and Important links to learn more about Research and Descriptive Statistics:

https://towardsdatascience.com/descriptive-statistics-f2beeaf7a8df (https://towardsdatascience.com/descriptive-statistics-f2beeaf7a8df) - By Satyapriya Chaudhury

https://www.scribbr.com/methodology/sampling-methods/ (https://www.scribbr.com/methodology/sampling-methods/) - Scribbr (Types of Sampling)

https://www.edureka.co/blog/math-and-statistics-for-data-science/ (https://www.edureka.co/blog/math-and-statistics-for-data-science/) - Math and Statistics for Data Science by Edureka

https://www.youtube.com/watch?v=QoQbR4lVLrs (https://www.youtube.com/watch?v=QoQbR4lVLrs) - by Teresa Johnson - Youtube

Useful Interview Questions:

https://www.educba.com/statistics-interview-questions/ (https://www.educba.com/statistics-interview-questions/)

https://www.analyticsvidhya.com/blog/2017/05/41-questions-on-statisitics-data-scientists-analysts/ (https://www.analyticsvidhya.com/blog/2017/05/41-questions-on-statisitics-data-scientists-analysts/)