**Introduction to Statistics**

Statistics is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them.

Data can be defined as groups of information that represent the qualitative or quantitative attributes of a variable or set of variables. In layman's terms, data in statistics can be any set of information that describes a given entity. An example of data can be the ages of the students in a given class. When you collect those ages, that becomes your data.

As we have seen in the definition of statistics, data collection is a fundamental aspect and therefore there are different methods of collecting data which when used on one set will result in different kinds of data. Let's move on to look at these individual methods of collection to better understand the types of data that will result.

**1. Census Data Collection**

Census data collection is a method of collecting data whereby all the data from every member of the population is collected.

**2. Sample Data Collection**

Sample data collection, which is commonly just referred to as sampling, is a method which collects data from only a chosen portion of the population.

Sampling assumes that the portion that is chosen to be sampled is a good estimate of the entire population. Thus, one can save resources and time by only collecting data from a small part of the population. But this raises the question of whether sampling is accurate or not. The answer is that for the most part, sampling is approximately accurate. This is only true if you choose your sample carefully to be able to closely approximate what the true population consists of.

Sampling is used commonly in everyday life, for example, all the different research polls that are conducted before elections. Pollsters don't ask all the people in a given state who they'll vote for, but they choose a small sample and assume that these people represent how the entire population of the state is likely to vote. History has shown that these polls are almost always close to accuracy, and as such sampling is a very powerful tool in statistics.

**3. Experimental Data Collection**

Experimental data collection involves one performing an experiment and then collecting the data to be further analysed. Experiments involve tests and the results of these tests are your data.
An example of experimental data collection is rolling a die one hundred times while recording the outcomes. Your data would be the results you get in each roll. The experiment could involve rolling the die in different ways and recording the results for each of those different ways.

**4. Observational Data Collection**
Observational data collection method involves not carrying out an experiment but observing without influencing the population at all. Observational data collection is popular in studying trends and behaviors of society where, for example, the lives of a bunch of people are observed and data is collected for the different aspects of their lives. Analysis of data collected in such ways can broadly categorized into 2 categories called descriptive and inferential statistics.

**Descriptive vs Inferential Statistics**

Descriptive statistics deals with the processing of data without attempting to draw any inferences from it. The data are presented in the form of tables and graphs. 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.
Inferential statistics is a scientific discipline that uses mathematical tools to make forecasts and projections by analyzing the given data. This is of use to people employed in such fields as engineering, economics, biology, the social sciences, business, agriculture and communications.

**Descriptive Statistics**
Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.
Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data. For example, if we had the results of 100 records of students' marks, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this. Typically, there are two general types of statistic that are used to describe data:

**1. Measures of Central Tendency:**
These are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest. We can describe this central position using the mode, median, and mean.

**2. Measures of Spread:**
These are ways of summarizing a group of data by describing how spread out the scores are. For example, the mean score of our 100 students may be 65 out of 100. However, not all students will have scored 65 marks. Rather, their scores will be spread out. Some will be lower and others higher. Measures of spread help us to summarize how spread out these scores are. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

**Measures of Central Tendencies -**
A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.
The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to
use than others.

**Mean**
The mean is the average of all numbers and is sometimes called the arithmetic. To calculate mean, add together all of the numbers in a set and then divide the sum by the total count of numbers.

   

Where (x1, x2, x3,…..xn) are all the elements in the sample and n is the size of the sample.

For example, in a data center rack, five servers consume 100 watts, 98 watts, 105 watts, 90 watts and 102 watts of power, respectively. The mean power use of that rack is calculated as (100 + 98 + 105 + 90 + 102 W)/5 servers = a calculated mean of 99 W per server.

**Median**
In a data center, means and medians are often tracked over time to spot trends, which inform capacity planning or power cost predictions. The statistical median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median.
For the five servers in the rack, arrange the power consumption figures from lowest to highest: 90 W, 98 W, 100 W, 102 W and 105 W. The median power consumption of the rack is 100 W. If there is an even set of numbers, average the two middle numbers. For example, if the rack had a sixth server that used 110 W, the new number set would be 90 W, 98 W, 100 W, 102 W, 105 W and 110 W. Find the median by averaging the two middle numbers: (100 + 102)/2 = 101 W.

**Mode**
The mode is the number that occurs most often within a set of numbers. For the server power consumption examples above, there is no mode because each element is different. But suppose the administrator measured the power consumption of an entire network operations center (NOC) and the set of numbers is 90 W, 104 W, 98 W, 98 W, 105 W, 92 W, 102 W, 100 W, 110 W, 98 W, 210 W and 115 W. The mode is 98 W since that power consumption measurement occurs most often amongst the 12 servers. Mode helps identify the most common or frequent occurrence of a characteristic. It is possible to have two modes (bimodal), three modes (trimodal) or more modes within larger sets of numbers.

**Measures of spread**
A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population. It is usually used in conjunction with a measure of central tendency, such as the mean or median, to provide an overall description of a set of data.
The range, the variance, and the standard deviation are the most common measures of spread or variation.
The range is the length of the smallest interval which contains all the data. It is calculated by subtracting the smallest observation (sample minimum) from the greatest (sample maximum). Alternatively, the range can be articulated as simply listing the lowest to highest value (i.e. range of 2 – 10).

The variance is the average difference of each value in the sample from the mean.



The standard deviation is simply the square root of the variance.

# Inferential Statistics

# Drawing Inferences from Data

The objective of making inference from data is to make **intelligent assertion** like -

1. People who don’t smoke live longer than people who smoke
2. 80% of all vehicles in USA are 4 wheelers

**Why making inference from data is important ?**

In our professional life, we make decision driven by data. It is always a better idea to have data to back our decision. In case if we don’t have data to back our decision, it can be easy that we can make wrong conclusion. It is a tangible way which you can use to defend yourself from the consequence of a decision which was correct based on information available at the point when decision was made, and which then went wrong later.

**How to make assert**

Let’s take the above example statement, People who don’t smoke live longer than people who smoke.

We all know how long a person lives is subject to chance. No one will know how he or she will live. Any variable which is subjected to chance is called **random variable.**  A random variable could take any one of different value. And the behavior of random variable is governed by their **probability distributions**.

Usually we will study a small group of people who smoke daily and then compare them with another small group of people who don’t smoke. This is called **sampling**.

For each of these 2 small groups, we will try and form a **Hypothesis** what might be true for all people and see if this Hypothesis is supported or not supported by statistics.

Our findings about the sample would allow us to test our Hypothesis. And this testing would rely on probability distributions of the underlying random variable.

This process of testing hypothesis is key, to make defensible assertion which are backed by data.

# Random Variables

**Random variables are variable**

- Whose value cannot be determined before an event happens.
- Whose value is subject to variation due to chance, but we know the value is restricted to a finite set of values.

**Example of Random variable**

- A person’s blood type
- Number of leaves on a tree
- Number of times a user visits LinkedIn in a day
- Length of a tweet.

**Types of random variable**

- Discrete, which can take only integer values (like 0,1,2,…)
- Continuous, which can take any value from a range of values
- Categorical, which can take one of a limited, fixed set of values. Eg., red, blue

Now let’s take a problem like fraud detection. To detect any fraudulent card transaction, we need to identify those variables which are related to a card transaction and can influence the problem. Below are few such variables and their types.

1. Amount spent (0 to ∞) – Continuous
2. IP address (set of all IP address in the world) - Categorical
3. Number of failed attempts on using the card (0, 1, 2.. ) - Discrete
4. Time since last transactions (0 … ∞) - Continuous
5. Location of transaction (Austin, Dallas, New York) – Categorical

As we see, each of these variables can have some influence on the transaction. We will not know the value of these variables before the transaction occur but will know the range for their values. And for each transaction, each variable’s value will be different, but it has to be from with in this set of range.

**Statistical Experiment**

Below statistic report showing LinkedIn user’s top 10 geographical distribution. Let say we need to pick a user at random from the entire group of LinkedIn user, and tell what country they are likely to be from. This is a statistical experiment – whose set of outcomes can be specified beforehand, but the actual outcome of the experiment is subject to chance.



Please note here the country of the user will be random variable which usually will be represented as X. Probability distribution is a table or function that links each outcome of a statistical experiment with its probability of occurrence.

                                               P (the person picked is from USA) = P(X) = 0.3 (from above table)

**Tossing a Die**

Another statistical experiment is tossing a die. When we toss a die, we will not know which value will come up, but we will know it has to be one of the values from 1 to 6. The outcome of a statistical experiment is represented by random variable.
In this case let say the outcome of the toss is X. Now X will be a discrete variable as it can take value from 1 to 6 all integers. Below table links each possible value of X with its probability.



As we see, all of the outcomes have equal probability and hence it is called uniform probability distribution. A Uniform distribution is a distribution that has a constant probability or function.

So as random variable can be Discrete or Continuous, so can their probability distribution also. Statisticians and Mathematicians have studied a lot of different random variables in nature and realized that there are some recurring themes. They have defined some standard distribution and most random variables that you would ever encounter would fall into one of these below distributions.



**Normal Distribution**

Let’s talk about the normal distribution which is very commonly seen in many instances of machine learning and statistical problems. Normal distribution is a distribution pattern which happens to occur lot in many natural phenomena. Below are some examples of normally distributed random variables which when plotted on a curve will results in an inverted curved which is a normal distribution.

- Height of a person
- Blood pressure
- Performance of students in a class

In normal distribution, most of the measurement (say height of a person) will be concentrated in the central peak. And there will be very few measurements that are very far off from the central point. Now these measurements are drawn from probability distribution which is basically a normal distribution.

- X axis is the value of the random variable.
- Y axis is the probability that it can take.
- The peak is the mean or the average value.
- Most of the measurement will be concentrated around the mean or average value.
- The spread of the distributions is described by the standard deviation.
- Mean and standard deviation are sufficient to completely describe the normal distribution.

# Normal distribution

**Normal Distribution** is also called as gaussian distribution or bell curve. It plays a special role in statistic as many phenomena in natural life just follows the normal curve. Below are few data sets whose values are distributed normally.

- Weights of a group of football players
- The sizes of houses in a neighborhood
- The IQs of a group of students.

If we took all of the values of any of these data set and plot a histogram, then the resulting histogram will look very much similar like below. Then if we draw a smooth curve through the histogram, we will get a normal curve as shown. This normal curve is mathematically significant.



**Mean –** The peak of this curve occurs at the MEAN and it is represented as µ. When a variable is normal, its value will most likely be close to the mean.

**Standard deviation -** The spread of this curve is defined by the standard deviation represented as σ

The normal distribution is entirely defined by 2 parameters µ & σ with the following formula. In other word, given the mean and SD, we can tell the probability of any value.

        Or  simply we can define as f(x) = F(x, μ, σ)

If we know the mean, SD you can tell the probability of any value that can occur in the normal distribution.

- Regardless of the actual values of mean and SD some characteristic remains same.
- Probability that a value lies within 1 SD either direction from the mean is always 68%
- Probability that a value lies within 2 SD in either direction from the mean is always 95%
- Probability that a value lies within 3 SD in either direction from the mean is always 99.7%

Above rule is very useful for
1. Testing whether a distribution is NORMAL
2. Finding outliers: Any values more than 3 σ away from the mean can be treated as outliers.



# Sampling

Let’s take a Microbiologist who learn about Fish. Normally he does the following to learn about all fishes in the sea.

1. Catch **some fish**. We refer this some fish as **Sample** and the process of catching some fish is called **Sampling.**
2. Then he studies the caught fish.
3. And finally draw conclusions about all the fish in the sea. We call all the fish in the sea as **population** and the process of drawing conclusions about the population by observing the sample is called **Generalization.**

In simple word, Sampling means drawing conclusions about the population by observing the sample and generalizing. The conclusion which we draw is called inference. To draw meaning inference, we learn few techniques and hypothesis testing.

Below are few sampling use cases which are becoming very common across industry.

- Pschological studies
- Polling
- Drug Trails
- A/B Tests
- Market Research Surveys

**Sample statistics** – We describe the sample which is subset of the entire population using sample statistics. The sample statistics usually characterize the sample and not the population.

Sample statistics can broadly classified into two type -

1. Sample Mean - used when the variable is continuous like
   • Height of a group of people
   • User engagement on a website

2. Sample percentage – normally used when the variable is Binary (yes/no) like
   • Do you support this candidate?
   • Is this drug an effective treatment?

Standard deviation (SD) of the sample is critical as it will allow us to determine the confidence interval around the assertion which we will manner.

The manner in which we calculate the standard deviation of the sample is different for sample mean and sample percentage.

Below shows standard deviation of the same when we are interested in sample mean



Below shows standard deviation of the same when we are interested in sample percentage



where p is the % of YESes

# Sample Statistics and Sampling Distribution

Let us pick 100 different samples of any dataset from any population. For each of these samples, you then compute the sample mean or % as shown below. These computed sample mean or % will vary as it is different for each sample and hence, we call it as random variable.



If we plot the histogram of these values, that will represent its probability distribution. It turns out whenever you take a large number of samples, the sample % or sample mean of those samples follows a normal distribution. This normal distribution is also known as the sampling distribution.



Now we will see how to calculate sample mean and SD of the sampling distribution with one sample, as in near life we usually have limited samples. Sampling distribution of the standard deviation is termed as Standard error σ.



The best estimate of a sample distribution’s mean is its sample average ie, μ = x̄

For sample mean, σ = SD / SQRT (N), where N is the number of points in our datasets

For Sample percentage, σ = SD

So as of now, we have

1. Sample mean
2. Standard deviation of the sample
3. Sample standard error

Using these 3 numbers, we can now make assertion or inference about the population. Most of the inference fall under few specific types and there are standard procedures involved for each type to draw inference. Below are the inference types and an example for each types

1. Identify the population mean
- Indian Police are on average 80KG +/- 5 KG with 95% confidence

2. Identify the population percentage
- 20% +/- 2% of software engineers in a given city goes for morning walk

3. Verify if population mean is equal to a certain value
- Is the average life expectancy of Indians is 70 years?

4. Verify if population percentage is equal to a certain value
- 20% of people who took the drug has a side effect

5. Verify if 2 population means are different
- Indians are on average taller than Chinese

6. Verify if 2 population percentage are different
- Only 10% of people who don’t take the drug get better, but 80% of people who take the drug better.