Learning Outcome

Introduction to descriptive statistics and inferential statistics, measure of central tendency and spread.

Types of distributions-uniform, binomial, normal, log, exp

Sampling techniques, population

## 

## 

# Introduction to Descriptive statistics

## What are descriptive statistics?

In Descriptive statistics you are describing, presenting, summarizing, and organizing your data, either through numerical calculations or graphs or tables.

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

## Understanding Descriptive Statistics

Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set.

For example, the sum of the following data set is 20: (2, 3, 4, 5, 6). The mean is 4 (20/5). The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set. It is the figure separating the higher figures from the lower figures within a data set. However, there are fewer common types of descriptive statistics that are still very important.

People use descriptive statistics to repurpose hard-to-understand quantitative insights across a large data set into bite-sized descriptions. A student's grade point average (GPA), for example, provides a good understanding of descriptive statistics. The idea of a GPA is that it takes data points from a wide range of exams, classes, and grades, and averages them together to provide a general understanding of a student's overall academic performance. A student's personal GPA reflects their mean academic performance.

## Types of Descriptive Statistics

All descriptive statistics are either measures of central tendency or measures of variability, also known as measures of dispersion.

## Measure of Central Tendency

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analyzed data.

Measures of central tendency describe the center position of a distribution for a data set. A person analyzes the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analyzed data set.

There are three main measures of central tendency:

1. **Mean:** It is the sum of the observation divided by the sample size. It is not a robust statistic as it is affected by extreme values. So, very large or very low value (i.e., Outliers) can distort the answer.
2. **Median:** It is the middle value of data. It splits the data in half and also called 50th percentile. It is much less affected by the outliers and skewed data than mean. If the no. of elements in the dataset is odd, the middle most element is the median. If the no. of elements in the dataset is even, the median would be the average of two central elements.
3. **Mode:** It is the value that occurs more frequently in a dataset. Therefore, a dataset has no mode, if no category is the same and also possible that a dataset has more than one mode. It is the only measure of central tendency that can be used for categorical variables.

Were,

L = Lower limit Mode of modal class

fm = Frequency of modal class

f1 = Frequency of class preceding the modal class

f2= Frequency of class succeeding the modal class

h = Size of class interval

## Measures of Variability

Measures of variability (or the measures of spread) aid in analyzing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.

So, while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).

## Why do we need statistics that simply describe data?

Descriptive statistics are used to describe or summarize the characteristics of a sample or data set, such as a variable's mean, standard deviation, or frequency. Inferential statistics can help us understand the collective properties of the elements of a data sample. Knowing the sample mean, variance, and distribution of a variable can help us understand the world around us.

## What are mean and standard deviation?

These are two commonly employed descriptive statistics. Mean is the average level observed in some piece of data, while standard deviation describes the variance, or how dispersed the data observed in that variable is distributed around its mean.

## Can descriptive statistics be used to make inference or prediction?

No. While these descriptive helps understand data attributes, inferential statistical techniques—a separate branch of statistics—are required to understand how variables interact with one another in a data set.

## Measures of Spread

## Introduction

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.

## When can we measure spread?

The spread of the values can be measured for quantitative data, as the variables are numeric and can be arranged into a logical order with a low end value and a high end value.

## Why is it important to measure the spread of data?

There are many reasons why the measure of the spread of data values is important, but one of the main reasons regards its relationship with measures of central tendency. A measure of spread gives us an idea of how well the mean, for example, represents the data. If the spread of values in the data set is large, the mean is not as representative of the data as if the spread of data is small. This is because a large spread indicates that there are probably large differences between individual scores. The most common are:

1. The range (including the interquartile range and the interdecile range),
2. The standard deviation,
3. The variance,
4. Quartiles.

Calculating the Range

**Dataset A -** 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

The range is 4, the difference between the highest value (8) and the lowest value (4).

**Dataset B -** 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

The range is 10, the difference between the highest value (11) and the lowest value (1).

**Dataset A**

0 1 2 3 4 5 6 7 8 9 10 11 12 13

**Dataset B**

0 1 2 3 4 5 6 7 8 9 10 11 12 13

On a number line, you can see that the range of values for Dataset B is larger than Dataset A. Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

**Quartiles**

25% of values Q1 25% of values Q2 25% of values Q3 25% of values

The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.

The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.

The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.

**Calculating Quartiles**

**Dataset A**

4 5 5 Q1 5 6 6 Q2 6 6 7 Q3 7 7 8

As the quartile point falls between two values, the mean (average) of those values is the quartile value:

Q1 = (5+5) / 2 = 5

Q2 = (6+6) / 2 = 6

Q3 = (7+7) / 2 = 7

**Dataset B**

1 2 3 Q1 4 5 6 Q2 6 7 8 Q3 9 10 11

As the quartile point falls between two values, the mean (average) of those values is the quartile value:

Q1 = (3+4) / 2 = 3.5

Q2 = (6+6) / 2 = 6

Q3 = (8+9) / 2 = 8.5

The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest. The IQR is often seen as a better measure of spread than the range as it is not affected by outliers.

**Interquartile Range**

25% of values Q1 25% of values Q2 25% of values Q3

25% of values

Calculating the Interquartile Range

The IQR for Dataset A is = 2

IQR = Q3 - Q1 = 7 – 5 = 2

The IQR for Dataset B is = 5

IQR = Q3 - Q1 = 8.5 - 3.5 = 5

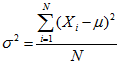
The variance and the standard deviation are measures of the spread of the data around the mean. They summaries how close each observed data value is to the mean value.

In datasets with a small spread all values are very close to the mean, resulting in a small variance and standard deviation. Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation.

The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero.

The standard deviation of a normal distribution enables us to calculate confidence intervals. In a normal distribution, about 68% of the values are within one standard deviation either side of the mean and about 95% of the scores are within two standard deviations of the mean.

The population Variance σ2 (pronounced sigma squared) of a discrete set of numbers is expressed by the following formula:



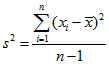
where:

Xi represents the ith unit, starting from the first observation to the last

μ represents the population mean

N represents the number of units in the population

The Variance of a sample s2 (pronounced s squared) is expressed by a slightly different formula:



where:

xi represents the ith unit, starting from the first observation to the last

x̅ represents the sample mean

n represents the number of units in the sample

The standard deviation is the square root of the variance. The standard deviation for a population is represented by σ, and the standard deviation for a sample is represented by s.

**Calculating the Population Variance σ2 and Standard Deviation σ**

**Dataset A**

Calculate the population mean (μ) of Dataset A.

(4 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 8) / 12

mean (μ) = 6

Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset = -2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 1, 2

Square each individual deviation value = 4, 1, 1, 1, 0, 0, 0, 0, 1,1,1, 4

Calculate the mean of the squared deviation values

= (4 + 1 +1 +1 + 0 + 0 + 0 + 0 +1 +1 +1 + 4) / 12

**Variance σ2= 1.17**

Calculate the square root of the variance

**Standard deviation σ = 1.08**

**Dataset B**

Calculate the population mean (μ) of Dataset B.

(1 + 2 + 3 + 4 + 5 + 6 + 6 + 7 + 8 + 9 + 10 + 11) / 12

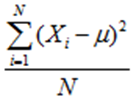
mean (μ) = 6

Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset  = -5, -4, -3, -2, -1, 0, 0, 1, 2, 3, 4, 5,

Square each individual deviation value

= 25, 16, 9, 4, 1, 0, 0, 1, 4, 9, 16, 25

Calculate the mean of the squared deviation values

 = (25 + 16 + 9 + 4 + 1 + 0 + 0 + 1 + 4 + 9 + 16 + 25) / 12

**Variance σ2 = 9.17**

Calculate the square root of the variance

**Standard deviation σ = 3.03**

The larger Variance and Standard Deviation in Dataset B further demonstrates that Dataset B is more dispersed than Dataset A.

## Inferential Statistics basics

## What is Inferential Statistic?

If you see based on the language, inferential means can be concluded.

In general, inferential statistics are a type of statistics that focus on processing sample data so that they can make decisions or conclusions on the population.

Inferential statistics focus on analyzing sample data to infer the population.

The flow of using inferential statistics is the sampling method, data analysis, and decision making for the entire population.

Inferential statistics are used by many people (especially scientist and researcher) because they are able to produce accurate estimates at a relatively affordable cost.

## Advantages of using inferential statistics

Inferential statistics have different benefits and advantages.

1. A precise tool for estimating population - The main purpose of using inferential statistics is to estimate population values. With the use of this method, of course, we expect accurate and precise measurement results and are able to describe the actual conditions.
2. Highly structured analytical methods - Inferential statistics have a very neat formula and structure. The method used is tested mathematically and can be regarded as an unbiased estimator.

## Knowing and using inferential statistics

## Inferential Statistics Examples

There are lots of examples of applications and the application of inferential statistics in life. However, in general, the inferential statistics that are often used are:

**1. Regression Analysis -** Regression analysis is one of the most popular analysis tools. Regression analysis is used to predict the relationship between independent variables and the dependent variable.

Using this analysis, we can determine which variables have a significant effect in a study.

For example, you want to know what factors can influence the decline in poverty. You use variables such as road length, economic growth, electrification ratio, number of teachers, number of medical personnel, etc.

After analysis, you will find which variables have an influence in reducing the poverty rate.

**2. Hypothesis test -** Hypothesis testing is a statistical test where we want to know the truth of an assumption or opinion that is common in society. Usually, this test is used to find out about the truth of a claim circulating in the community.

Hypothesis testing also helps us to prove whether the opinions or things we believe are true or false.

For example, we often hear the assumption that female students tend to have higher mathematical values ​​than men. Is that right?

To prove this, you can take a representative sample and analyze the mathematical values ​​of the samples taken.

By using a hypothesis test, you can draw conclusions about the actual conditions.

Can you use the entire data on the overall mathematics value of students and analyze the data? Certainly, very allowed.

But, of course, you will need a longer time in reaching conclusions because the data collection process also requires substantial time.

**3. Confidence Interval -** Confidence interval or confidence level is a statistical test used to estimate the population by using samples. With this level of trust, we can estimate with a greater probability what the actual population value is.

When using confidence intervals, we will find the upper and lower limits of a statistical test that we believe there is a population value we estimate.

When we use 95 percent confidence intervals, it means we believe that the test statistics we use are within the range of values ​​we have obtained based on the formula.

For example, we want to estimate what the average expenditure is for everyone in city X. Therefore, research is conducted by taking a number of samples. The results of this study certainly vary.

Therefore, we must determine the estimated range of the actual expenditure of each person. The hope is, of course, the actual average value will fall in the range of values ​​that we have calculated before.

At the last part of this article, I will show you how confidence interval works as inferential statistics examples.

**4. Time series analysis -** As you know, one type of data based on time is time series data. Sometimes, often a data occurs repeatedly or has special and common patterns so it is very interesting to study more deeply.

Time series analysis is one type of statistical analysis that tries to predict an event in the future based on pre-existing data. With this method, we can estimate how predictions a value or event that appears in the future.

Example: every year, policymakers always estimate economic growth, both quarterly and yearly. By using time series analysis, we can use data from 20 to 30 years to estimate how economic growth will be in the future.

## Procedure for using inferential statistics

1. Determine the population data that we want to examine
2. Determine the number of samples that are representative of the population
3. Select an analysis that matches the purpose and type of data we have
4. Make conclusions on the results of the analysis

## Differences in Inferential Statistics and Descriptive Statistics

Inferential statistics and descriptive statistics have very basic differences in the analysis process. In general, these two types of statistics also have different objectives.

1. Descriptive statistics aim to describe the characteristics of the data. While statistical inferencing aims to draw conclusions for the population by analyzing the sample.
2. Descriptive statistics are usually only presented in the form of tables and graphs. The test statistics used are fairly simple, such as averages, variances, etc. While inferential statistics, the statistics used are classified as very complicated. Not everyone is able to use inferential statistics so special seriousness and learning are needed before using it.

Therefore, we cannot use any analytical tools available in descriptive analysis to infer the overall data.

## How to make inferential statistics as a stronger tool?

Probably, the analyst knows several things that can influence inferential statistics in order to produce accurate estimates. The main key is good sampling.

Samples taken must be random or random. That is, there should not be certain trends in taking who, what, and how the condition of the sample.

The selected sample must also meet the minimum sample requirements. Actually, there is no specific requirement for the number of samples that must be used to be able to represent the population. However, many experts agree that the number of samples used must be at least 30 units.

Samples must also be able to meet certain distributions. Usually, the commonly used sample distribution is a normal distribution. Although sometimes, there are cases where other distributions are indeed more suitable.

Make sure the above three conditions are met so that your analysis results don’t disappoint later.

## Why we should use inferential statistics?

Case Study of Inferential Statistics

There are several types of inferential statistics examples that you can use. But in this case, I will just give an example using statistical confidence intervals.

Suppose a regional head claims that the poverty rate in his area is very low. To prove this, he conducted a household income and expenditure survey that was theoretically able to produce poverty.

Considering the survey period and budget, 10,000 household samples were selected from a total of 100,000 households in the district.

Based on the survey results, it was found that there were still 5,000 poor people. Of course, this number is not entirely true considering the survey always has errors.

Therefore, confidence intervals were made to strengthen the results of this survey.

Based on the results of calculations, with a confidence level of 95 percent and the standard deviation is 500, it can be concluded that the number of poor people in the city ranges from 4,990 to 5010 people.

## Types of Distributions

## Bernoulli Distribution

Let’s start with the easiest distribution that is Bernoulli Distribution. It is actually easier to understand than it sounds!

All you cricket junkies out there! At the beginning of any cricket match, how do you decide who is going to bat or ball? A toss! It all depends on whether you win or lose the toss, right? Let’s say if the toss results in a head, you win. Else, you lose. There’s no midway.

A **Bernoulli distribution** has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So, the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure.

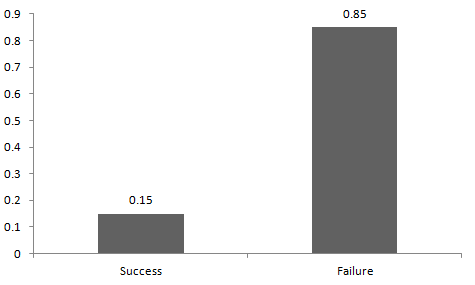
Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

The probability mass function is given by: px(1-p)1-x where x € (0, 1).

It can also be written as



The probabilities of success and failure need not be equally likely, like the result of a fight between me and Undertaker. He is pretty much certain to win. So in this case probability of my success is 0.15 while my failure is 0.85



Here, the probability of success(p) is not same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight.

Here, the probability of success = 0.15 and probability of failure = 0.85. The expected value is exactly what it sounds. If I punch you, I may expect you to punch me back. Basically expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Bernoulli distribution is found as follows:

E(X) = 1\*p + 0\*(1-p) = p

The variance of a random variable from a bernoulli distribution is:

V(X) = E(X²) – [E(X)]² = p – p² = p(1-p)

There are many examples of Bernoulli distribution such as whether it’s going to rain tomorrow or not where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

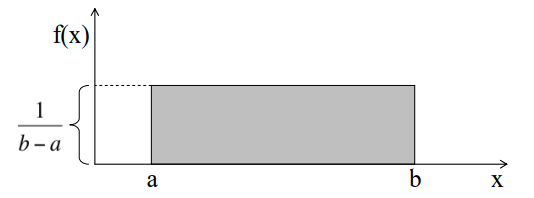
## Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniformly distributed if the density function is:



The graph of a uniform distribution curve looks like



You can see that the shape of the Uniform distribution curve is rectangular, the reason why Uniform distribution is called rectangular distribution.

For a Uniform Distribution, a and b are the parameters.

The number of bouquets sold daily at a flower shop is uniformly distributed with a maximum of 40 and a minimum of 10.

Let’s try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)\*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is = 0.667

The mean and variance of X following a uniform distribution is:

Mean -> E(X) = (a+b)/2

Variance -> V(X) = (b-a)²/12

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:



## Binomial Distribution

Let’s get back to cricket. Suppose that you won the toss today and this indicates a successful event. You toss again but you lost this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Let’s assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

The outcomes need not be equally likely. Remember the example of a fight between me and Undertaker? So, if the probability of success in an experiment is 0.2 then the probability of failure can be easily computed as q = 1 – 0.2 = 0.8.

Each trial is independent since the outcome of the previous toss doesn’t determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

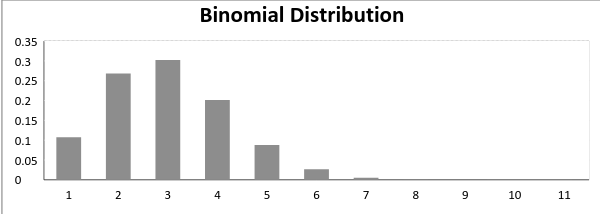
On the basis of the above explanation, the properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

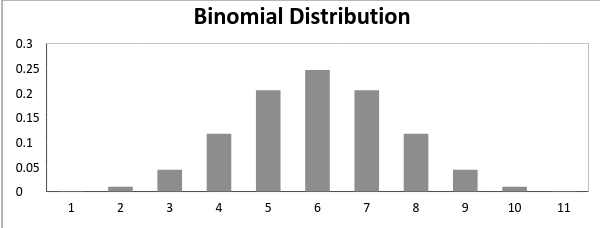
The mathematical representation of binomial distribution is given by:



A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Now, when probability of success = probability of failure, in such a situation the graph of binomial distribution looks like



The mean and variance of a binomial distribution are given by:

Mean -> µ = n\*p

Variance -> Var(X) = n\*p\*q

## Normal Distribution

Normal distribution represents the behavior of most of the situations in the universe (That is why it’s called a “normal” distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

1. The mean, median and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center and the other half to the right.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity, then the shapes will be quite similar.

The PDF of a random variable X following a normal distribution is given by:



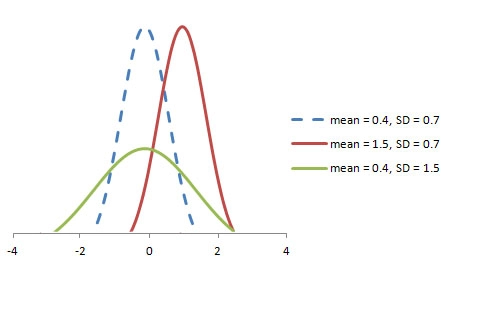
The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean -> E(X) = µ

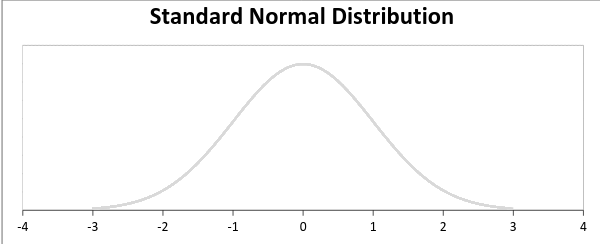
Variance -> Var(X) = σ^2

Here, µ (mean) and σ (standard deviation) are the parameters.

The graph of a random variable X ~ N (µ, σ) is shown below.



A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1. For such a case, the PDF becomes:



## Poisson Distribution

Suppose you work at a call center; approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are

1. The number of emergency calls recorded at a hospital in a day.
2. The number of thefts reported in an area on a day.
3. The number of customers arriving at a salon in an hour.
4. The number of suicides reported in a particular city.
5. The number of printing errors at each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called Poisson distribution when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

* λ is the rate at which an event occurs,
* t is the length of a time interval,
* And X is the number of events in that time interval.

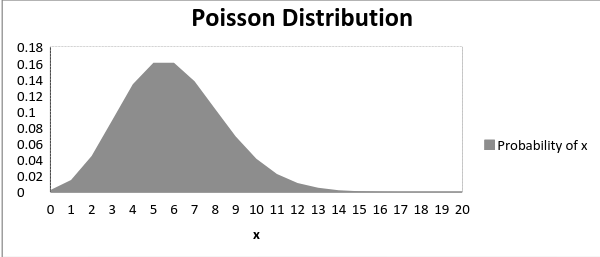
Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Let µ denote the mean number of events in an interval of length t. Then, µ = λ\*t.

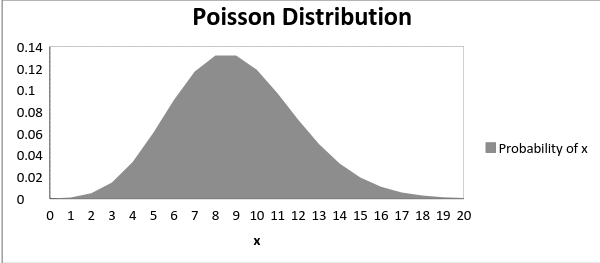
The PMF of X following a Poisson distribution is given by:



The mean µ is the parameter of this distribution. µ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:



The graph shown below illustrates the shift in the curve due to increase in mean.



It is perceptible that as the mean increases, the curve shifts to the right.

The mean and variance of X following a Poisson distribution:

Mean -> E(X) = µ

Variance -> Var(X) = µ

## Exponential Distribution

Let’s consider the call center example one more time. What about the interval of time between the calls? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

Other examples are:

1. Length of time between metro arrivals,
2. Length of time between arrivals at a gas station
3. The life of an Air Conditioner

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

A random variable X is said to have an exponential distribution with PDF:

f(x) = { λe-λx, x ≥ 0

and parameter λ>0 which is also called the rate.

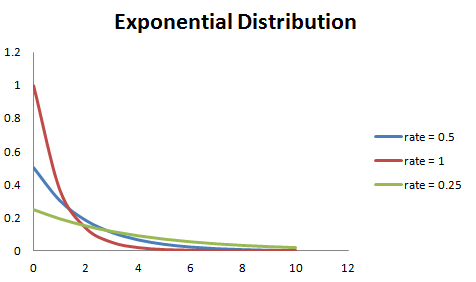
For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

Mean and Variance of a random variable X following an exponential distribution:

Mean -> E(X) = 1/λ

Variance -> Var(X) = (1/λ)²

Also, the greater the rate, the faster the curve drops and the lower the rate, flatter the curve. This is explained better with the graph shown below.



To ease the computation, there are some formulas given below.

P{X≤x} = 1 – e-λx, corresponds to the area under the density curve to the left of x.

P{X>x} = e-λx, corresponds to the area under the density curve to the right of x.

P{x1<X≤ x2} = e-λx1 – e-λx2, corresponds to the area under the density curve between x1 and x2.

## What is sampling?

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country’s population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug’s behavior.

**Types of sampling: sampling methods**

Sampling in market research is of two types – probability sampling and non-probability sampling. Let’s take a closer look at these two methods of sampling.

**Probability sampling:** Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.

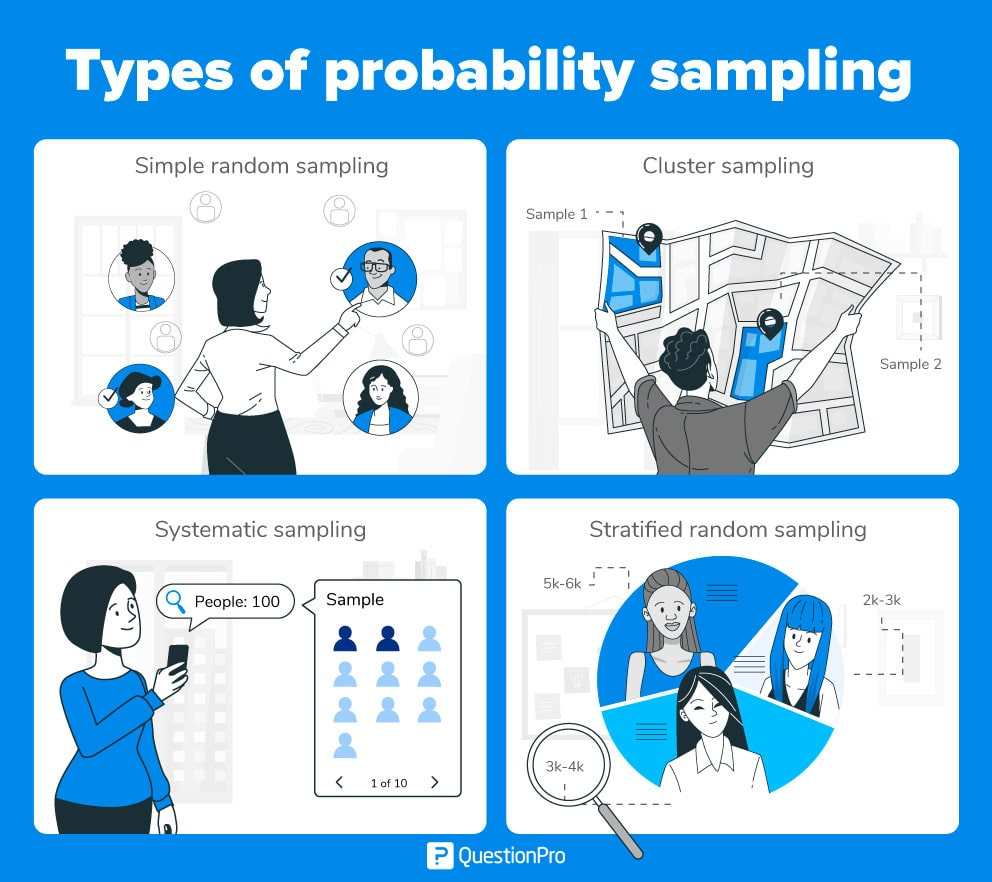
**Non-probability sampling:** In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

**Types of probability sampling with examples:**

Probability sampling is a sampling technique in which researchers choose samples from a larger population using a method based on the theory of probability. This sampling method considers every member of the population and forms samples based on a fixed process.

For example, in a population of 1000 members, every member will have a 1/1000 chance of being selected to be a part of a sample. Probability sampling eliminates bias in the population and gives all members a fair chance to be included in the sample.

There are four types of probability sampling techniques:



Types of probability sampling

**Simple random sampling:** One of the best probability sampling techniques that helps in saving time and resources, is the Simple Random Sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.

For example, in an organization of 500 employees, if the HR team decides on conducting team building activities, it is highly likely that they would prefer picking chits out of a bowl. In this case, each of the 500 employees has an equal opportunity of being selected.

**Cluster sampling:** Cluster sampling is a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback.

For example, if the United States government wishes to evaluate the number of immigrants living in the mainland US, they can divide it into clusters based on states such as California, Texas, Florida, Massachusetts, Colorado, Hawaii, etc. This way of conducting a survey will be more effective as the results will be organized into states and provide insightful immigration data.

**Systematic sampling:** Researchers use the systematic sampling method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.

For example, a researcher intends to collect a systematic sample of 500 people in a population of 5000. He/she numbers each element of the population from 1-5000 and will choose every 10th individual to be a part of the sample (Total population/ Sample Size = 5000/500 = 10).

**Stratified random sampling:** Stratified random sampling is a method in which the researcher divides the population into smaller groups that don’t overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.

For example, a researcher looking to analyze the characteristics of people belonging to different annual income divisions will create strata (groups) according to the annual family income. Eg – less than $20,000, $21,000 – $30,000, $31,000 to $40,000, $41,000 to $50,000, etc. By doing this, the researcher concludes the characteristics of people belonging to different income groups. Marketers can analyze which income groups to target and which ones to eliminate to create a roadmap that would bear fruitful results.

## Uses of probability sampling

There are multiple uses of probability sampling:

**Reduce Sample Bias:** Using the probability sampling method, the bias in the sample derived from a population is negligible to non-existent. The selection of the sample mainly depicts the understanding and the inference of the researcher. Probability sampling leads to higher quality data collection as the sample appropriately represents the population.

**Diverse Population:** When the population is vast and diverse, it is essential to have adequate representation so that the data is not skewed towards one demographic. For example, if Square would like to understand the people that could make their point-of-sale devices, a survey conducted from a sample of people across the US from different industries and socio-economic backgrounds helps.

Create an Accurate Sample: Probability sampling helps the researchers plan and create an accurate sample. This helps to obtain well-defined data.

## Types of non-probability sampling with examples

The non-probability method is a sampling method that involves a collection of feedback based on a researcher or statistician’s sample selection capabilities and not on a fixed selection process. In most situations, the output of a survey conducted with a non-probable sample leads to skewed results, which may not represent the desired target population. But, there are situations such as the preliminary stages of research or cost constraints for conducting research, where non-probability sampling will be much more useful than the other type.

Four types of non-probability sampling explain the purpose of this sampling method in a better manner:

**Convenience sampling:** This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as convenience sampling, because of the researcher’s ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it’s purely done based on proximity and not representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

For example, startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the mall entrance and giving out pamphlets randomly.

**Judgmental or purposive sampling:** Judgemental or purposive samples are formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master’s degree. The selection criteria will be: “Are you interested in doing your masters in …?” and those who respond with a “No” are excluded from the sample.

**Snowball sampling:** Snowball sampling is a sampling method that researchers apply when the subjects are difficult to trace. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.

**Quota sampling:** In Quota sampling, the selection of members in this sampling technique happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.

## Uses of non-probability sampling

Non-probability sampling is used for the following:

**Create a hypothesis:** Researchers use the non-probability sampling method to create an assumption when limited to no prior information is available. This method helps with the immediate return of data and builds a base for further research.

**Exploratory research:** Researchers use this sampling technique widely when conducting qualitative research, pilot studies, or exploratory research.

**Budget and time constraints:** The non-probability method when there are budget and time constraints, and some preliminary data must be collected. Since the survey design is not rigid, it is easier to pick respondents at random and have them take the survey or questionnaire.

## How do you decide on the type of sampling to use?

For any research, it is essential to choose a sampling method accurately to meet the goals of your study. The effectiveness of your sampling relies on various factors. Here are some steps expert researchers follow to decide the best sampling method.

* Jot down the research goals. Generally, it must be a combination of cost, precision, or accuracy.
* Identify the effective sampling techniques that might potentially achieve the research goals.
* Test each of these methods and examine whether they help in achieving your goal.
* Select the method that works best for the research.

## Difference between probability sampling and non-probability sampling methods

We have looked at the different types of sampling methods above and their subtypes. To encapsulate the whole discussion, though, the significant differences between probability sampling methods and non-probability sampling methods are as below:

|  |  |  |
| --- | --- | --- |
|  | Probability Sampling Methods | Non-Probability Sampling Methods |
| Definition | Probability Sampling is a sampling technique in which samples from a larger population are chosen using a method based on the theory of probability. | Non-probability sampling is a sampling technique in which the researcher selects samples based on the researcher’s subjective judgment rather than random selection. |
| Alternatively Known as | Random sampling method. | Non-random sampling method |
| Population selection | The population is selected randomly. | The population is selected arbitrarily. |
| Nature | The research is conclusive. | The research is exploratory. |
| Sample | Since there is a method for deciding the sample, the population demographics are conclusively represented. | Since the sampling method is arbitrary, the population demographics representation is almost always skewed. |
| Time Taken | Takes longer to conduct since the research design defines the selection parameters before the market research study begins. | This type of sampling method is quick since neither the sample or selection criteria of the sample are undefined. |
| Results | This type of sampling is entirely unbiased and hence the results are unbiased too and conclusive. | This type of sampling is entirely biased and hence the results are biased too, rendering the research speculative. |
| Hypothesis | In probability sampling, there is an underlying hypothesis before the study begins and the objective of this method is to prove the hypothesis. | In non-probability sampling, the hypothesis is derived after conducting the research study. |

References

1. <https://www.investopedia.com/terms/d/descriptive_statistics.asp>
2. <https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language+-+measures+of+spread>
3. <https://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php>
4. <https://worldsustainable.org/inferential-statistics-examples/>
5. <https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/>
6. <https://www.questionpro.com/blog/types-of-sampling-for-social-research/>