Introduction to Statistics

Source: https://collegelearners.com/msc-statistics-and-data-science-edinburgh/

Introduction to Statistics

Statistics is a discipline of mathematics that is globally agreed to be prerequisite for a more profound understanding of machine learning.

Although Statistics is a large field with many obscure theories and findings, the nuts and bolt tools, and notations taken from the field are required for machine learning practitioners. With a firm foundation of what statistics is, it is possible to focus on just the good or relevant parts.

To be the best Data Scientists you can be, your skills in statistical understanding should be well-established. The more you appreciate statistics, the better you will understand how machine learning performs its apparent magic.

Introduction to Statistics

Statistic is the key concept for Data Science, without Stats we never think about data, so we can say that Stats is the heart of Data Science. We all learn **Statistic** in our school days but after school we never use **Statistic for** solving real world problem, but in the field of **data** we have to use **Statistic** to solve the business problem ,to see the insight or pattern from the **Data.**

We will learn how to learn stats for Data Science in a much simpler way,.

What is Statistics?

Statistics is a branch or field of a scientific study, which consists of:

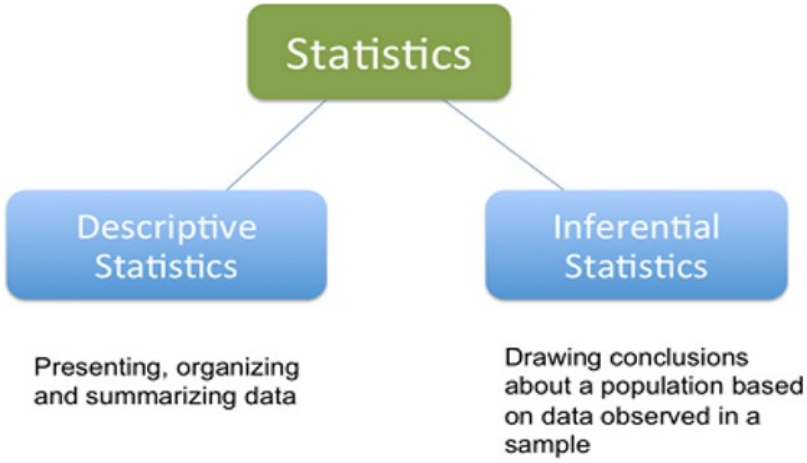
- Collecting
- Classifying
- Organizing
- Summarizing
- Interpreting the results

Introduction to Statistics

When is comes to the statistical tools that we use in practice, it can be helpful to divide the field of statistics into two large groups of methods: descriptive statistics for summarizing data, and inferential statistics for drawing conclusions from samples of data.

Descriptive Statistics is the branch of statistics that involves organizing, displaying, and describing data. It is a major branch of statistics that supports describing a huge amount of data with charts and tables. It neither permits us to draw inferences about the population nor reach an inference regarding any of hypothesis.

Inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population.

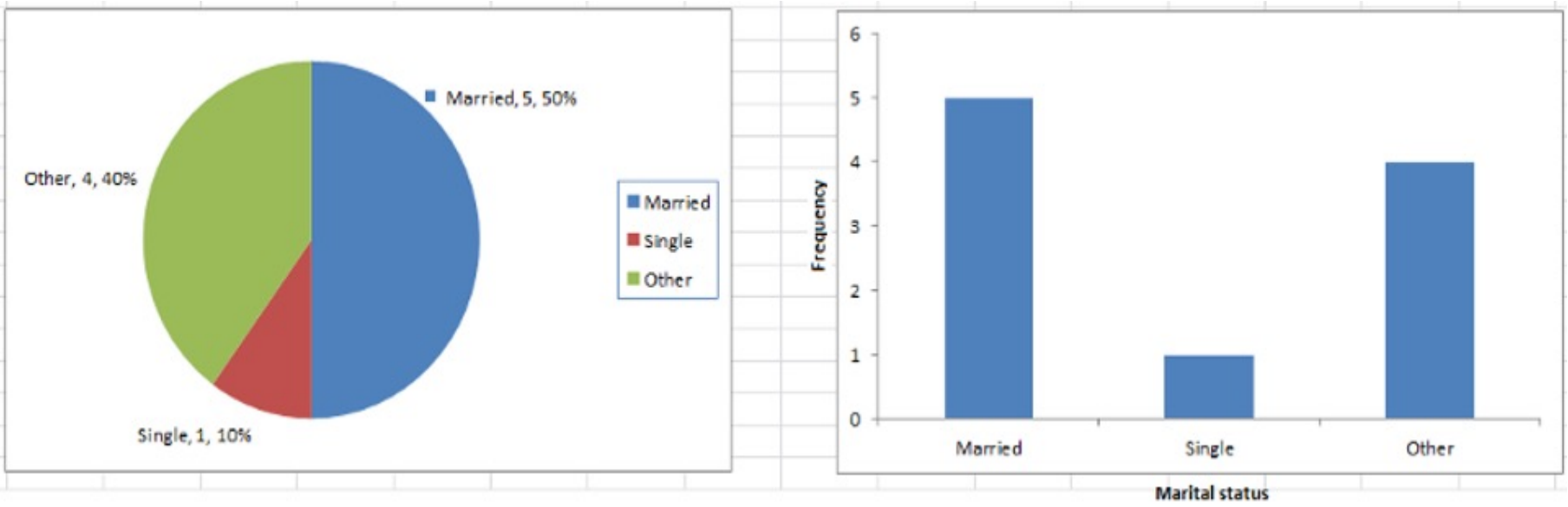


Descriptive Statistics

Descriptive statistics refers to methods for summarizing and organizing the information is a data set. It's of two types:

Uni-variate Descriptive Statistics

Different ways one can describe patterns found in uni-variable data include central tendency: mean, mode, and median and dispersion: range, variance, maximum, minimum, quartiles, and standard deviation. The various plots used to visualize uni-variate data typically are Bar Charts, histograms, Pie Charts etc.



Bi-variate Descriptive Statistics

The bi-variate analysis involves the analysis of two variables for the purpose of determining the empirical relationship between them. The various plots used to visualize bi-variate data typically are scatter-plot, box-plot.

Introduction to Statistics

Before we dive deep into basics of statistics let us introduce the concept of Population and Sample.

We begin with a simple example. There are millions of passenger automobiles in the United States. What is their average value? It is obviously impractical to attempt to solve this problem directly by assessing the value of every single car in the country, add up all those values, then divide by the number of values, one for each car. In practice the best we can do would be to *estimate* the average value. A natural way to do so would be to randomly select *some* of the cars, say 200 of them, ascertain the value of each of those cars, and find the average of those 200 values. The set of all those millions of vehicles is called the *population* of interest, and the number attached to each one, its value, is a *measurement*. The average value is a *parameter*: a number that describes a characteristic of the population, in this case monetary worth. The set of 200 cars selected from the population is called a *sample*, and the 200 numbers, the monetary values of the cars we selected, are the *sample data*. The average of the data is called a *statistic.*

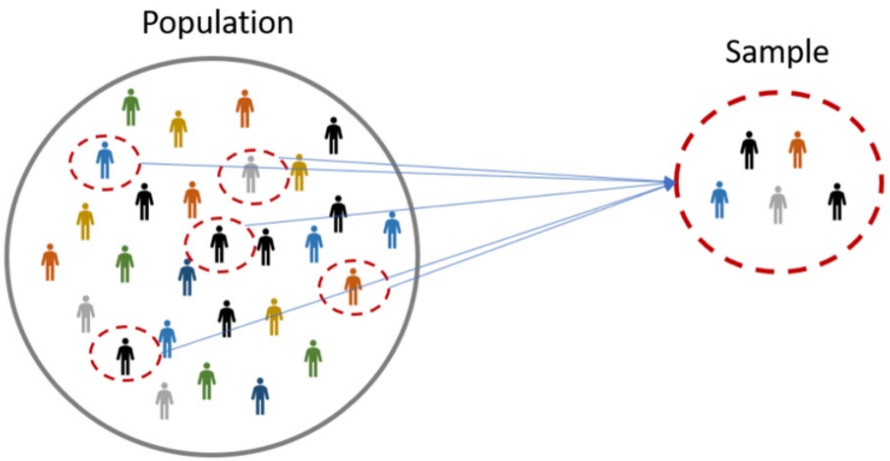
Introduction to Statistics

What is Population?

A collection or a group of entities. It is the entire group that you want to draw conclusions about. The description portion of the population is called a parameter. To describe and analyze population data is called **Descriptive Statistics.**

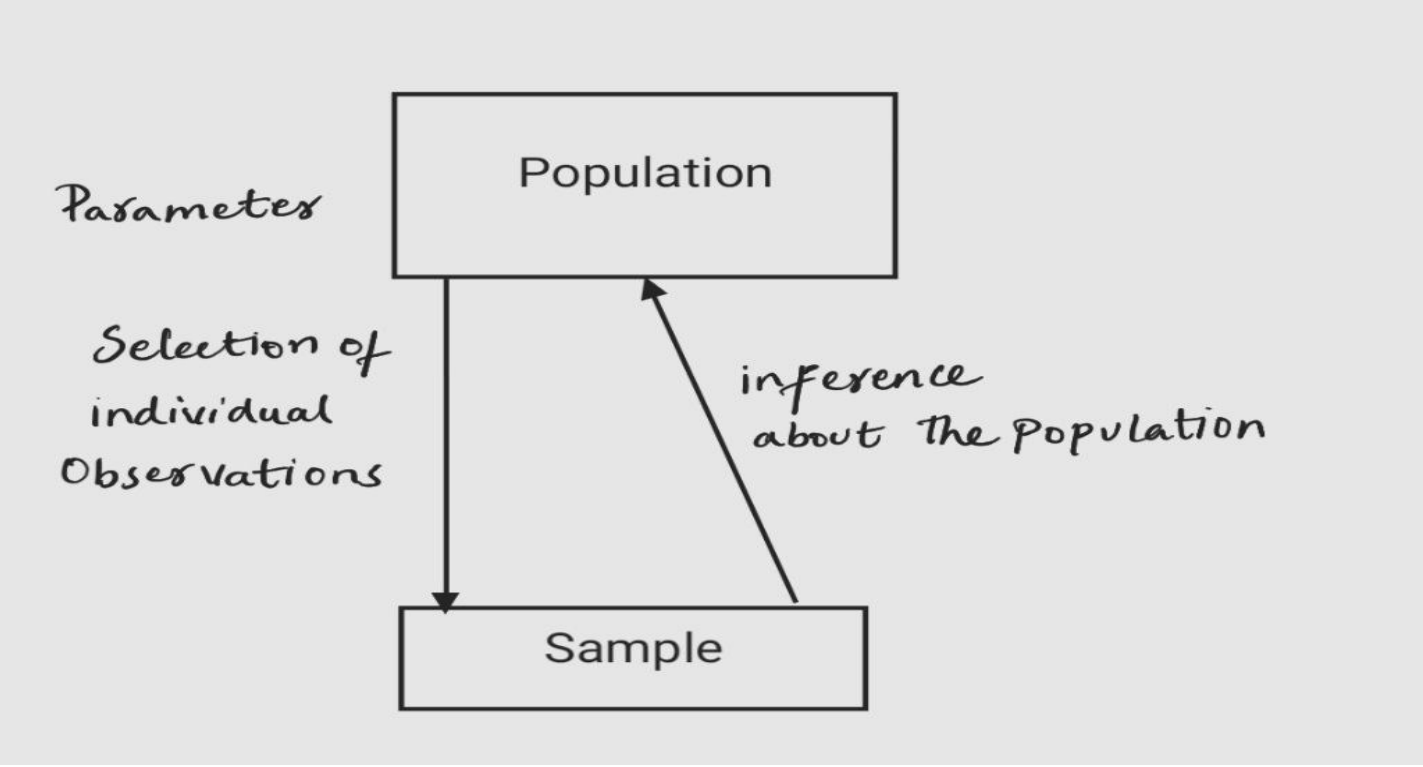
What is Sample?

When your population is large in size, geographically dispersed, or difficult to contact, it's necessary to use a sample. With analysis, you can use sample data to make estimates or test hypotheses about population data. A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.



Introduction to Statistics

Essentially, this boils down to:



Descriptive Statistics are also called Measure of Central Tendency.

Introduction to Statistics

Continuing with our example, if the average value of the cars in our sample was $8,357, then it seems reasonable to conclude that the average value of all cars is about $8,357. In reasoning this way, we have drawn an inference about the *population* based on information obtained from the *sample* . In general, *statistics* is a study of data: describing properties of the data, which is called ***descriptive statistics***, and drawing conclusions about a population of interest from information extracted from a sample, which is called ***inferential statistics***. Computing the single number $8,357 to summarize the data was an operation of descriptive statistics; using it to make a statement about the population was an operation of inferential statistics

.

Type of data

The measurement made on each element of a sample need not be numerical. In the case of automobiles, what is noted about each car could be its color, its make, its body type, and so on. Such data are categorical or qualitative, as opposed to numerical or quantitative data such as value or age. This is a general distinction.

Qualitative data are measurements for which there is no natural numerical scale, but which consist of attributes, labels, or other non-numerical characteristics. The qualitative variables are marital status, mortgage, rank and risk

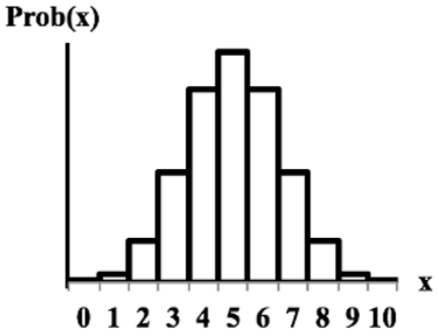
Quantitative data are numerical measurements that arise from a natural numerical scale. The quantitative variables are income and year.

Qualitative data can generate numerical sample statistics. In the automobile example, for instance, we might be interested in the proportion of all cars that are less than six years old. In our same sample of 200 cars, we could note for each car whether it is less than six years old or not, which is a qualitative measurement. If 172 cars in the sample are less than six years old, which is 0.860.or 86%, then we would estimate the parameter of interest, the population proportion, to be about the same as the sample statistic, the sample proportion, that is, about 0.860

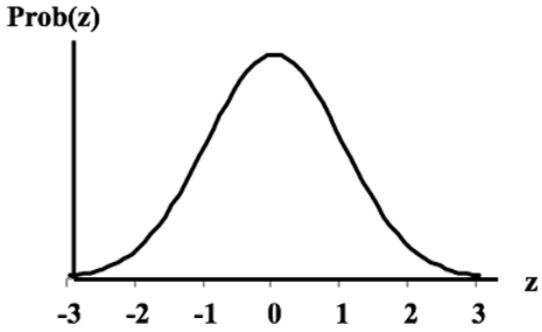
Type of data

Discrete Variable a numerical variable that can take either a finite or a countable number of values is a discrete variable, for which each value can be graphed as a separate point with space between each point. ‘year’ ais an example of discrete variable.

Continuous Variable Quant A numerical variable that can take infinitely many values is a continuous variable, whose possible values form an interval on the number lin, with no psace between the points. ‘income’ is an example of a continuous variable.

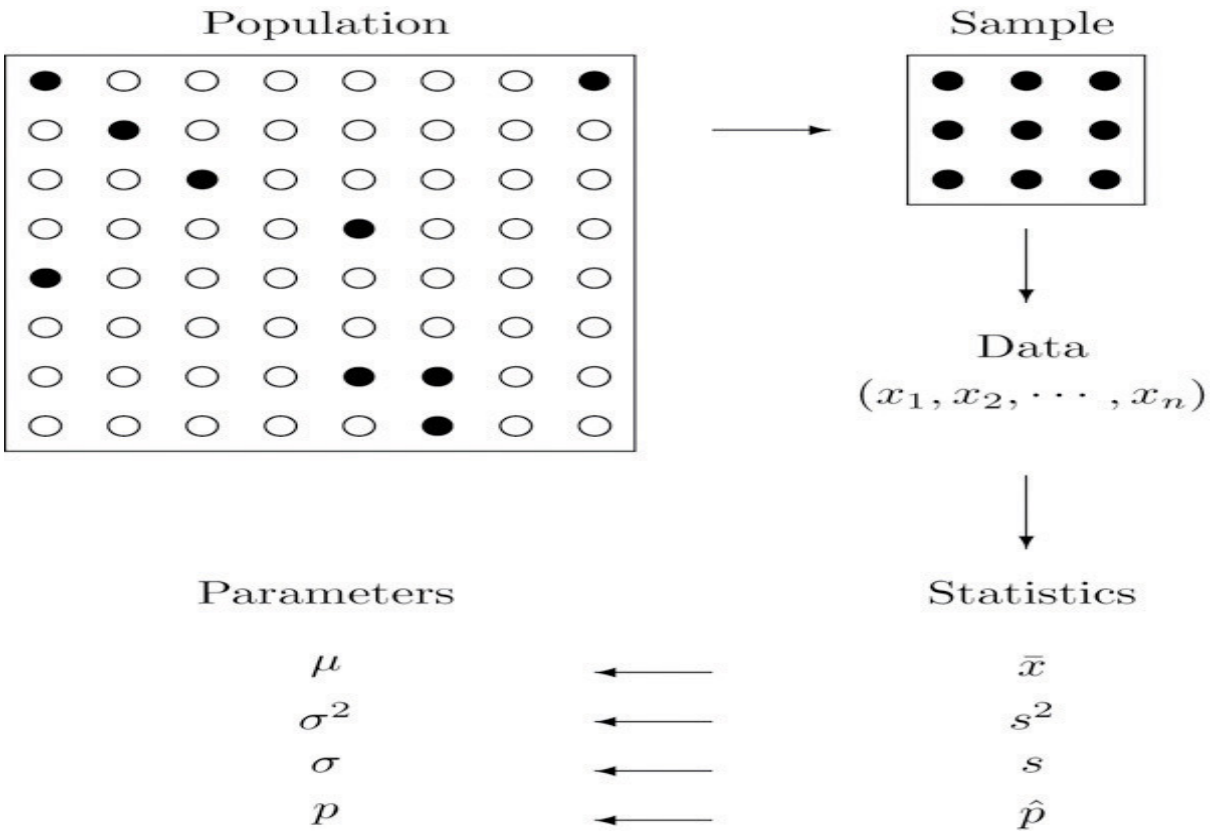


Binomial Distribution
Discrete Data & Discrete
Probability Curve



Standard Normal Distribution
Continuous Data and Continuous
Probability Curve

The relationship between a population of interest and a sample drawn from that population is perhaps the most important concept in statistics, since everything else rests on it. This relationship is illustrated graphically in the figure below. The circles in the large box represent elements of the population. In the figure there was room for only a small number of them but in actual situations, like our automobile example, they could very well number in the millions. The solid black circles represent the elements of the population that are selected at random and that together form the sample. For each element of the sample there is a measurement of interest, denoted by a lower case x (which we have indexed as $x1,\ldots,xn$ to tell them apart); these measurements collectively form the sample data set. From the data we may calculate various statistics. To anticipate the notation that will be used later, we might compute the sample mean $\bar{x}$ and the sample proportion $\hat{p}$, and take them as approximations to the population mean μ (this is the lowercase Greek letter mu, the traditional symbol for this parameter) and the population proportion p, respectively. The other symbols in the figure stand for other parameters and statistics that we will encounter.


Population
Sample
Data
$(x_1, x_2, \cdots, x_n)$
Parameters
Statistics
μ
$\bar{x}$
σ^2
s^2
σ
s
p
$\hat{p}$

Introduction to Statistics

What is Measure of Central Tendency?

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. We will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

Introduction to Statistics

Mean: Mean is the sum of observations divided by the sample size. It is not a robust statistic as it is affected by extreme values. So, very large or very low values, i.e., Outliers can distort the answer.

Introduction to Statistics

Mean: The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.. It is the sum of observations divided by the sample size.

$$\text{Mean} = \frac{\text{Sum of all individual data points}}{\text{total count of data}}$$

2 types of mean

$$\text{Population mean}(\mu) = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{Sample mean}\ (\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Introduction to Statistics

Example

A random sample of ten students is taken from the student body of a college and their GPAs are recorded as follows:

1.90,3.00,2.53,3.71,2.12,1.76,2.71,1.39,4.00,3.33

Find the mean.

Solution:

Mean = (1.90+3.00+2.53+3.71+2.12+1.76+2.71+1.39+4.00+3.33)/10=2.65

The mean of two numbers is the number that is halfway between them. For example, the average of the numbers 5 and 17 is (5+17)/2=11, which is 6 units above 5 and 6 units below 17. In this sense the average 11 is the "center" of the data set {5,17}. For larger data sets the mean can similarly be regarded as the "center" of the data.

Introduction to Statistics

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero

Introduction to Statistics

When not to use mean

It is not a robust statistic as it is affected by extreme values. So, very large or very low values, i.e. Outliers can distort the answer.

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values.

Introduction to Statistics

Median: The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

To see why another concept of average is needed, consider the following situation. Suppose we are interested in the average yearly income of employees at a large corporation. We take a random sample of seven employees, obtaining the sample data
(rounded to the nearest hundred dollars and expressed in thousands of dollars).

24.8,22.8,24.6,192.5,25.2,18.5,23.7

The mean (rounded to one decimal place) is $\bar{x}$=47.4, but the statement "the average income of employees at this corporation is $47,400" is surely misleading. It is approximately twice what six of the seven employees in the sample make and is nowhere near what any of them makes. It is easy to see what went wrong: the presence of the one executive in the sample whose salary is so large compared to everyone else's, caused the numerator in the formula for the sample mean to be far too large, pulling the mean far to the right of where we think that the average "ought" to be, namely around $24,000 or $25,000. The number 192.5 in our data set is called an outlier, a number that is far removed from most or all of the remaining measurements. Many times an outlier is the result of some sort of error, but not always, as is the case here.

Introduction to Statistics

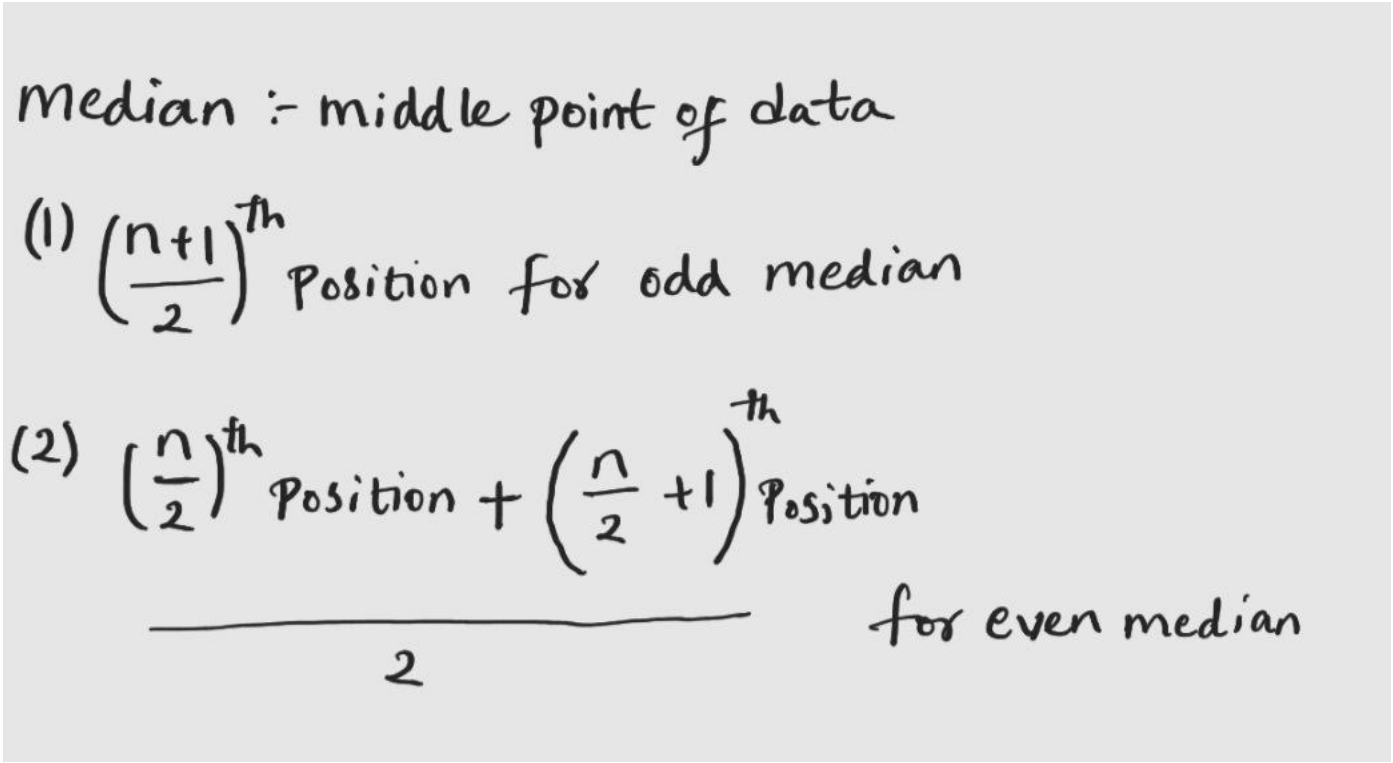
We would get a better measure of the “center” of the data if we were to arrange the data in numerical order:

18.5,22.8,23.7,24.6,24.8,25.2,192.5(2.2.4)

then select the middle number in the list, in this case 24.624.6. The result is called the median of the data set and has the property that roughly half of the measurements are larger than it is, and roughly half are smaller. In this sense it locates the center of the data. If there are an even number of measurements in the data set, then there will be two middle elements when all are lined up in order, so we take the mean of the middle two as the median.

Introduction to Statistics

Median: It is the middle value of the data. It splits the data in half and also called 50th Percentile. It is much less affected by the outliers and skewed data than the mean. If the number of elements in the dataset is odd, the middle most element is the median. If the number of elements in the dataset is even, the median would be the average of the two central elements.

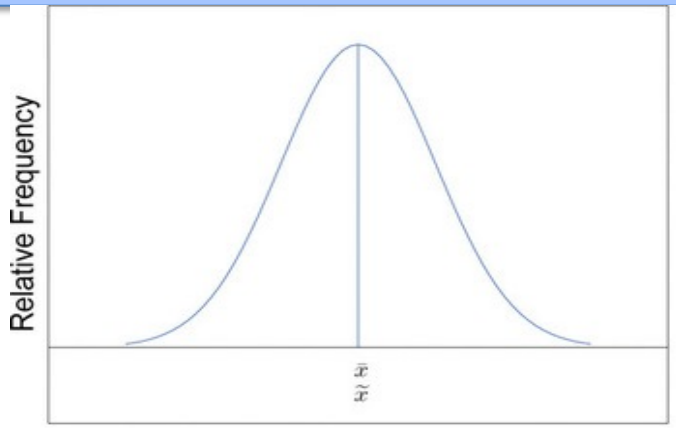


Introduction to Statistics

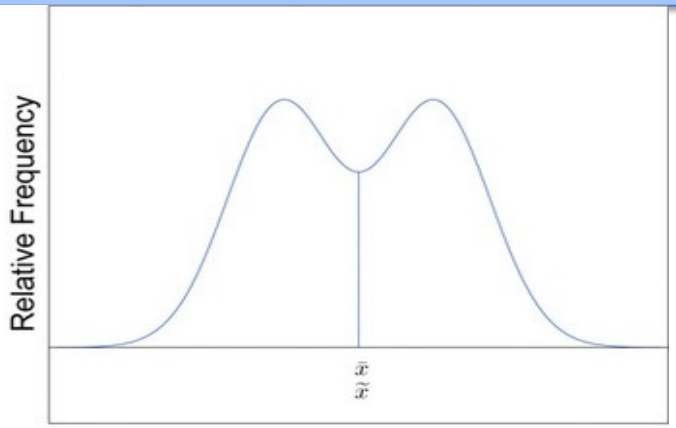
The relationship between the mean and the median for several common shapes of distributions is shown next. The distributions in panels (a) and (b) are said to be *symmetric* because of the symmetry that they exhibit. The distributions in the remaining two panels are said to be *skewed*. In each distribution we have drawn a vertical line that divides the area under the curve in half, which is located at the median. The following facts are true in general:

- When the distribution is symmetric, as in panels (a) and (b) of Figure, the mean and the median are equal.
- When the distribution is as shown in panel (c), it is said to be skewed right. The mean has been pulled to the right of the median by the long “right tail” of the distribution, the few relatively large data values.
- When the distribution is as shown in panel (d), it is said to be skewed left. The mean has been pulled to the left of the median by the long “left tail” of the distribution, the few relatively small data values.

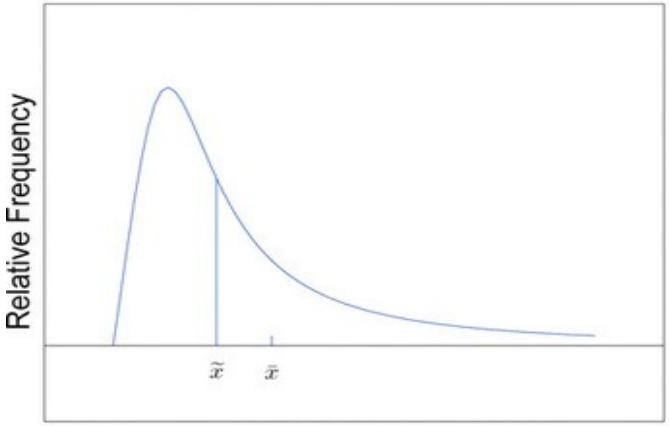
Introduction to Statistics



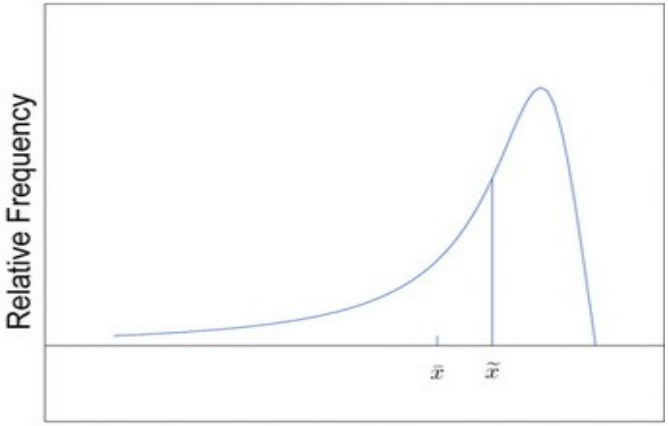
(a) $\bar{x} = \tilde{x}$



(b) $\bar{x} = \tilde{x}$

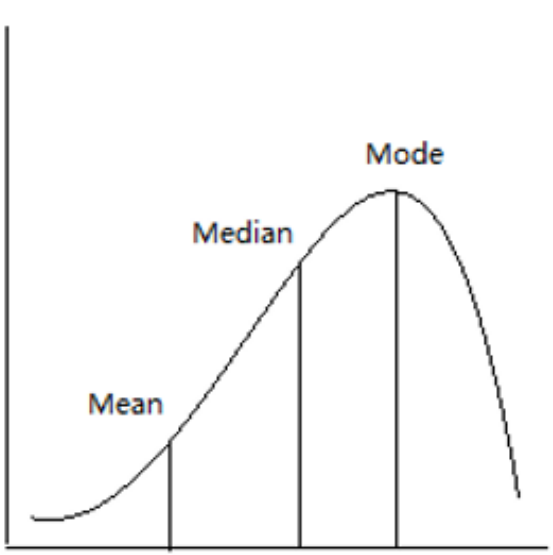


(c) $\bar{x} > \tilde{x}$

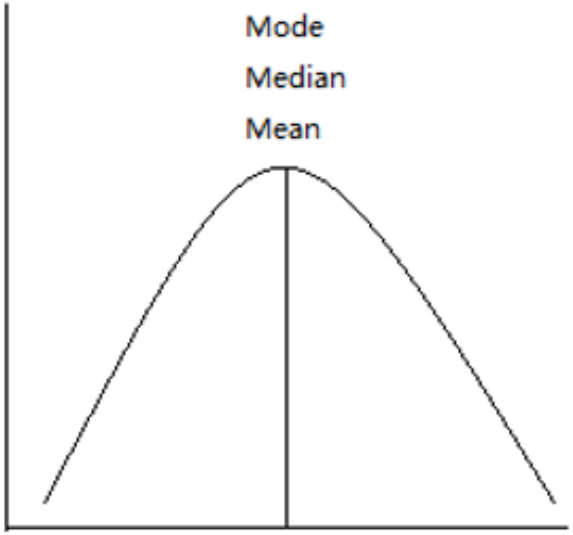


(d) $\bar{x} < \tilde{x}$

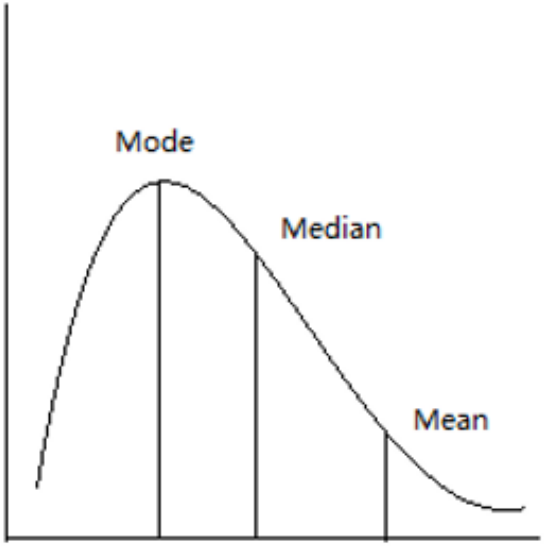
Comparison between mean median and mode



Left skew



Normal Distribution



Right skew

Introduction to Statistics

Mode: Perhaps you have heard a statement like "The average number of automobiles owned by households in the United States is 1.37," and have been amused at the thought of a fraction of an automobile sitting in a driveway. In such a context the following measure for central location might make more sense: Sample Mode

It is the value that occurs most 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.

The mode is a measure of central location since most real-life data sets have more observations near the center of the data range and fewer observations on the lower and upper ends. The value with the highest frequency is often in the middle of the data range.

If mean= median= mode. It is called **Symmetrical data**.

If mean ≠ median ≠ mode. It is called **Asymmetrical data.**

Introduction to Statistics

Summary of when to use mean, median and mode

Type of variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio(Skewed)	Median

Introduction to Statistics

Key Takeaway

The mean, the median, and the mode each answer the question “Where is the center of the data set?” The nature of the data set, as indicated by a relative frequency histogram, determines which one gives the best answer.

Introduction to Statistics

Drawbacks of measure of central tendency:

- By using measure of central tendency, we can calculate the middle values of the data, but we can't calculate min, max or how one data point is deviating from the other data point.

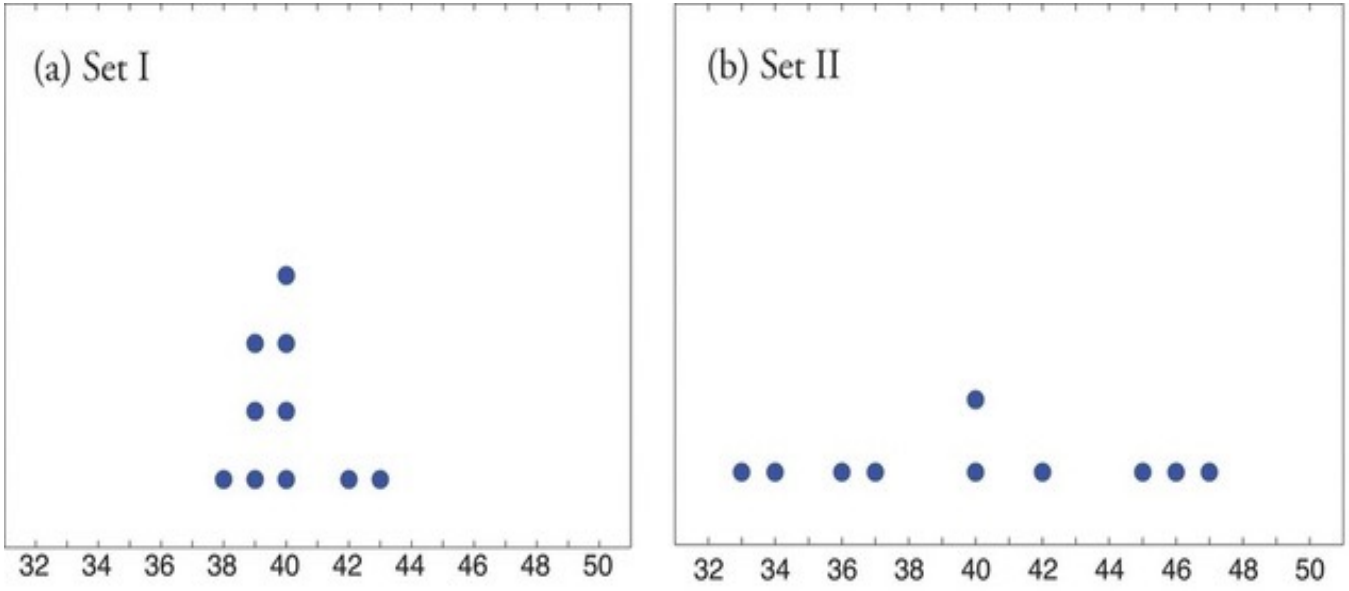
- To overcome these problems, we are going to use measure of dispersion or measure of variability.

Let us look at the two data sets and the graphical representation of each, called a dot plot

Data Set I:	**40**	**38**	**42**	**40**	**39**	**39**	**43**	**40**	**39**	**40**
Data Set II	46	37	40	33	42	36	40	47	34	45

Introduction to Statistics

The two sets of ten measurements each center at the same value: they both have mean, median, and mode 40. Nevertheless a glance at the figure shows that they are markedly different. In Data Set I the measurements vary only slightly from the center, while for Data Set II the measurements vary greatly. Just as we have attached numbers to a data set to locate its center, we now wish to associate to each data set numbers that measure quantitatively how the data either scatter away from the center or cluster close to it. These new quantities are called measures of variability, and we will discuss three of them.



Introduction to Statistics

What is the measure of variability?
Measure of variability also known as the spread of the data describes how similar or varied are the set of observations.

Let's discuss each measure in detail

Range: the range describes the difference between the largest and the smallest points in the data. The bigger the range the more spread out is the data.

$$\text{Range} = x_{max} - x_{min}$$

The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability among the data, whereas a larger range indicates the opposite.

Introduction to Statistics

Variance: It is the average squared deviation from the mean. The variance is computed by finding the difference between every data point and the mean, squaring them, summing them up and then taking the average of those numbers.

Note: The problem with variance is that because if the squaring. It is not the same unit of measurement as the original data. Below are the formula of variance

$$\text{Population Variance } (V_p) = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$\text{Sample Variance } (V_s) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Introduction to Statistics

To overcome the problem of variance, we square root the variance which is called **Standard deviation**.

Standard Deviation: Standard deviation is used more often because it is in the original unit. It is simply the square root of the variance and because of that, it is returned to the original unit of measurement.

When you have a low standard deviation, your data points tend to be close to the mean. A high standard deviation means that your data points are spread out over a wide range.

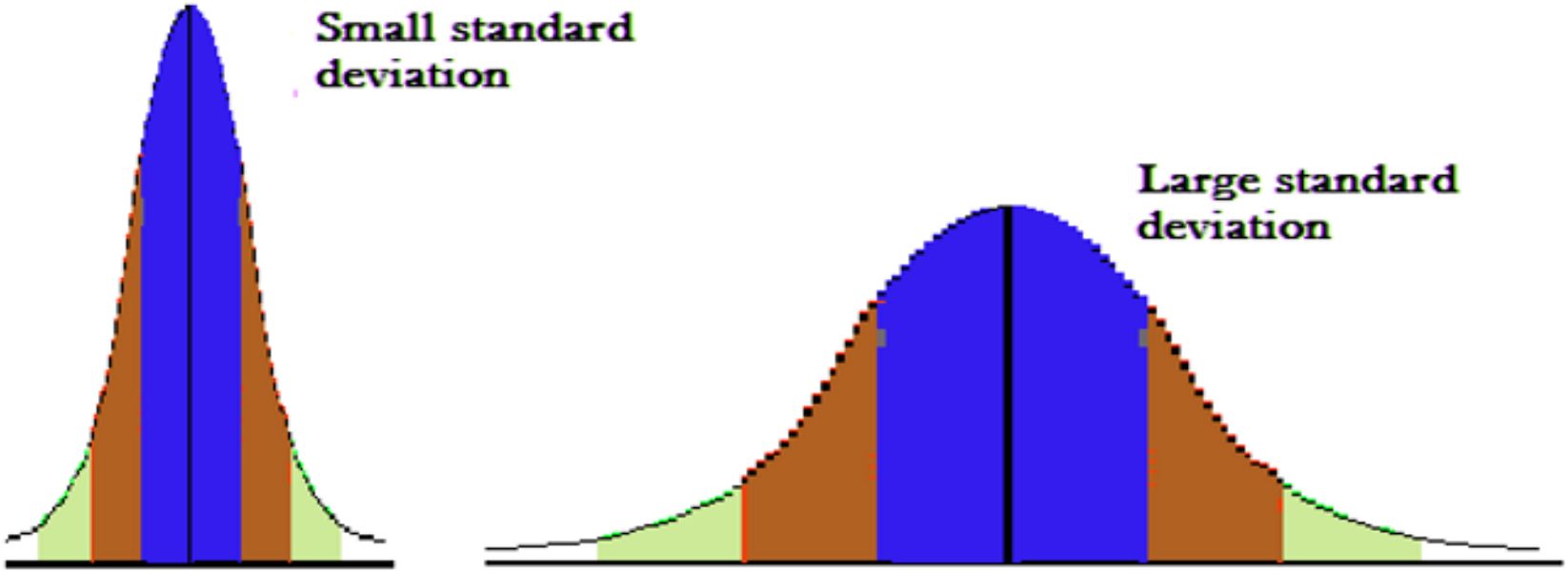
In a normal distribution, 68% of the data points fall between one standard deviation above and one standard deviation below the mean.

Approximately, 95% of the data fall between two standard deviation below the mean and two standard deviation above the mean.

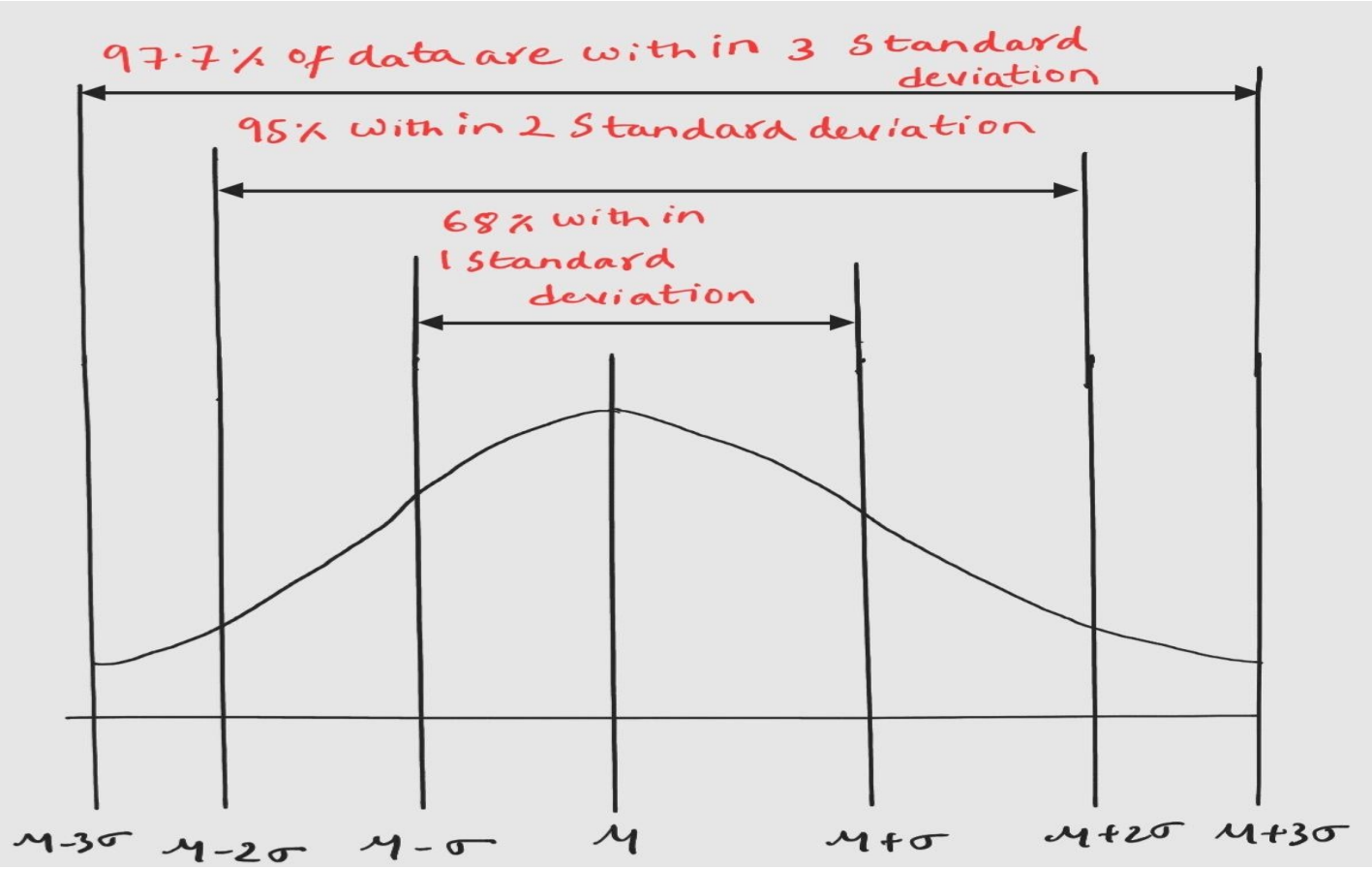
And Approximately 99.7% fall between three standard deviations above and three standard deviations below the mean.

Introduction to Statistics

The smaller the deviation, the narrower the peak, the data points are closer to the mean, The further the data points are from the mean, the greater the standard deviation.



Introduction to Statistics



Introduction to Statistics

Formula for calculating the standard deviation:

$$\sigma_P = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

$$\sigma_S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

n = The number of data points

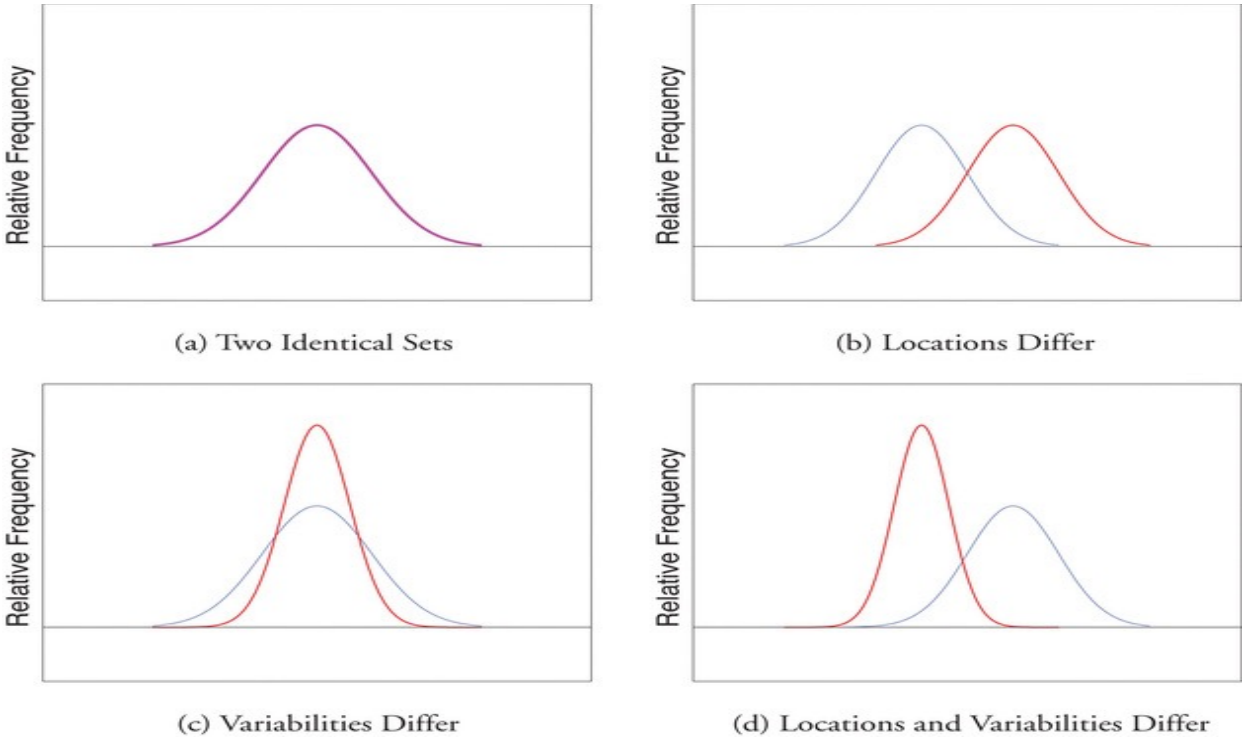
x_i = Each of the values of the data

$\bar{x}$ = The mean of x_i

Note that the denominator in the fraction is the full number of observations, not that number reduced by one, as is the case with the sample standard deviation. Since most data sets are samples, we will always work with the sample standard deviation and variance.

Introduction to Statistics

Finally, in many real-life situations the most important statistical issues have to do with comparing the means and standard deviations of two data sets. Figure illustrates how a difference in one or both of the sample mean and the sample standard deviation are reflected in the appearance of the data set as shown by the curves derived from the relative frequency histograms built using the data.



Relative Position of Data

When you take an exam, what is often as important as your actual score on the exam is the way your score compares to other students' performance. If you made a 70 but the average score (whether the mean, median, or mode) was 85, you did relatively poorly. If you made a 70 but the average score was only 55 then you did relatively well. In general, the significance of one observed value in a data set strongly depends on how that value compares to the other observed values in a data set. Therefore, we wish to attach to each observed value a number that measures its relative position.

Percentiles and Quartiles

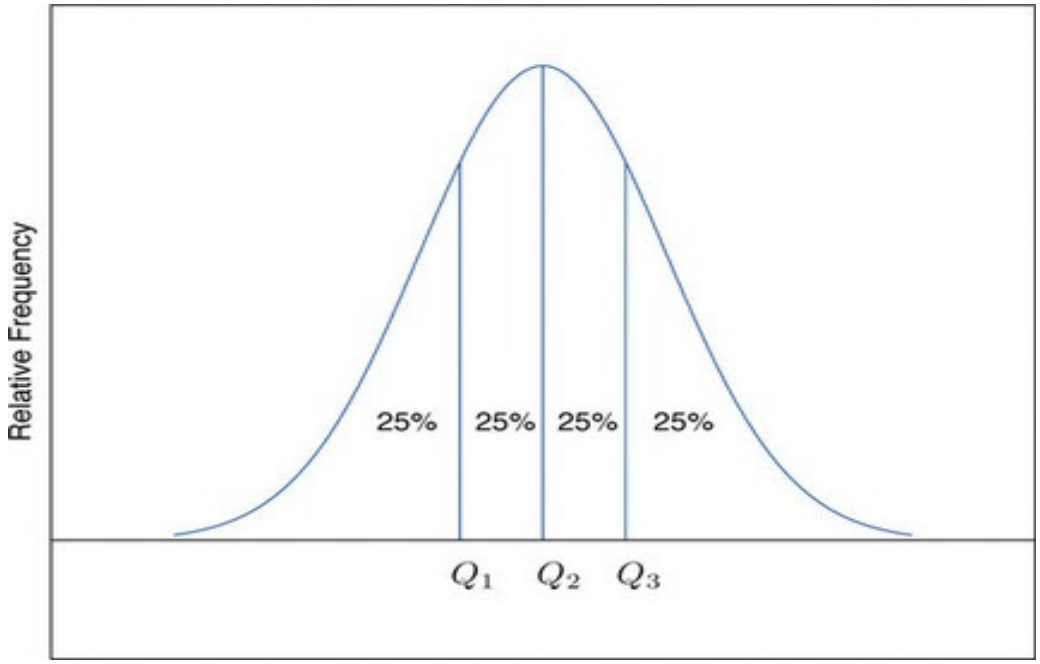
Anyone who has taken a national standardized test is familiar with the idea of being given both a score on the exam and a "percentile ranking" of that score. You may be told that your score was 625 and that it is the 85^{th} percentile. The first number tells how you actually did on the exam; the second says that 85% of the scores on the exam were less than or equal to your score, 625.

Percentile of data

Given an observed value x in a data set, x is the Pth percentile of the data if P% of the data are less than or equal to x. The number P is the percentile rank of x.

Relative Position of Data

The Pth percentile cuts the data set in two so that approximately P%of the data lie below it and (100−P)% of the data lie above it. In particular, the three percentiles that cut the data into fourths, as shown in Figure, are called the quartiles of a data set. The quartiles are the three numbers $Q1$, $Q2$, $Q3$ that divide the data set approximately into fourths



Introduction to Statistics

Quartile:
For any data set:

The second quartile $Q2$ of the data set is its median.

Define two subsets:
the lower set: all observations that are strictly less than $Q2$
the upper set: all observations that are strictly greater than $Q2$

The first quartile $Q1$ of the data set is the median of the lower set.

The third quartile $Q3$ of the data set is the median of the upper set.

Introduction to Statistics

Example:
Find the quartiles of the following data set

1.39 1.76 1.90 2.12 2.53 2.71 3.00 3.33 3.71 4.00

Solution:
This data set has n=10 observations. Since 10 is an even number, the median is the mean of the two middle observations:
$\tilde{x}$ =(2.53+2.71)/2=2.62.
Thus the second quartile is $Q2$=2.62. The lower and upper subsets are

Lower: L={1.39,1.76,1.90,2.12,2.53},

Upper: U={2.71,3.00,3.33,3.71,4.00}.

Each has an odd number of elements, so the median of each is its middle observation. Thus the first quartile is $Q1$=1.90, the median of L,

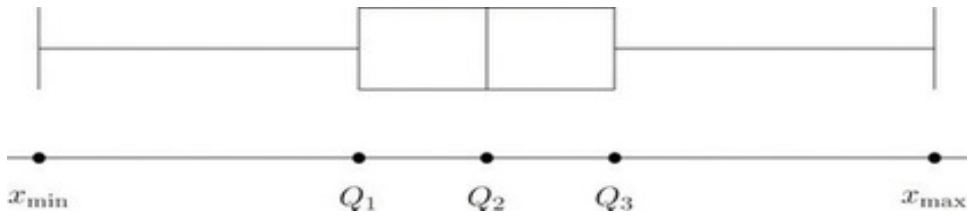
and the third quartile is $Q3$=3.33, the median of U.

Introduction to Statistics

In addition to the three quartiles, the two extreme values, the minimum x_{min} and the maximum x_{max} are also useful in describing the entire data set. Together these five numbers are called the five-number summary of a data set,

$$\{ Xmin, \mathrm{X}_{\mathrm{min}}, Q1, Q2, Q3\}$$

The five-number summary is used to construct a box plot. Each of the five numbers is represented by a vertical line segment, a box is formed using the line segments at $Q1$ and $Q3$ as its two vertical sides, and two horizontal line segments are extended from the vertical segments marking $Q1$ and $Q3$ to the adjacent extreme values. (The two horizontal line segments are referred to as “whiskers,” and the diagram is sometimes called a "box and whisker plot.") We caution the reader that there are other types of box plots that differ somewhat from the ones we are constructing, although all are based on the three quartiles .



Introduction to Statistics

Note that the distance from $Q1$ to $Q3$ is the length of the interval over which the middle half of the data range. Thus it has the following special name: Interquartile range

IQR = Q3-Q1

Z-Score

Another way to locate a particular observation x in a data set is to compute its distance from the mean in units of standard deviation known as z-score

The z-score indicates how many standard deviations an individual observation x is from the center of the data set, its mean. It is used on distributions that have been *standardized*, which allows us to better understand its properties. If z is negative then xx is below average.

If z is 0 then x is equal to the average. If z is positive then x is above the average

The z-score of an observation x is the number z given by

$$z = \frac{x - \mu}{\sigma}$$

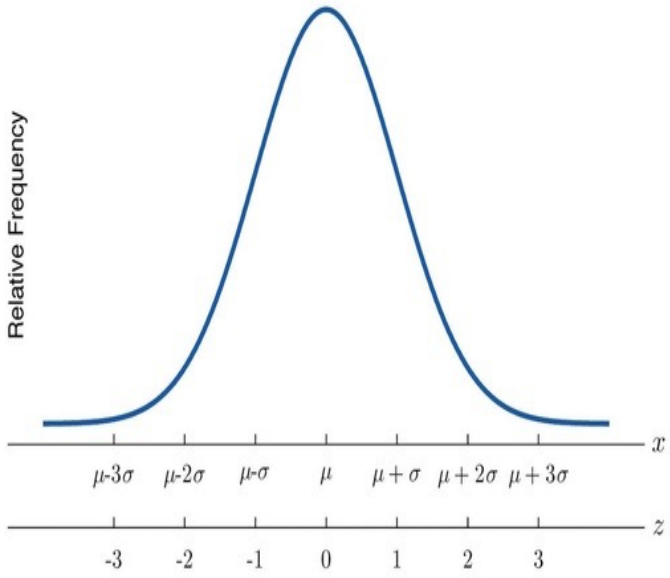
Z-Score

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the *Empirical Rule* says the following:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean $\mu\mu$ (within three standard deviations of the mean). Notice that almost all the xx values lie within three standard deviations of the mean.
- The z-scores for $+1\sigma$ and -1σ are +1 and –1, respectively.
- The z-scores for $+2\sigma$ and -2σ are +2 and –2, respectively.
- The z-scores for $+3\sigma$ and -3σ are +3 and –3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.



x-Scale versus *z*-Score

Z-Score

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0,1)$. The mean of the z-scores is zero and the standard deviation is one. If y is the z-score for a value x from the normal distribution $N(\mu,\sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Inferential Statistics

Why do we need Inferential Statistics?
In contrast to Descriptive Statistics, rather than having access to the whole population, we often have a limited number of data.

In such cases, Inferential Statistics come into action. For example, we might be interested in finding the average of the entire school's exam marks. It is not reasonable because we might find it impracticable to get the data we need. So, rather than getting the entire school's exam marks, we measure a smaller sample of students (for example, a sample of 50 students). This sample of 50 students will now describe the complete population of all students of that school.

Simply put, Inferential Statistics make predictions about a population based on a sample of data taken from that population.

The technique of Inferential Statistics involves the following steps:
First, take some samples and try to find one that represents the entire population accurately.
Next, test the sample and use it to draw generalizations about the whole population.

Inferential Statistics

There are two main objectives of inferential statistics:

Estimating parameters: We take a statistic from the collected data, such as the standard deviation, and use it to define a more general parameter, such as the standard deviation of the complete population.

Hypothesis testing: Very beneficial when we are looking to gather data on something that can only be given to a very confined population, such as a new drug. If we want to know whether this drug will work for all patients ("complete population"), we can use the data collected to predict this (often by calculating a z-score).

Z-Score

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0,1)$. The mean of the z-scores is zero and the standard deviation is one. If y is the z-score for a value x from the normal distribution $N(\mu,\sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Counting the Number of Combinations

You want to calculate the number of combinations of n items taken r at a time

C(n,r) = n!/(n-r)!r!

Python has built in function

```
from scipy.special import comb
comb(5,3)
```

Out: 10

Counting the Number of Combinations

Print the combinations

```
from itertools import combinations
combi = combinations ([1,2,3,4,5],3)
#len(list(combi))

# print the list of combinations
for i in list(combi):
    print(i)
```

Counting the Number of Combinations

You want to calculate the number of permutations of n items taken r at a time

P(n,r) = n!/(n-r)!

In combination order does not matter whereas in permutation combination matters. There is function in python which can give permutations also.

```
from itertools import permutations

Perm_number = permutations([1,2,3])

#print(len(list(Perm_number)))
```

Counting the Number of Combinations

You want to calculate the number of permutations of n items taken r at a time

P(n,r) = n!/(n-r)!

In combination order does not matter whereas in permutation combination matters. There is function in python which can give permutations also.

```
from itertools import permutations

Perm_number = permutations([1,2,3])

#print(len(list(Perm_number)))

for i in list(Perm_number):
    print(i)
```

Counting the Number of Combinations

```
# Defining the number of elements chosen

from itertools import permutations

Perm_number1 = permutations([1,2,3,4, 5],2 )

#print(len(list(Perm_number1)))

for i in list(Perm_number1):
    print(i)
```

Random Variable

A random variable is a numerical quantity that is generated by a random experiment. Usually it is denoted by capital letters, such as X, and the actual values that they can take by lowercase letters, such as x.

Discrete random variable

A random variable is called a discrete if it has either a finite or a countable number of possible values.

A random variable is called continuous if its possible values contain a whole interval of numbers.

Random Variable

The distinction between discrete and continuous random variables is important because of the different mathematical techniques associated with the two kinds of random variables.

In most of the cases we will consider a *discrete random variable.* The number of students in a statistics class is a discrete random variable. Values such as 15, 25, 50, and 250 are all possible. However, 25.5 students is not a possible value for the number of students.

Most of the *continuous random variables* we will see will occur as the result of a measurement on a continuous scale. For example, the air pressure in an automobile tire represents a continuous random variable. The air pressure could, in theory, take on any value from 0-2 (psi) to the bursting pressure of the tire. Values such as 20.126 psi, 20.12678 psi, and so forth are possible.

Associated to each possible value of x of a discrete random variable X is the probability $P(x)$ that X will take the value x in one trial of the experiment.

Probability Distribution

Probability Distributions are mathematical functions that describe all the possible values and likelihoods that a random variable can take within a given range.

Probability distributions help model random phenomena, enabling us to obtain estimates of the probability that a certain event may occur.

Discrete Probability Distribution

Probability distribution

The probability distribution of a discrete random variable X is a list of each possible value of X together with the probability that X takes that value in one trial of the experiment.

The probabilities in the probability distribution of a random variable X must satisfy the following two conditions:

- Each probability $P(x)$ must be between 0 and 1:

$$0 \leq P(x) \leq 1.$$

- The sum of all the possible probabilities is 1:

$$\sum P(x) = 1.$$

Probability Distribution

Example

A fair coin is tossed twice. Let X be the number of heads that are observed.

1. Construct the probability distribution of X.
2. Find the probability that at least one head is observed.

Solution:

The possible values that X can take are 0, 1, and 2. Each of these numbers corresponds to an event in the sample space S={hh,ht,th,tt} of equally likely outcomes for this experiment:

X=0 to {tt}, X=1 to {ht,th},and X=2 to hh

The probability of each of these events, hence of the corresponding value of X, can be found simply by counting, to give

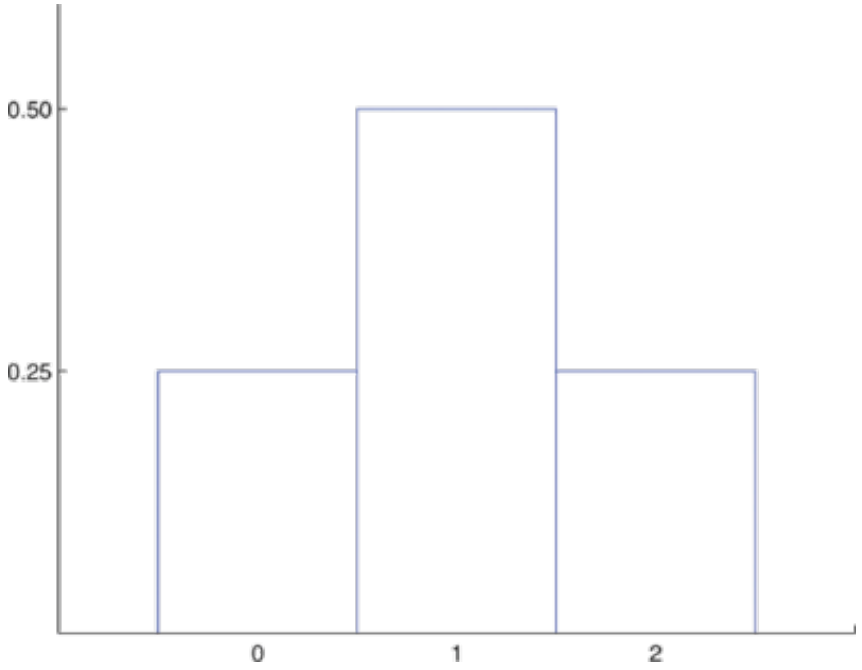
x	0	1	2
$P(x)$	0.25	0.50	0.25

This table is the probability distribution of X.

Probability Distribution

"At least one head" is the event $X \geq 1$, which is the union of the mutually exclusive events $X=1$ and $X=2$. Thus

$$P(X \geq 1) = P(1) + P(2) = 0.50 + 0.25 = 0.75$$



Probability Distribution for Tossing a Fair Coin Twice

Discrete Probability Distribution

Mean and Standard Deviation of a discrete variable

The *mean* (also called the "expectation value" or ” expected value") of a discrete random variable X is the number

$$\mu=E(X)=\sum xP(x)$$

The mean of a random variable may be interpreted as the average of the values assumed by the random variable in repeated trials of the experiment.

Variance:

The *variance* (σ^2) of a discrete random variable X is the number

$$\sigma^2=\sum(x-\mu)^2P(x)$$

which by algebra is equivalent to the formula

$$\sigma^2=[\sum x^2P(x)]-\mu^2$$

Discrete Probability Distribution

Standard Deviation

The standard deviation, σ, *of a discrete random variable* X *is the square root of its variance.*

The variance and standard deviation of a discrete random variable X may be interpreted as measures of the variability of the values assumed by the random variable in repeated trials of the experiment. The units on the standard deviation match those of X.

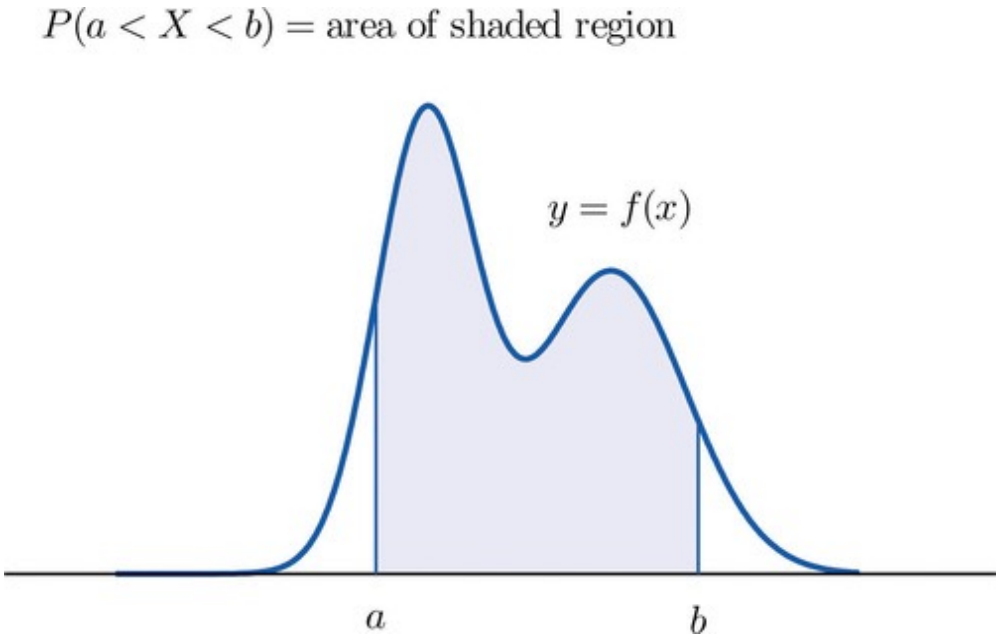
Probability Distribution of a Continuous Variable

For a discrete random variable X the probability that X assumes one of its possible values on a single trial of the experiment makes good sense. This is not the case for a continuous random variable. For example, suppose X denotes the length of time a commuter just arriving at a bus stop has to wait for the next bus. If buses run every 30 minutes without fail, then the set of possible values of X is the interval denoted [0,30], the set of all decimal numbers between 0 and 30. But although the number 7.211916 is a possible value of X, there is little or no meaning to the concept of the probability that the commuter will wait precisely 7.211916 for the next bus. If anything the probability should be zero, since if we could meaningfully measure the waiting time to the nearest millionth of a minute it is practically inconceivable that we would ever get exactly 7.211916 minutes. More meaningful questions are those of the form: What is the probability that the commuter's waiting time is less than 10 minutes, or is between 5 and 10 minutes? In other words, with continuous random variables one is concerned not with the event that the variable assumes a single particular value, but with the event that the random variable assumes a value in a particular interval.

Probability Distribution of a Continuous Variable

Density Function

The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a function $f(x)$f(x), called a density function, in the following way: the probability that X assumes a value in the interval $[a,b]$ is equal to the area of the region that is bounded above by the graph of the equation $y=f(x)$, bounded below by the x-axis, and bounded on the left and right by the vertical lines through a and b



Probability Distribution of a Continuous Variable

Every density function $f(x)$ must satisfy the following two conditions:

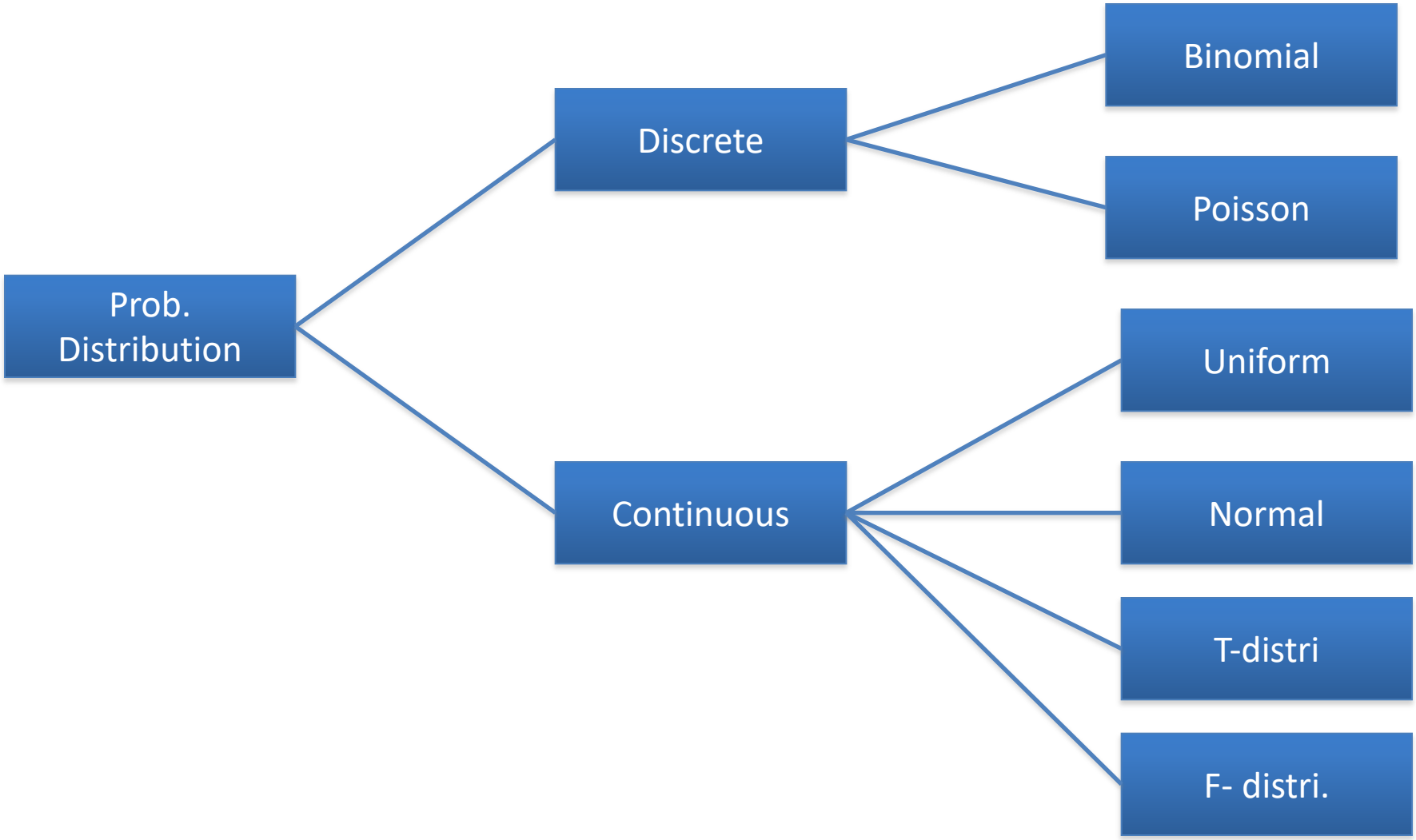
- For all numbers x, $f(x)\geq 0$, so that the graph of $y=f(x)$ never drops below the x-axis.
- The area of the region under the graph of $y=f(x)$ and above the x-axis is 1.

Because the area of a line segment is 0, the definition of the probability distribution of a continuous random variable implies that for any particular decimal number, say a, the probability that X assumes the exact value a is 0. This property implies that whether or not the endpoints of an interval are included makes no difference concerning the probability of the interval.

For any continuous random variable X:

$$P(a\leq X\leq b)=P(a<X\leq b)=P(a\leq X<b)=P(a<X<b)$$

Common Probability Distribution



Common Probability Distribution

Uniform Distribution – where a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n. Another way of saying it " a known, finite number o outcomes equally likely to happen".

Normal Distribution - A theoretical frequency distribution for a set of variable data, usuall represented by a bell-shaped curve symmetrical about the mean. Also called *Gaussian distributio*

Bernoulli Trials - a random experiment with exactly two possible outcomes, "success" an "failure", in which the probability of success is the same every time the experiment is conducted

Binomial Distribution - a frequency distribution of the **possible number of successful outcome in a given number of trials** in each of which there is the same probability of success

All of the common distributions are available in Python, and for every distribution python can calculate various properties.

Probability Distribution

Distribution in python with their function names and function requirements (generally values of distribution parameters)

Distribution	Specifications
Normal	Mean μ and standard deviation σ
Uniform	Min and Max of the distribution
Binomial	Number of trials and probability of success on each trial

Common Probability Distribution

The normal distribution is the most important of all probability distributions. It is applied directly to many practical problems such as: the income distribution in the economy, students average reports, the average height in populations, etc. and several very useful distributions are based on it.

A **Normal Distribution** is also known as a **Gaussian distribution** or famously **Bell Curve**. People use both words interchangeably, but it means the same thing. It is a continuous probability distribution.

Many empirical frequency distributions have the following characteristics:

- They are approximately symmetrical, and the mode is close to the center of the distribution
- The mean, median and mode are close together.
- The shape of the distribution can be approximated by bell: nearly flat on the top, then decreasing more quickly, then decreasing more slowly toward the tails of the distribution.

Normal Probability Distribution

The probability density function (pdf) for Normal Distribution:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi * \sigma^2}} * e^{-\frac{1}{2} * (\frac{x-\mu}{\sigma})^2}$$

where, μ = Mean , σ = Standard deviation , x = input value.

If a random variable x follows the normal distribution, then we write:

$x \sim N(\mu,\sigma^2)$

Normal Probability Distribution

- The normal distribution is a continuous probability distribution.

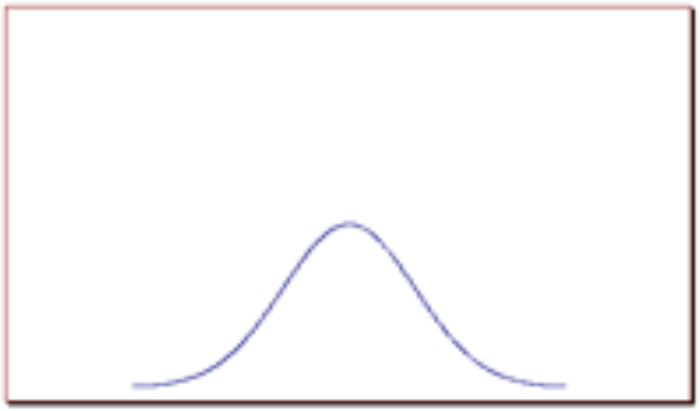
- The total area under the normal curve is equal to 1.

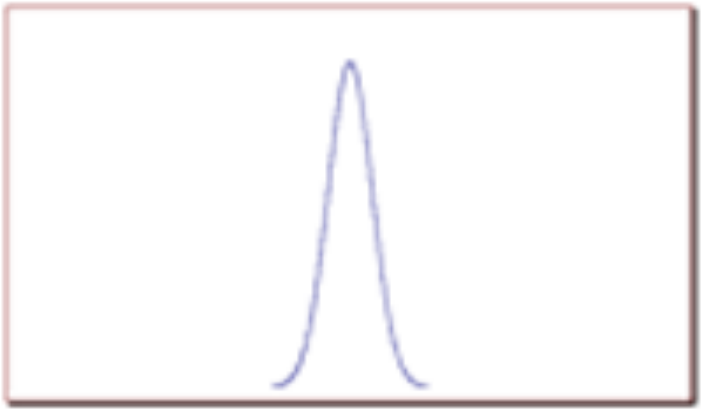
- Because the normal density function is symmetrical, the mean, median and mode coincide at $x = \mu$, thus the value of μ determines the location of the center of the distribution and the value of σ determines its spread.

Properties of Normal Probability Distribution

The graph of the normal distribution depends on two factors – the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. We can alter the shape of the bell curve by changing the mean and standard deviation. Changing the mean will shift the curve towards that mean value, this means we can change the position of the curve by altering the mean value while the shape of the curve remains intact. The shape of the curve can be controlled by the value of Standard deviation. A smaller standard deviation will result in a closely bounded curve while a high value will result in a more spread out curve.

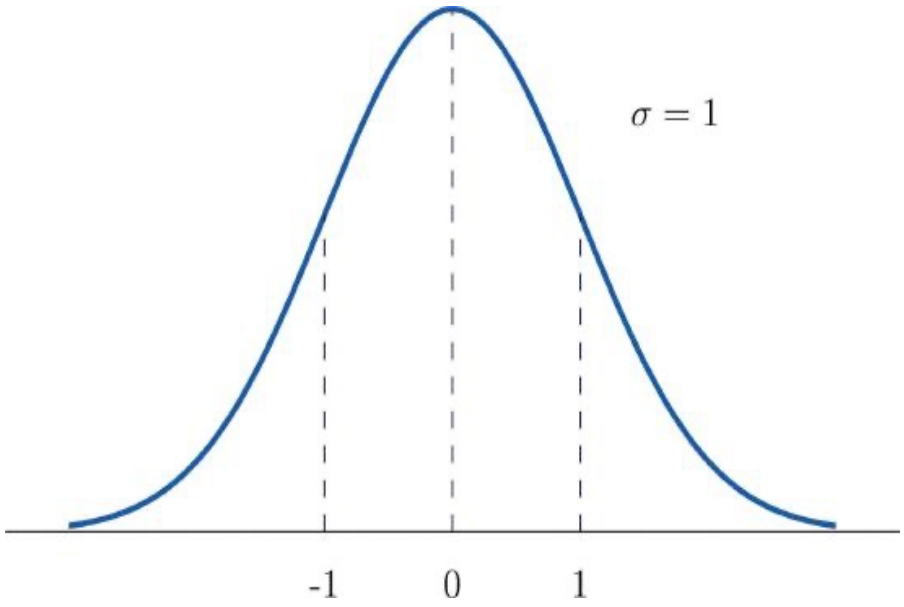
When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow.





Standard Normal Probability Distribution

A *standard normal random variable* is a normally distributed random variable with mean μ=0 and standard deviation σ=1. The density function for a standard normal random variable is shown below

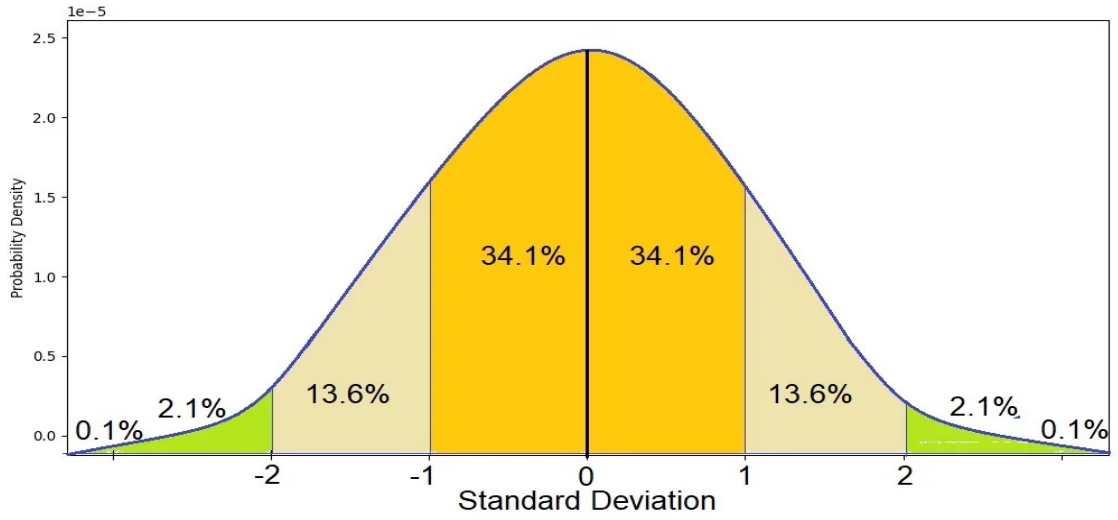


Properties of Normal Probability Distribution

- The mean, mode, and median are all equal.
- The total area under the curve is equal to 1.
- The curve is symmetric around the mean.

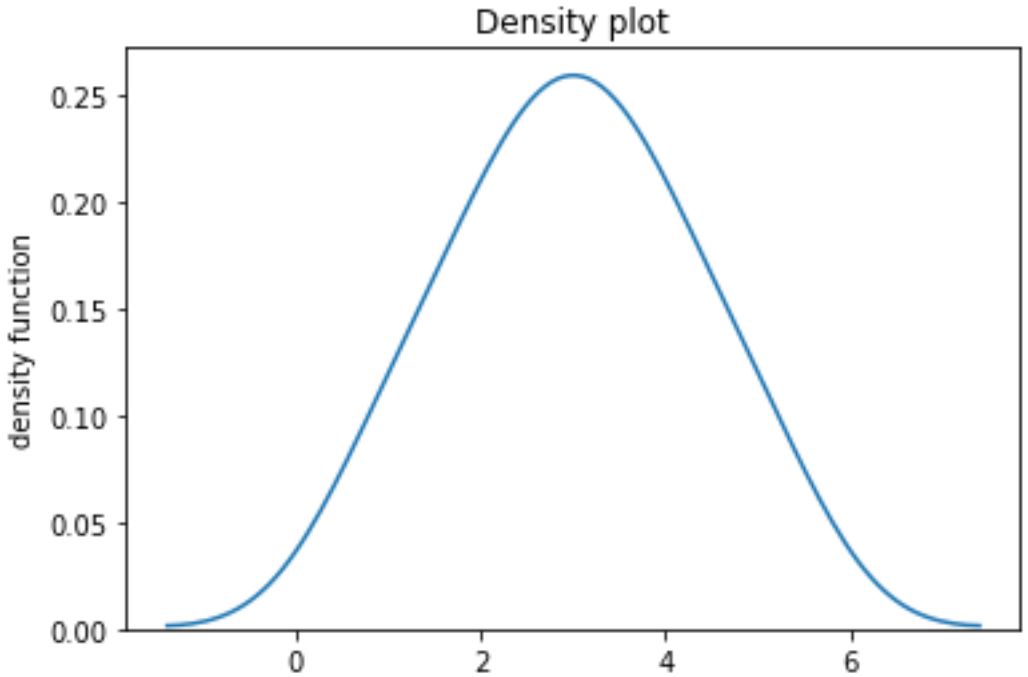
Empirical rule tells us that:

- 68% of the data falls within one standard deviation of the mean.
- 95% of the data falls within two standard deviations of the mean.
- 99.7% of the data falls within three standard deviations of the mean.



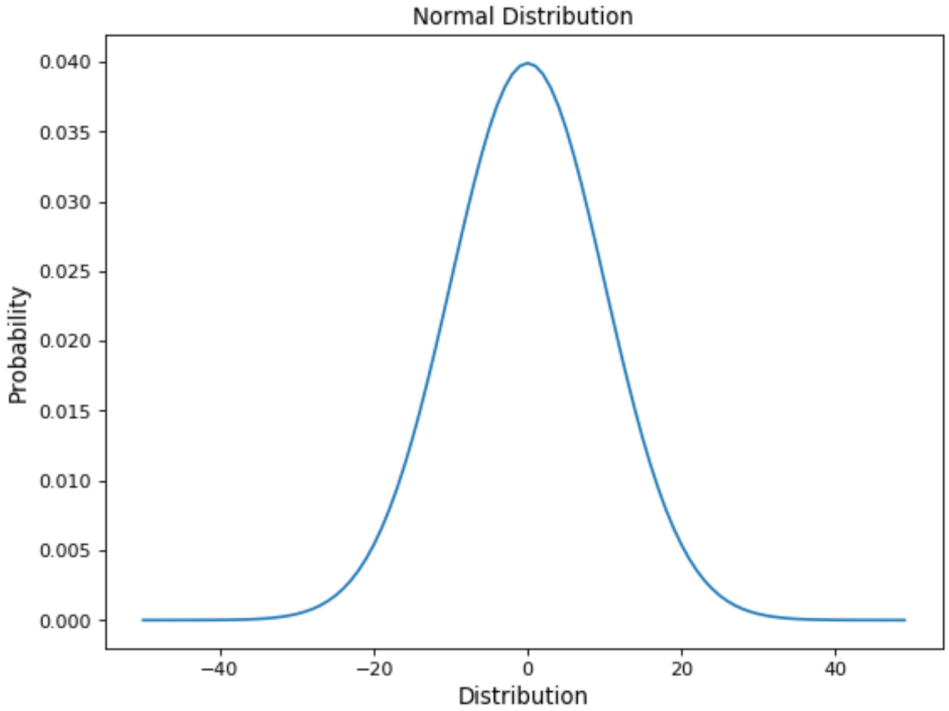
Normal Distribution

```
x = pd.Series([1, 2, 2.5, 3, 3.5, 4, 5])
# Plot Normal Distribution
Import seaborn as sns
sns.distplot(x, hist = False, kde=True)
plt.title('Density plot')
plt.xlabel('x')
plt.ylabel('density function')
```



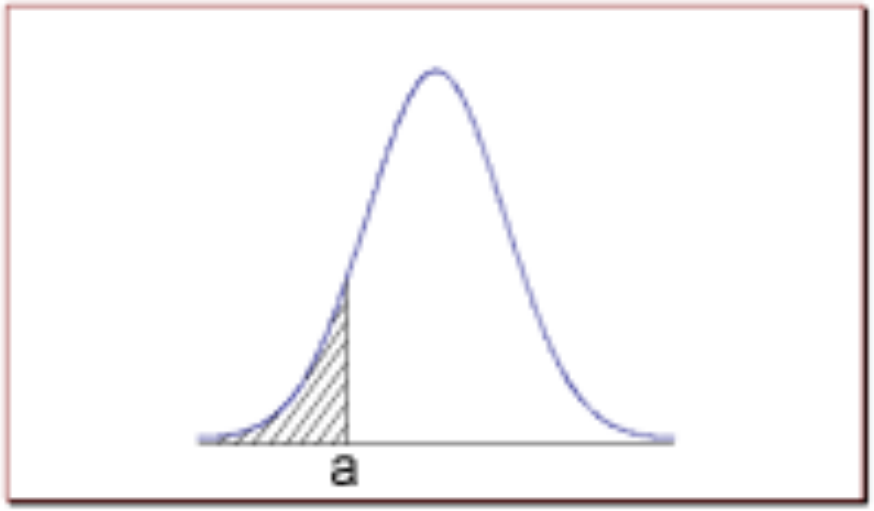
Normal Distribution

```
x = np.arrange(-50,50)
mean = 0
normal = stats.norm.pdf(n,mean,10)
plt.plot(x,normal)
plt.title(‘Normal Distribution’)
plt.xlabel(‘Distribution')
plt.xlabel(‘Probability')
```



Normal Distribution

The probability that x is greater than a equals the area under the normal curve bounded by a and plus infinity (as indicated by the non-shaded area in the figure below).

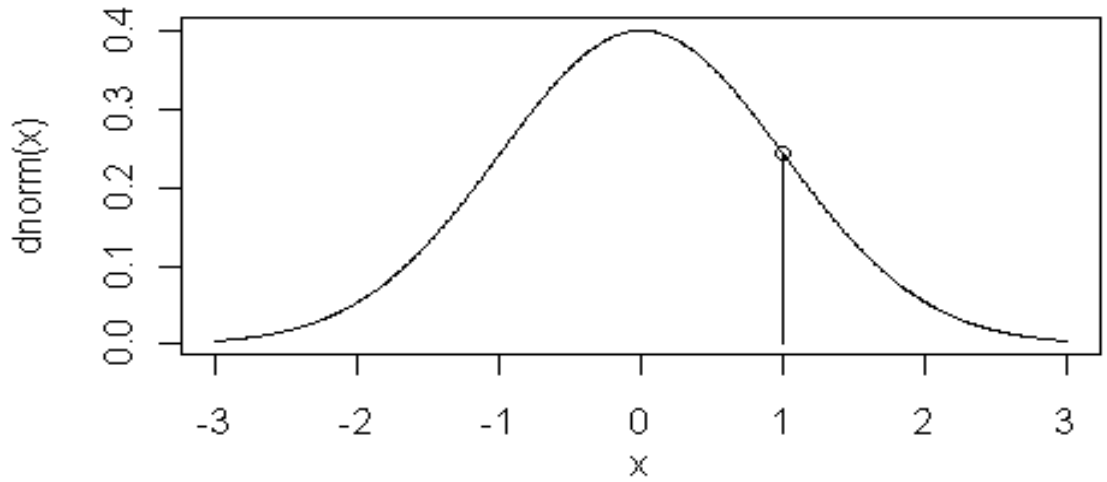


Normal Distribution

The pdf function returns the height of the normal curve at some value along the x-axis. This is illustrated in the figure at left. Here the value of norm.pdf(1) is shown by the vertical line at x=1.

```
from scipy.stats import norm
norm.pdf(1)
0.2419707
```

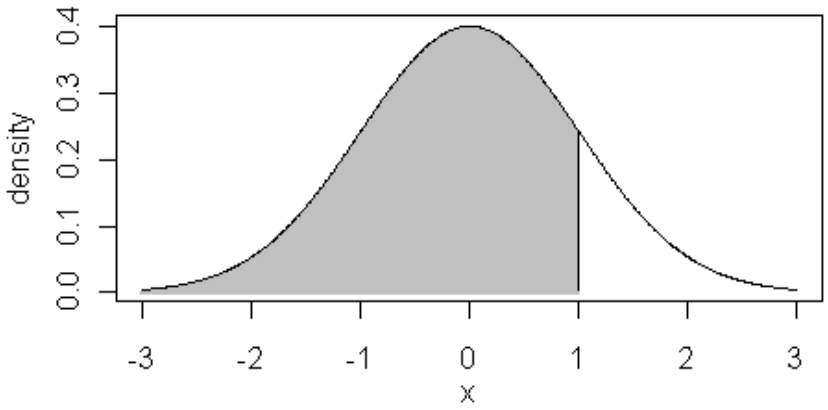
With no options specified, the value of "x" is treated as a standard score or z-score. To change this, you can specify ”loc (equivalent to mean)=" and ”scale (equivalent sd)=" options. In other words, returns the probability density function or pdf.



Normal Distribution

The cdf() function is the cumulative density function or cdf. It returns the area below the given value of "x", or for x=1, the shaded region in the figure at right.

> norm.cdf(1)
[1] 0.8413447
Once again, the defaults for mean and sd are 0 and 1 respectively. These can be set to other values as in the case of pdf().



To find the area above the cutoff x-value, cdf subtract from 1,

> 1 – norm.cdf(1)
[1] 0.1586553

Normal Distribution

Problem

Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

```
1-norm.cdf(84, mean=72, sd=15.2)
[1] 0.21492
```

cdf gives cumulative probability function.

Answer

The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

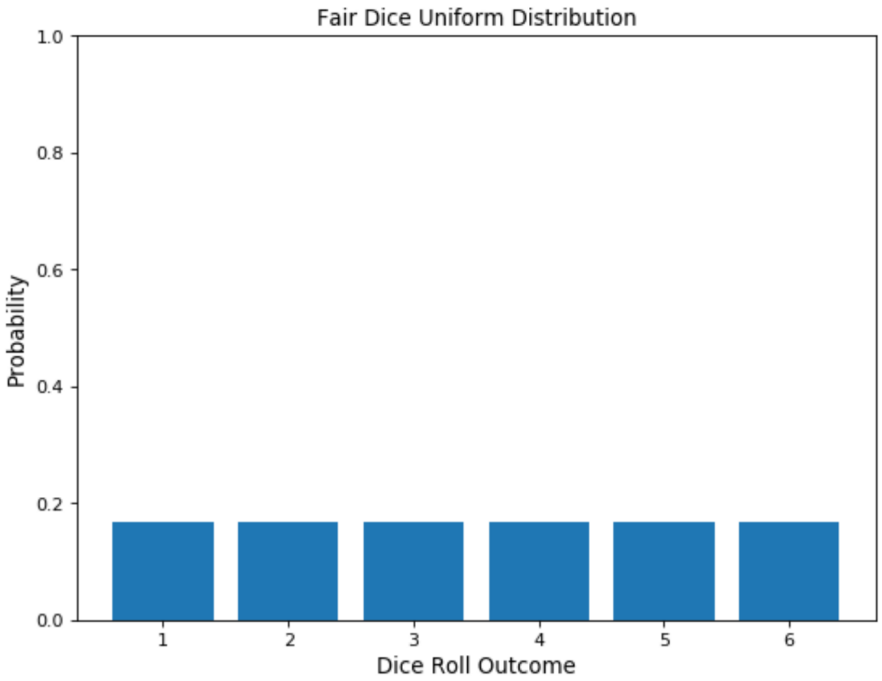
Uniform Distribution

The Uniform Distribution can be easily derived from the Bernoulli Distribution. In this case, a possibly unlimited number of outcomes are allowed and all the events hold the same probability to take place.

As an example, imagine the roll of a fair dice. In this case, there are multiple possible events with each of them having the same probability to happen.

Uniform Distribution

```
probs = np.full((6,1/6)
face =[1,2,3,4,5,6]
plt.bar(face, probs)
plt.ylabel('Probability')
plt.xlabel('Dice Roll Outcome')
plt.title('Fair Dice Uniform Distribution')
axes = plt.gca()
axes.set_ylim([0,1])
```



Binomial Distribution

- The Binomial Distribution can instead be thought as the sum of outcomes of an event following a Bernoulli distribution. The Binomial Distribution is therefore used in binary outcome events and the probability of success and failure is the same in all the successive trials. This distribution takes two parameters as inputs: the number of times an event takes place and the probability assigned to one of the two classes.

- This important distribution applies in some cases to repeated trials where there are only two possible outcomes: heads or tails, success or failure, defective item or good item, or many other possible pairs.

- The requirements for using the binomial distribution are as follows:
  - The outcome is determined completely by chance
  - There are only two possible outcomes
  - All trials have the same probability for a particular outcome in a single trial. That is, the probability in a subsequent trial is independent of the outcome of a previous trial
  - The number of trials must be fixed, regardless of the outcome of each trial.

Binomial Distribution

To understand binomial distributions and binomial probability, it helps to understand binomial experiment.

- A binomial experiment (also known as a Bernoulli trial) is a statistical experiment that has the following properties:
  - The experiment consists of n repeated trials.
  - Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
  - The probability of success, denoted by P, is the same on every trial.
  - The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials

Binomial Distribution

- Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:
  - The experiment consists of repeated trials. We flip a coin 2 times.
  - Each trial can result in just two possible outcomes - heads or tails.
  - The probability of success is constant - 0.5 on every trial.
  - The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Binomial Distribution

- The binomial distribution has the following properties:
  - The mean of the distribution (μx) is equal to n * P .
  - The variance (σ^2_x) is n * P * (1 - P).
  - The standard deviation (σ_x) is sqrt[n * P * (1 - P)].

Binomial Distribution

The binomial distribution is a discrete probability distribution. It describes the outcome of n independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure. If the probability of a successful trial is p, then the probability of having k successful outcomes in an experiment of n independent trials is as follows.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Example

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Solution

Since only one out of five possible answers is correct, the probability of answering a question correctly by random is 1/5 = 0.2. we can find the probability of having exactly 4 correct answers by random attempts as follows:

```
from scipy.stats import binom
binom.pmf(4,12,0.2)
0.132875
```

Example

Cumulative Binomial Probability
A cumulative binomial probability refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit

To find the probability of having four or less correct answers by random attempts, we apply the function dbinom with x = 0,…,4.

binom.pmf(0,12,0.2)+binom.pmf(1,12,0.2)+binom.pmf(2,12,0.2)+binom.pmf(3,12,0.2)+binom.pmf(4,12,0.2)

0.9274

Example

Alternatively, we can use the cumulative probability function for binomial distribution .

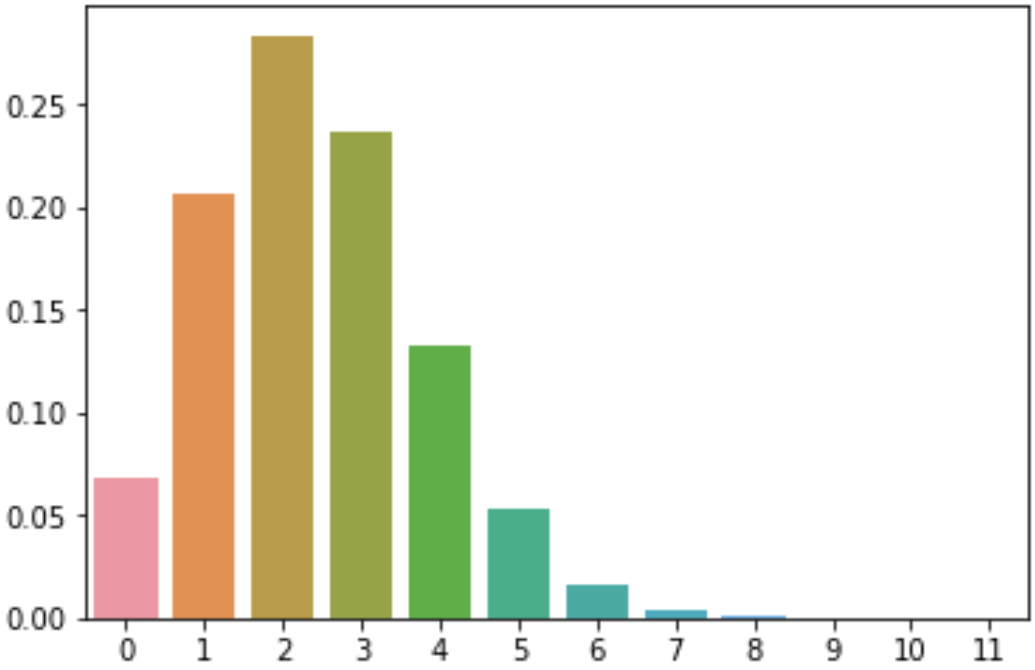
binom.cdf(4,12,0.2)

0.92744

Example

The distribution of all possible results on the test

```
x = np.arange(0,12)
y = binom.pmf(x,12,0.2)
sns.barplot(x,y)
```



Confidence Interval

Confidence Interval

Confidence Interval is used to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.
- A margin of error.

The confidence level describes the uncertainty of a sampling method (it's the probability part of the confidence interval). The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the ***sample statistic + margin of error*.**

Confidence Interval

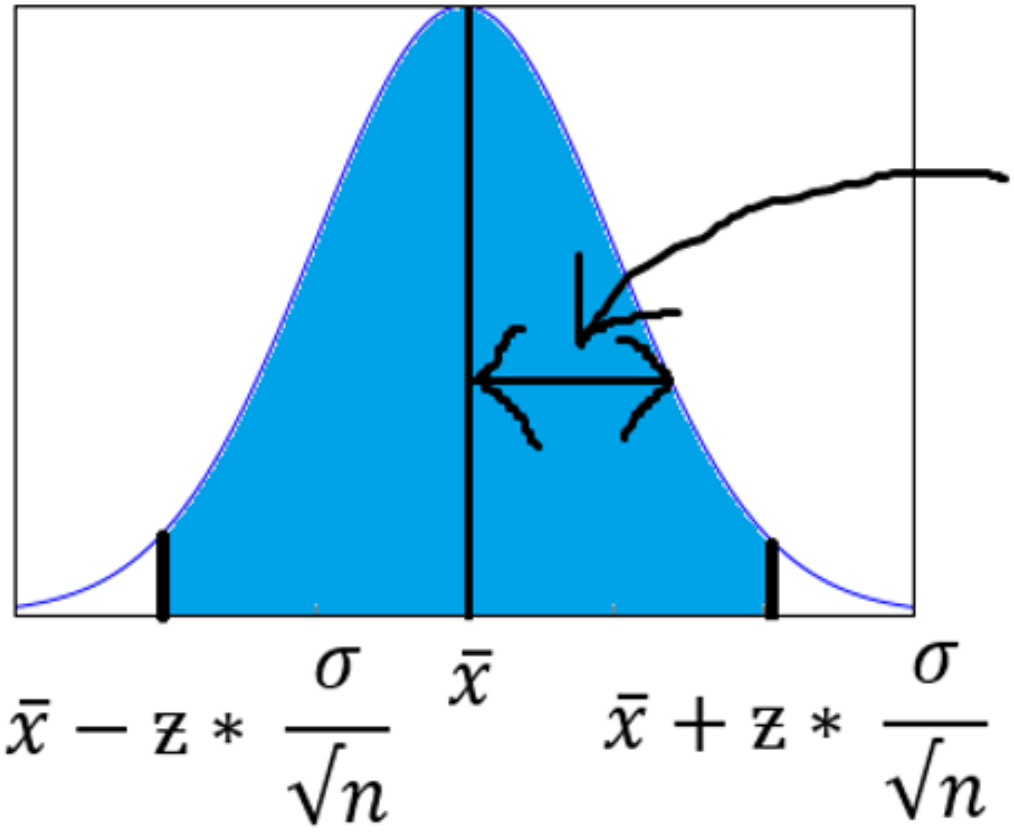
The margin of error is found by multiplying the standard error of the mean and the z-score.

$$\text{Margin of error} = Z \times \frac{\sigma}{\sqrt{n}}$$

And the Confidence interval is defined as:

$$\text{Confidence Interval} = \bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

Confidence Interval



Margin of error (half the width of confidence interval)

A confidence interval having a value of 90% indicates that we are 90% sure that the actual mean is within our confidence interval.

Hypothesis Testing

Hypothesis testing is a part of statistics in which we make assumptions about the population parameter. So, hypothesis testing mentions a proper procedure by analyzing a random sample of the population to accept or reject the assumption.

Hypothesis testing is the way of trying to make sense of assumptions by looking at the sample data.

Type of Hypothesis

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

Hypothesis Testing

There are two types of statistical hypotheses.

Null hypothesis. The null hypothesis, denoted by **Ho**, is usually the hypothesis that sample observations result purely from chance.

Alternative hypothesis. The alternative hypothesis, denoted by **H1 or Ha**, is the hypothesis that sample observations are influenced by some non-random cause.

Hypothesis Testing

Steps of Hypothesis Testing

The process to determine whether to reject a null hypothesis or to fail to reject the null hypothesis, based on sample data. This process, called **hypothesis testing**, consists of four steps.

- **State the hypotheses.** This involves stating the null and alternative hypotheses, and both should be mutually exclusive. That is, if one is true, the other must be false.
- **Formulate an analysis plan.** It describes how to use sample data to evaluate the null hypothesis. This evaluation often focuses on a single test statistic. Here we choose the significance level (α) among 0.01, 0.05, or 0.10 and also determine the test method.
- **Analyze sample data.** Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) and p-value described in the analysis plan.
- **Interpret results.** Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

Hypothesis Testing

Errors in hypothesis testing

We have explained what is hypothesis testing and the steps to do the testing. Now, while performing the hypothesis testing, there might be some errors.

Type I error. A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha** and is often denoted by **α**.

Type II error. A Type II error occurs when the researcher fails to reject a false null hypothesis. The probability of committing a Type II error is called **beta** and is often denoted by **β**. The probability of *not* committing a Type II error is called the **Power** of the test.

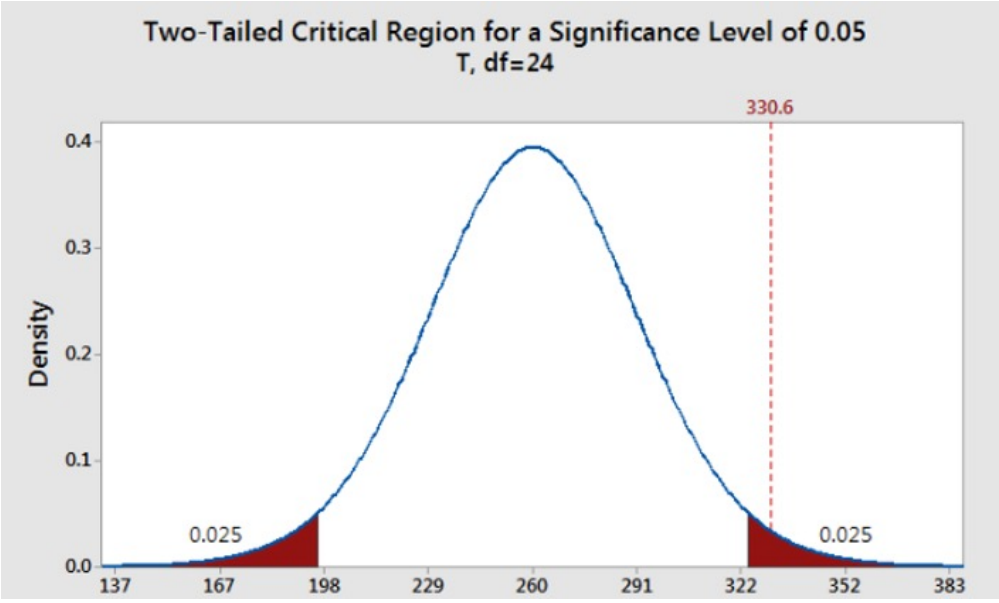
Hypothesis Testing

Terms in Hypothesis testing

Significance level

The significance level is defined as the probability of the case when we reject the null hypothesis, but in actuality, it is true. For example, a 0.05 significance level indicates that there is a 5% risk in assuming that there is some difference when, in actuality, there is no difference. It is denoted by alpha (**α**).

The figure shows that the two shaded regions are equidistant from the null hypothesis, each having a probability of 0.025 and a total of 0.05, which is our significance level. The shaded region in case of a two-tailed test is called the critical region.



Hypothesis Testing

T-test

T-tests are very much similar to the z-scores, the only difference being that instead of the Population Standard Deviation, we now use the Sample Standard Deviation. The rest is the same as before, calculating probabilities on the basis of t-values.

The Sample Standard Deviation is given as:

$$s = \frac{\sqrt{\sum(x-\bar{x})^2}}{(n-1)}$$

where *n*-1 is Bessel's correction for estimating the population parameter.

Another difference between z-scores and t-values is that t-values are dependent on the Degree of Freedom of a sample.

Hypothesis Testing

The Degree of Freedom — It is the number of variables that have the choice of having more than one arbitrary value. For example, in a sample of size 10 with a mean of 10, 9 values can be arbitrary, but the 10th value is forced by the sample mean.

Points to note about the t-tests:
The greater the difference between the sample mean and the population mean, the greater the chance of rejecting the Null Hypothesis.
Greater the sample size, the greater the chance of rejection of the Null Hypothesis.

Hypothesis Testing

Different types of T-test

One Sample T-test

The one-sample t-test compares the mean of sample data to a known value. So, if we have to compare the mean of sample data to the population mean, we use the One-Sample T-test.

We can run a one-sample T-test when we do not have the population S.D., or we have a sample of size less than 30.

t-statistic is given by:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

where, *X(bar)* is the sample mean, μ the population mean, *s* the sample standard deviation, and *N* the sample size.

Hypothesis Testing

Two sample T-test

We use a two-sample T-test when we want to evaluate whether the mean of the two samples is different or not. In a two-sample T-test, we have another two categories:
Independent Sample T-test: Independent sample means that the two different samples should be selected from two completely different populations. In other words, we can say that one population should not be dependent on the other population.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

$\bar{x}_1 - \bar{x}_2$ is the difference between the sample means

$\mu_1 - \mu_2$ is the difference between the hypothesized population means

$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is the standard error of the difference between the sample means

Hypothesis Testing

Paired T-test: If our samples are connected in some way, we have to use the paired t-test. Here, 'connecting' means that the samples are connected as we are collecting data from the same group two times, e.g., blood tests of patients of a hospital before and after medication.

$$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

where, $\bar{\mathbf{d}}$ is mean of the case wise difference between before and after case,

S_d = standard deviation of the difference

n = sample size.

Hypothesis Testing

Chi-square test

The Chi-square test is used in the case when we have to compare categorical data.

The Chi-square test is of two types. Both use chi-square statistics and distribution for different purposes.

The goodness of fit: It determines if sample data of categorical variables match with population or not.

Test of Independence: It compares two categorical variables to find whether they are related to each other or not.

Chi-square statistic is given by:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

Hypothesis Testing

ANOVA (Analysis of variance)

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means. Now, the question arises — Why do we need another test for checking the difference of means between independent groups? Why can we not use multiple t-tests to check for the difference in means?

The answer is simple. Multiple t-tests will have a compound effect on the error rate of the result. Performing a t-test thrice will give an error rate of ~15%, which is too high, whereas ANOVA keeps it at 5% for a 95% confidence interval.

To perform an ANOVA, you must have a continuous response variable and at least one categorical factor with two or more levels. ANOVA requires data from approximately normally distributed populations with equal variances between factor levels.

There are two types of ANOVA test:

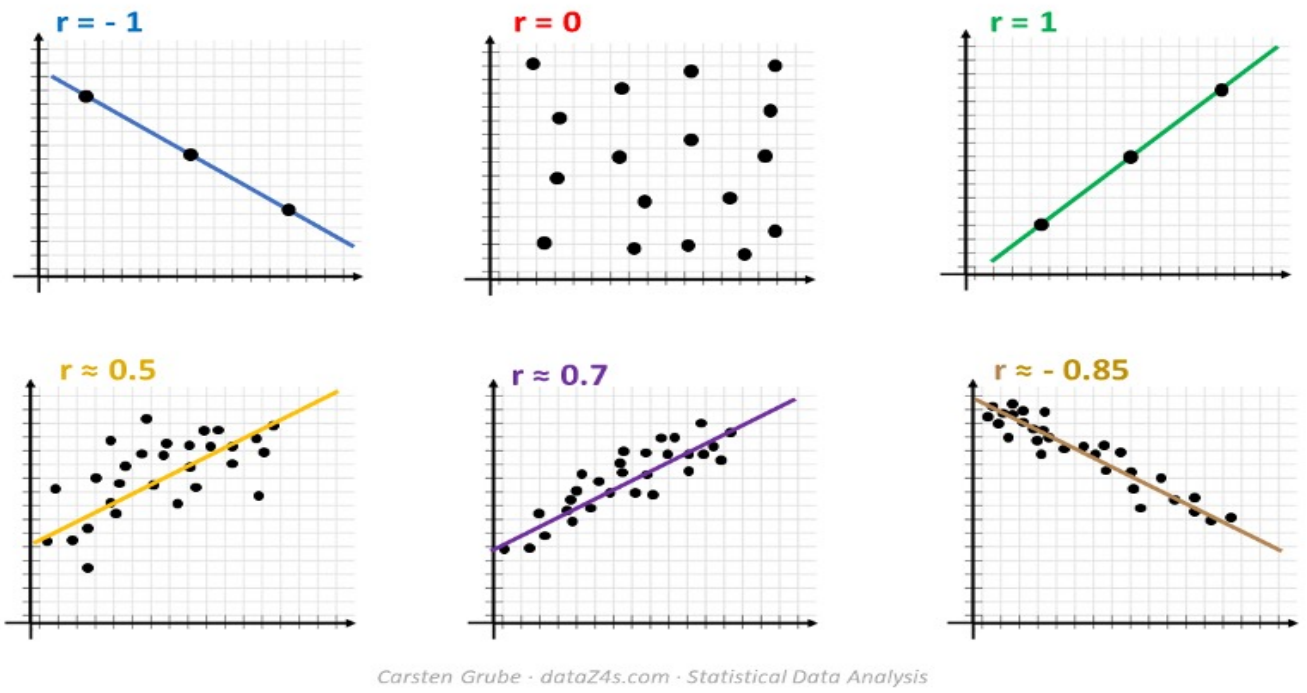
One-way ANOVA: when only 1 independent variable is considered.

Two-way ANOVA: when 2 independent variables are considered.

N-way ANOVA: when *N* number of independent variables are considered.

Correlation coefficient (R or r)

It is used to measure the strength between two variables. It is simply the square root of the coefficient of Determination and ranges from -1 to 1 where 0 represents no correlation, and 1 represents positive strong correlation while -1 represents negative strong correlation.



Q-Q (quantile-quantile) Plots

Before understanding QQ plots first understand what is a **Quantile**?

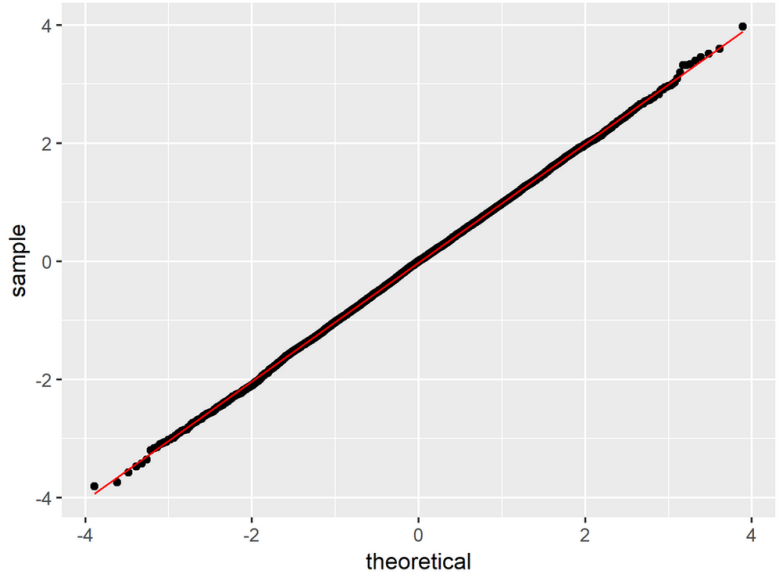
A quantile defines a particular part of a data set, i.e. a quantile determines how many values in a distribution are above or below a certain limit. Special quantiles are the quartile (quarter), the quintile (fifth), and percentiles (hundredth).

Q-Q (quantile-quantile) Plots

An example:

If we divide a distribution into four equal portions, we will speak of four quartiles. The first quartile includes all values that are smaller than a quarter of all values. In a graphical representation, it corresponds to 25% of the total area of distribution. The two lower quartiles comprise 50% of all distribution values. The interquartile range between the first and third quartile equals the range in which 50% of all values lie that are distributed around the mean.

In Statistics, A **Q-Q(quantile-quantile)** plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight(y=x).



Q-Q (quantile-quantile) Plots

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A **45-degree** angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

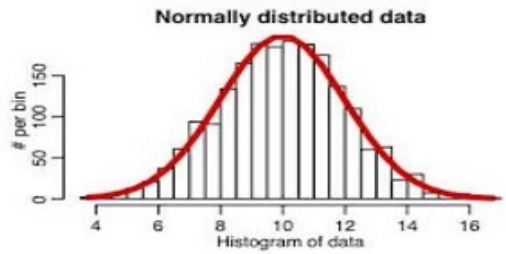
It's very important for you to know whether the distribution is normal or not so as to apply various statistical measures on the data and interpret it in much more human-understandable visualization and their Q-Q plot comes into the picture. The most fundamental question answered by the Q-Q plot is if the curve is Normally Distributed or not.

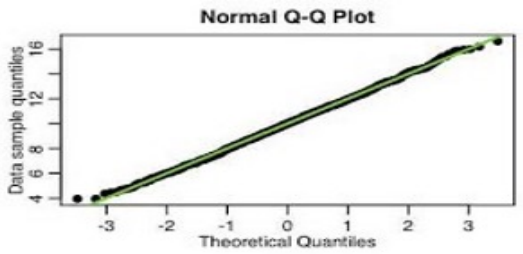
Normally distributed, but why?

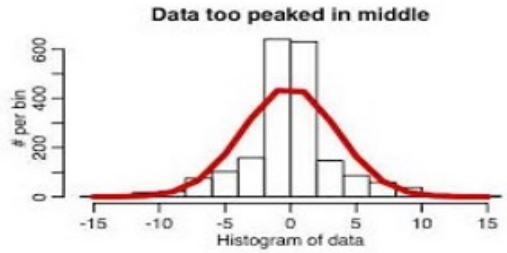
The Q-Q plots are used to find the type of distribution for a random variable whether it is a Gaussian Distribution, Uniform Distribution, Exponential Distribution, or even Pareto Distribution, etc.

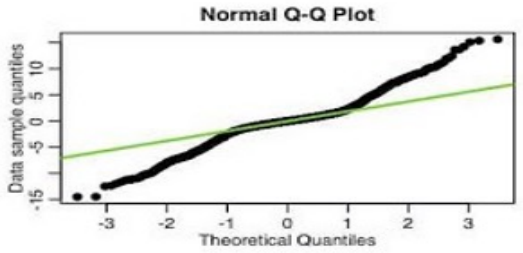
Q-Q (quantile-quantile) Plots

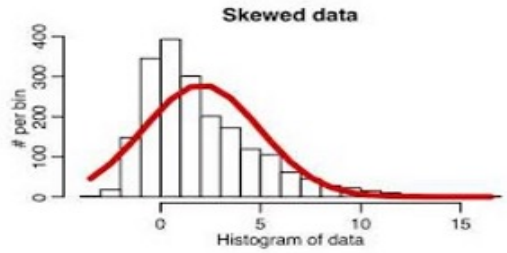
You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. In general, we are talking about Normal distributions only because we have a very beautiful concept of the 68–95–99.7 rule which perfectly fits into the normal distribution So we know how much of the data lies in the range of the first standard deviation, second standard deviation and third standard deviation from the mean. So knowing if a distribution is Normal opens up new doors for us to experiment with

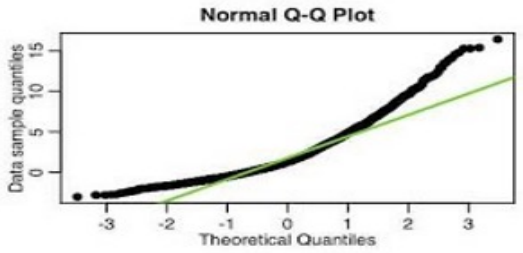












Q-Q (quantile-quantile) Plots

Skewed Q-Q plots

Q-Q plots can find skewness(measure of asymmetry) of the distribution.
If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then the distribution is **Left skewed(Negatively skewed).**

Now if upper end of the Q-Q plot deviates from the staright line and the lower is not, then the distribution is **Right skewed(Positively skewed).**

Tailed Q-Q plots

Q-Q plots can find Kurtosis(measure of tailedness) of the distribution.
The distribution with the fat tail will have both the ends of the Q-Q plot to deviate from the straight line and its centre follows the line, where as a thin tailed distribution will term Q-Q plot with very less or negligible deviation at the ends thus making it a perfect fit for normal distribution.

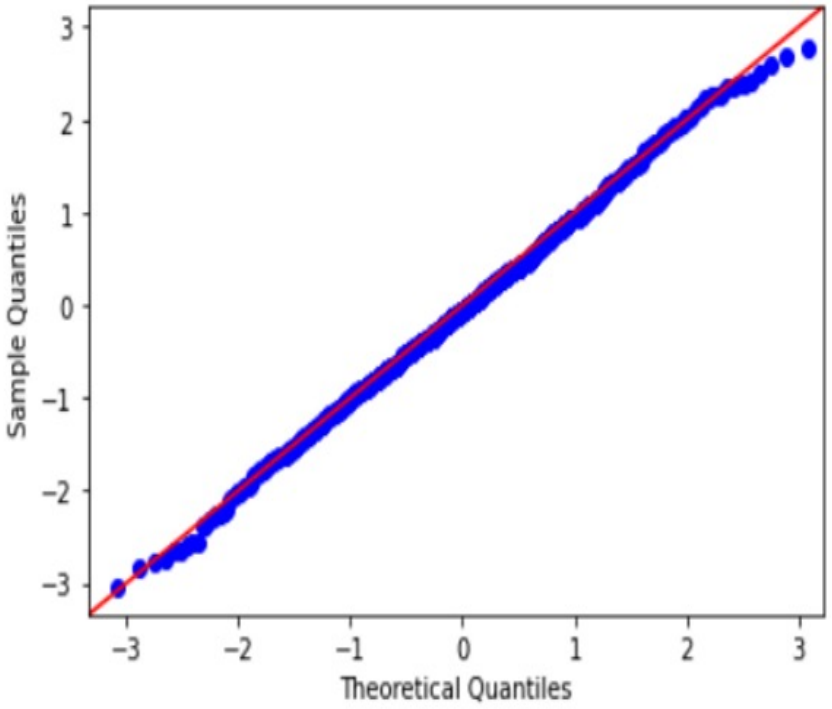
Q-Q plots in Python

```
import numpy as np

#create dataset with 1000 values that follow a normal distribution
np.random.seed(0) data = np.random.normal(0,1, 1000)
```

To create a Q-Q plot, we can use function from the statsmodels library

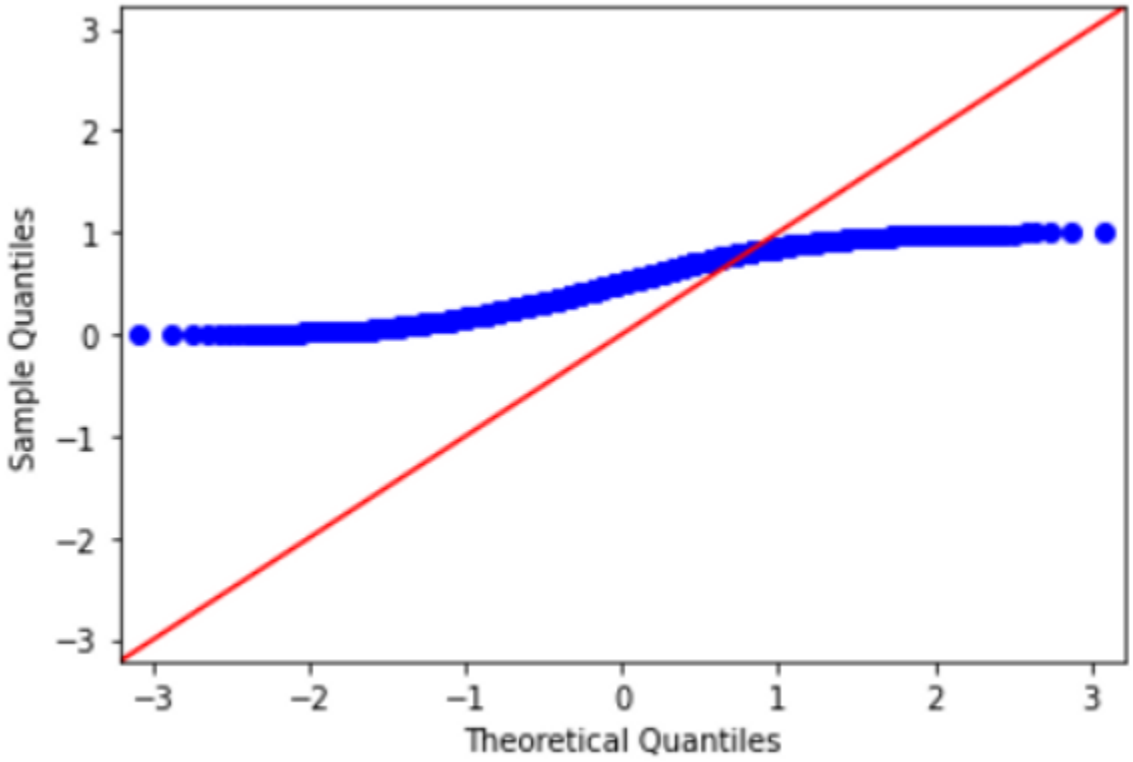
```
import statsmodels.api as sm
import matplotlib.pyplot as plt
#create Q-Q plot with 45-degree line added to plot

fig = sm.qqplot(data, line='45’)
plt.show()
```

Q-Q plots in Python



Q-Q plots in Python

```
#create dataset of 100 uniformally distributed values data =
np.random.uniform(0,1, 1000)
#generate Q-Q plot for the dataset
fig = sm.qqplot(data, line='45’)
plt.show()
```



Log-Normal Distribution

In probability theory, a **Log-normal distribution** also known as **Galton's distribution** is a continuous probability distribution of a random variable whose logarithm is normally distributed.

Thus, if the random variable **X** is log-normally distributed, then **Y = ln(X)** has a normal distribution. Equivalently, if **Y** has a normal distribution, then the exponential function of **Y i.e, X = exp(Y),** has a log-normal distribution.

Skewed distributions with low mean and high variance and all positive values fit under this type of distribution. A random variable that is log-normally distributed takes only positive real values.

The general formula for the probability density function of the lognormal distribution is:

$$f(x) = \frac{e^{-((\ln((x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}} \qquad x > \theta; m, \sigma > 0$$

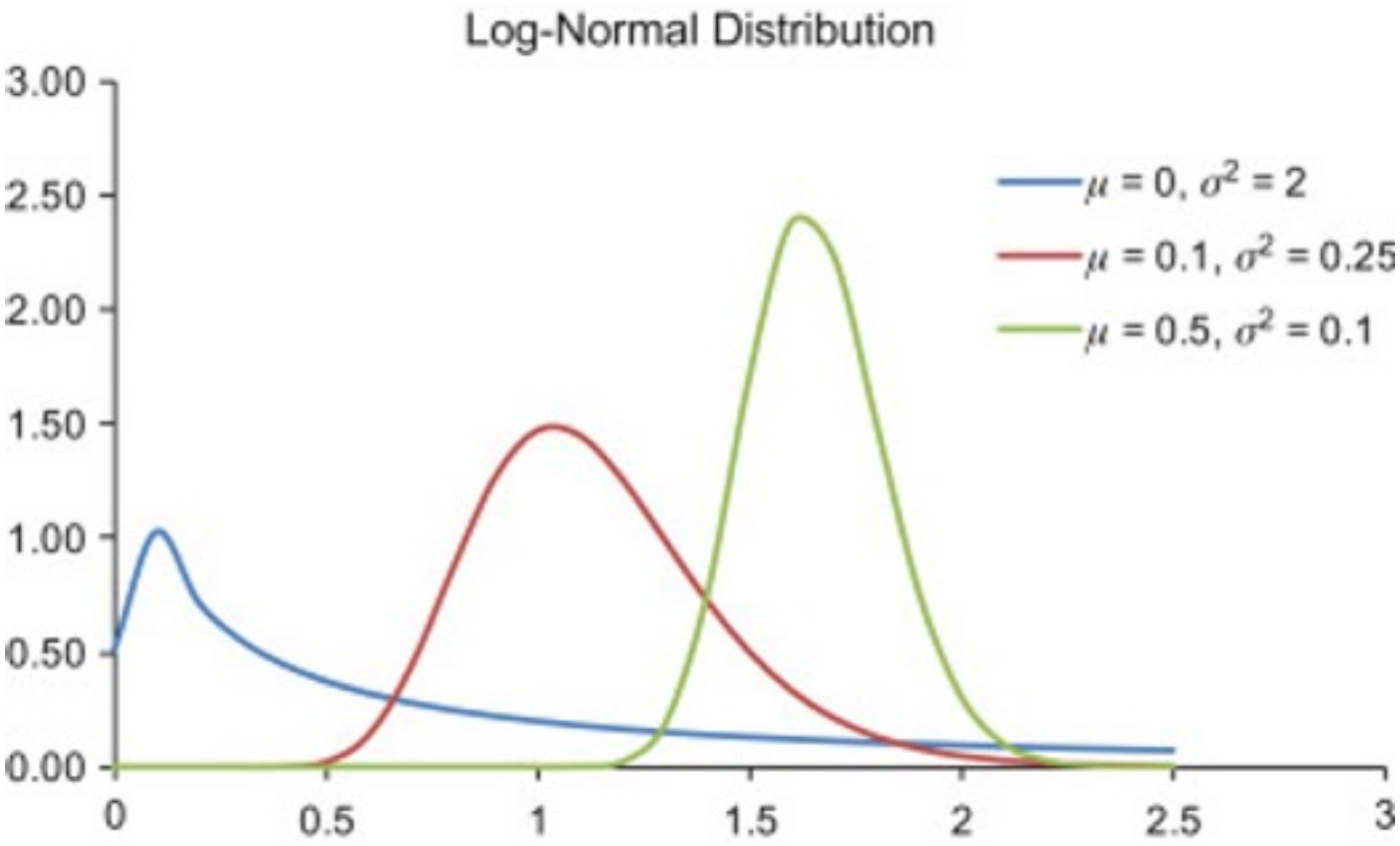
Q-Q plots in Python

The shape of Lognormal distribution is defined by 3 parameters:
σ is the shape parameter, (and is the standard deviation of the log of the distribution)
θ or **μ** is the location parameter (and is the mean of the distribution)
m is the scale parameter (and is also the median of the distribution)
The location and scale parameters are equivalent to the mean and standard deviation of the logarithm of the random variable as explained above.

If **x = θ**, then **f(x) = 0**. The case where **θ = 0** and **m = 1** is called the **standard lognormal distribution**. The case where **θ** equals zero is called the **2-parameter lognormal distribution**.

The following graph illustrates the effect of the **location(μ)** and **scale(*σ*)** parameter on the probability density function of the lognormal distribution:

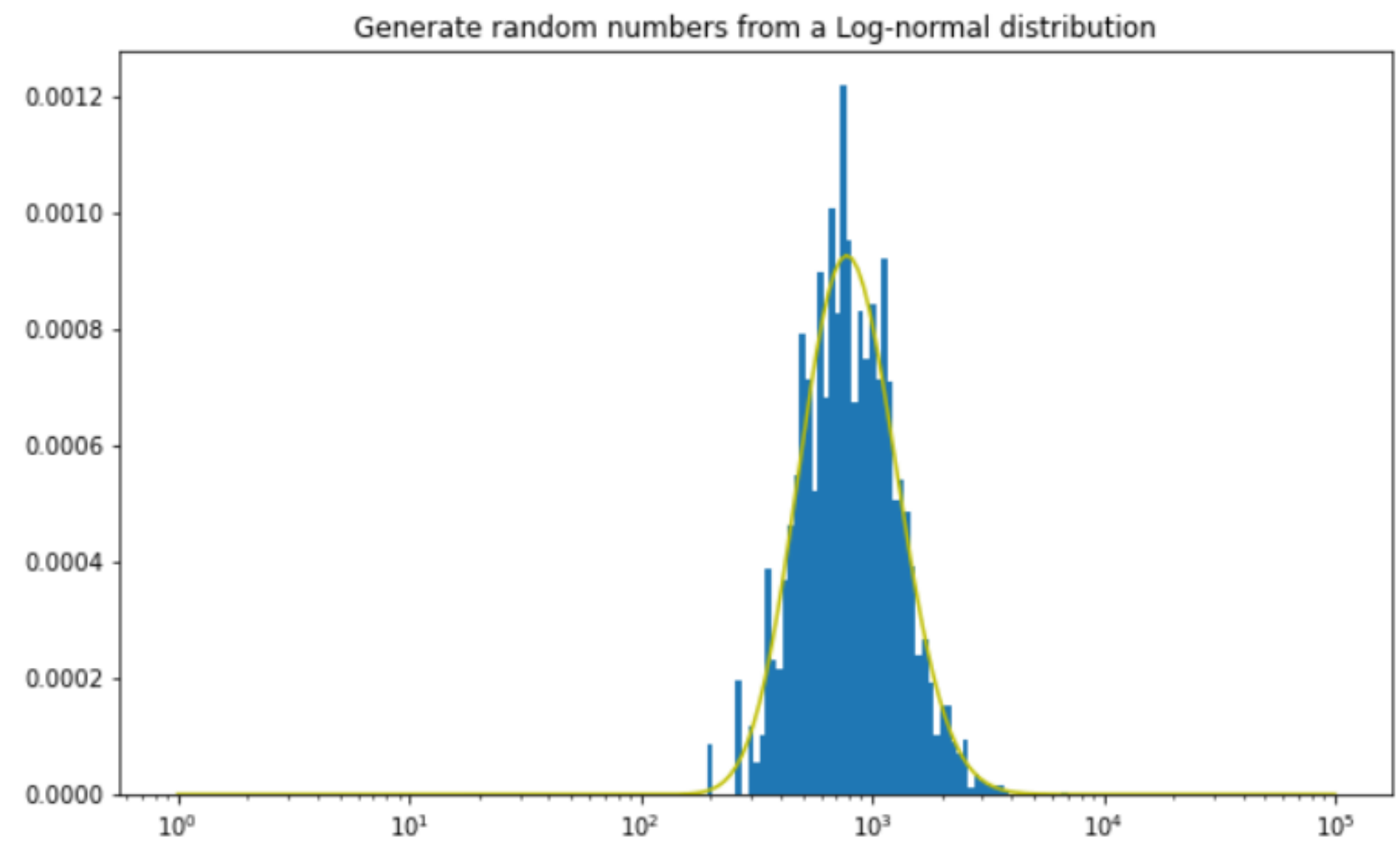
Log-Normal Distribution



Log-Normal dictribution in Python

```
import numpy as np
 import matplotlib.pyplot as plt
from scipy.stats import lognorm
np.random.seed(68)
data = lognorm.rvs(s=0.5, loc=1, scale=1000, size=1000)
plt.figure(figsize=(10,6))
ax = plt.subplot(111)
plt.title('Generate random numbers from a Log-normal distribution')
ax.hist(data, bins=np.logspace(0,5,200), density=True)
ax.set_xscale("log")
shape,loc,scale = lognorm.fit(data)
x = np.logspace(0, 5, 200)
pdf = lognorm.pdf(x, shape, loc, scale)
ax.plot(x, pdf, 'y')
plt.show()
```

Log-Normal Distribution in Python

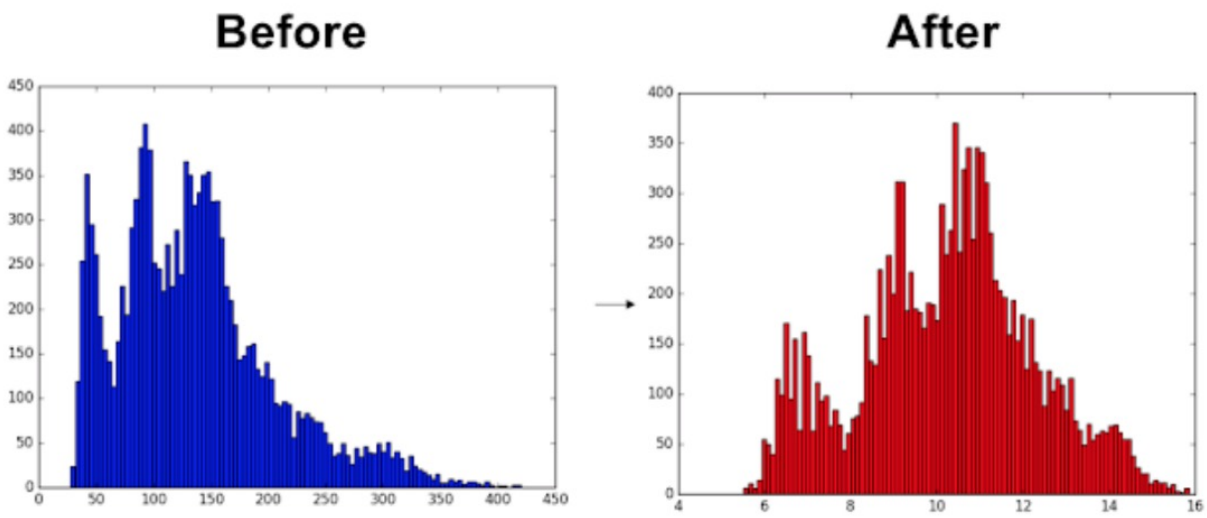


Box-Cox transformation

The **Box-Cox transformation** transforms our data so that it closely resembles a normal distribution.

The one-parameter Box-Cox transformations are defined as In many statistical techniques, We assume that the errors are normally distributed. This assumption allows us to construct confidence intervals and conduct hypothesis tests. By transforming your target variable, we can (hopefully) normalize our errors (if they are not already normal).

Additionally, transforming our variables can improve the predictive power of our models because transformations can cut away white noise.



Box-Cox transformation

At the core of the Box-Cox transformation is an exponent, **lambda (λ)**, which varies from -5 to 5. All values of **λ** are considered and the optimal value for your data is selected; The "optimal value" is the one that results in the best approximation of a normal distribution curve.

The one-parameter Box-Cox transformations are defined as:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$$

Box-Cox transformation

and the two-parameter Box-Cox transformations as:

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \lambda_1 \neq 0 \\ \log(y+\lambda_2), & \lambda_1 = 0 \end{cases}$$

Moreover, the one-parameter Box-Cox transformation holds for **y > 0**, i.e. only for positive values and two-parameter Box-Cox transformation for **y > -λ**, i.e. negative values.

The parameter **λ** is estimated using the profile likelihood function and using goodness-of-fit tests.

Box-Cox transformation

If we talk about some drawbacks of Box-cox transformation, then if interpretation is what you want to do, then Box-cox is not recommended. Because if $\boldsymbol{\lambda}$ is some non-zero number, then the transformed target variable may be more difficult to interpret than if we simply applied a log transform.

A second stumbling block is that the Box-Cox transformation usually gives the median of the forecast distribution when we revert the transformed data to its original scale. Occasionally, we want the mean and not the median.

Box-Cox transformation

If we talk about some drawbacks of Box-cox transformation, then if interpretation is what you want to do, then Box-cox is not recommended. Because if $\boldsymbol{\lambda}$ is some non-zero number, then the transformed target variable may be more difficult to interpret than if we simply applied a log transform.

A second stumbling block is that the Box-Cox transformation usually gives the median of the forecast distribution when we revert the transformed data to its original scale. Occasionally, we want the mean and not the median.

Source: https://www.theaidream.com/post/advanced-statistical-concepts-in-data-science

Box-Cox transformation I Python

```
#load necessary packages
import numpy as np
from scipy.stats import boxcox
import seaborn as sns
#make this example reproducible
np.random.seed(0)
#generate dataset
data = np.random.exponential(size=1000)
fig, ax = plt.subplots(1, 2)
#plot the distribution of data values
sns.distplot(data, hist=False, kde=True, kde_kws = {'shade': True, 'linewidth': 2}, label =
"Non-Normal", color ="red", ax = ax[0])
#perform Box-Cox transformation on original data
transformed_data, best_lambda = boxcox(data)
sns.distplot(transformed_data, hist = False, kde = True, kde_kws = {'shade': True,
'linewidth': 2}, label = "Normal", color ="red", ax = ax[1])
```

Box-Cox transformation I Python

```
#adding legends to the subplots
plt.legend(loc = "upper right")
#rescaling the subplots
fig.set_figheight(5)
 fig.set_figwidth(10)
#display optimal lambda value
print(f"Lambda value used for Transformation: {best_lambda}")
```

```
Lambda value used for Transformation: 0.2420131978174143
```