

## Statistics

1. Statistics: It is the science of conducting studies to collect, organize, summarize, analyze, and draw a conclusion out of data.
2. It deals with collective informative data, interpreting those data, and drawing a conclusion from that data.
3. It is used in many disciplines like marketing, business, healthcare, telecom, etc.

In any data science project, data helps us to analyze the initial level of insight.

## Types of Statistics

1. Descriptive
2. Inferential

### Descriptive

1. It helps us to organize and summarize data using numbers and graphs to look for a pattern in the data set.
2. Measures of Central tendency: Mean, Median, Mode.
3. The measure of Variability: Standard Deviation, Variance & Range

(\*Central Tendency- it is the single value which attempts to describe a set of data)

### Inferential

1. To make an inference or draw a conclusion from the population, sample data is used.
2. Using probability to determine how confident we can be that the conclusion we make is correct. (Confidence Interval & margin of error)

Example: Our primary concern is to find out how many people like blue cars in the data set.

Suppose, in a city, 1 lakh people are there. For our analysis, we have taken 100 people from the data set. Out of 100, 20 people like blue cars. i.e., 20/100 means 20% population like blue cars. This 20% is descriptive Statistics.

If we say 20%  $\pm$  2%, i.e., 20% people with 2% margin of error like blue cars. So in this, we are 98% sure that this is correct. This is called inferential.

## **Introduction to basic Statistics terms**

- **Population:** A collection, or set, of individuals or objects or events whose properties are to be analyzed ( Raw data that we get from the client).
- **Sample:** A subset of the population. It should be representative of the population.
- **Variable:** A characteristic of each element of a population or sample.
- **Data (singular):** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.
- **Data (plural):** The set of values collected for the variable from each of the elements belonging to the sample.
- **Experiment:** A planned activity whose results yield a collection of data.
- **Parameter:** A numerical value summarizing all the data of an entire population (Mean, Median, Mode).
- **Statistic:** A numerical value summarizing the sample data.

## **Descriptive Statistics**

**Descriptive Statistics:** It is a method of organizing, summarizing, and presenting data in an informative way.

**Example:** You have all the data on how the business is going on, how much inventory you keep, how many customers come to your store, In which month it has been more, at what day of the week it occurs more. Which product is being sold more at what point of time, on what hours is your product sold more. What kind of customers come, do male customers come more at a certain point in time, or female customers come then. People with children come more, cigarette buyers come more, or beer buyers come more, or grocery item buyers come more.

Descriptive Statistics answers all these questions based on data.

## **Types of data**

- 1) Categorical
- 2) Numerical

- 1) Categorical Data represents groups or categories

Examples: \* Car brands: Audi, BMW and Mercedes.

\* Answers to yes/no questions: yes or no.

2) Numerical data represent numbers. It is divided into two groups: Discrete and Continuous.

Discrete data can be usually counted finitely, while continuous is infinite and impossible to count.

**Examples:**

Discrete: \* Number of children you want to have: 1, 2, 3.

\* Grades at University: 0 to 100 %

\* Number of Objects: Bottles, glasses, tables or cars

Continuous: Height, Area, Distance, Time

**Levels of measurement**

1) Qualitative: A variable that categorizes or describes a population element.

Note that arithmetic operations such as addition and averaging are meaningless for data resulting from a qualitative variable.

2) Quantitative: A variable that quantifies a population element.

Note that arithmetic operations, such as addition and average, are meaningful for data resulting from a quantitative variable.

1) Qualitative is classified into the following two levels:

- Nominal
- Ordinal

The nominal level represents categories that cannot be put in any order, while ordinal represents categories that can be ordered.

Examples: \* Nominal: four seasons (winter, spring, summer, autumn)

\* Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)

2) There are two quantitative levels: interval and ratio.

They both represent “numbers”; however, ratios have a true zero, while intervals don’t.


Examples:

\* Interval: degrees Celsius and Fahrenheit

\* Ratio: degree Kelvin.

## Measures of Central Tendency

# Measures of Central Tendency



most *representative* or *typical* of all values in a group  
“average”

MODE	MEDIAN	MEAN
<ul style="list-style-type: none"> <li>• most frequent data point</li> <li>• mode exists as a data point</li> <li>• unaffected by extreme values</li> <li>• useful for qualitative data</li> <li>• may have more than 1 value</li> </ul>	<ul style="list-style-type: none"> <li>• value that divides ranked data points into halves: 50% larger than it, 50% smaller</li> <li>• may not exist as a data point in the set</li> <li>• influenced by position of items, but not their values</li> </ul>	$\bar{x} = \frac{\sum x}{N}$ <ul style="list-style-type: none"> <li>• most stable measure</li> <li>• affected by extreme values</li> <li>• may not exist as a data point in the set</li> </ul>

### Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Here n is the size of the data set,  $\bar{x}$  is the sample mean, and  $x_i$  the numbers in sequence.

$\sum$  is the summation of the entire data set

Similarly, for a data population of size N, the population mean is:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

**Example:** The systolic blood pressure of seven middle-aged men in:

150, 123, 134, 170, 146, 124 and 113.

The Mean is =  $(150+123+134+170+146+124+113)/7 = 137.14$

## Mode and Median

- The median for the sample data arranged in increasing order is defined as :
  - i. If "n" is an odd number - Middle value
  - ii. If "n" is an even number - Midway between the two middle values
- The mode is the most commonly occurring value.
- Mode exists as a data point.
- Useful for qualitative data.

### Example – if n is odd

The re-ordered systolic blood pressure data:

113,124,125,132,146,151 and 170.

-> The median here is 132.

-> Two individuals have systolic blood pressure = 124mm Hg, so the Mode is 124.

### Example – if n is even

Six men with high cholesterol participated in the study to investigate the effects of diet on cholesterol levels. At the beginning of the study, their cholesterol levels (mg/dl) were as follows:

366, 327, 274, 292, 274 and 230

Rearrange the data in ascending order as follows:

230, 274, 274, 292, 327 and 366.

-> The median is 283(average of 274 and 292).

-> The mode between the two men having the same cholesterol level = 274.

## Mean, Mode and Median in Brief:

### MEAN

The "mean" is the "average". To find the mean, you add up all the numbers and then divide by the number of numbers.

TO FIND THE MEAN FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13  
average the set of numbers:

$(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) \div 9 = 15$

Note that the mean isn't a value from the original list. This is a common result. DO NOT assume that the mean will be one of the original numbers.

### MEDIAN

The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in **numerical order**, so you may have sort the list first.

FOR AN ODD NUMBER OF VALUES: 1,5,2,8,7  
**Sort the numbers 1, 2, 5, 7, 8**

FOR AN EVEN NUMBER OF VALUES: 1,5,2,10,8,7  
**Sort the numbers: 1, 2, 5, 7, 8, 10.**

TAKE THE AVERAGE OF THE TWO MEAN NUMBERS:  $(5+7)/2 = 6$

### MODE

The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.

TO FIND THE MODE FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13

**Sort the numbers: 13, 13, 13, 13, 14, 14, 16, 18, 21**

Note:

- Mean is highly sensitive to outliers
  - Example:
    - 1,2,3,4,5
      - > Mean: 3
      - > Median: 3
    - 1,2,3,4,5,100
      - > Mean: 51.5
      - > Median: 3.5

### Calculating Mean (Python Code):

```
import numpy as np
expenditure = np.random.normal(25000, 15000, 10000)
np.mean(expenditure)
```

24820.023388888958

### Calculating Median:

```
np.median(expenditure)
```

```
24691.98032372038
```

Now, we are adding a large number to the sample.

```
expenditure = np.append(expenditure, [10000000000])
```

```
np.median(expenditure)
```

```
24698.883118187983
```

```
np.mean(expenditure)
```

```
424650.16332356015
```

Here, the Median did **not** change much, but the Mean did.

### Calculating Mode:

Let's generate a random expenditure set data using the script below.

```
expenditure = np.random.randint(15, high=50, size=200)
expenditure
```

This gives us the following output:

```
array([40, 15, 46, 22, 17, 45, 23, 29, 15, 24, 41, 34, 36, 37, 32, 17, 38,
       31, 25, 28, 37, 48, 45, 31, 41, 47, 36, 39, 37, 29, 48, 38, 46, 28,
       26, 39, 17, 18, 37, 22, 31, 30, 15, 24, 15, 23, 19, 31, 20, 38, 39,
       42, 47, 27, 19, 24, 27, 34, 21, 20, 38, 15, 28, 48, 32, 41, 29, 21,
       32, 19, 38, 28, 39, 32, 45, 22, 15, 24, 41, 23, 15, 30, 31, 47, 29,
       48, 35, 32, 16, 33, 37, 43, 34, 25, 39, 22, 45, 34, 20, 45, 36, 15,
       24, 24, 29, 42, 41, 47, 48, 33, 44, 25, 49, 39, 41, 41, 39, 34, 48,
       47, 29, 22, 45, 24, 32, 46, 44, 47, 43, 24, 28, 15, 15, 47, 46, 29,
       35, 21, 15, 40, 31, 37, 44, 38, 15, 48, 48, 17, 15, 29, 25, 40, 37,
       35, 33, 47, 26, 48, 16, 20, 37, 32, 37, 30, 44, 25, 49, 19, 41, 15,
       19, 27, 36, 25, 16, 49, 21, 49, 36, 44, 31, 23, 34, 35, 15, 43, 44,
       36, 29, 22, 27, 49, 46, 31, 24, 40, 43, 41, 36, 28])
```

```
from scipy import stats
stats.mode(expenditure)
```

```
ModeResult(mode=array([15]), count=array([15]))
```



## Measures of Dispersion

### MEASURES OF DISPERSION

The mean, median or mode is usually not by itself a sufficient measure to reveal the shape of a distribution of a data set. We also need a measure that can provide some information about the variation among data set values.

The measures that help us to know about the spread of a data set are called **measures of dispersion**.

The measures of central tendency and dispersion taken together give a better picture of a data set.

We consider 3 measures of dispersion:

1. **Range**
2. **Variance**
3. **Standard Deviation**

### Range

- The difference between the smallest and the largest observations in the sample is called Range.
- For example, the minimum and maximum blood pressure are 113 and 170, respectively. Hence the range is 57.
- It is easy to calculate.
- It's implemented for both "best" or "worst" case scenarios.
- Too sensitive for extreme values.

### Sample and Population

- Population - The entire set of objects or individuals or interests or the measurements obtained from all individuals or objects of interest.

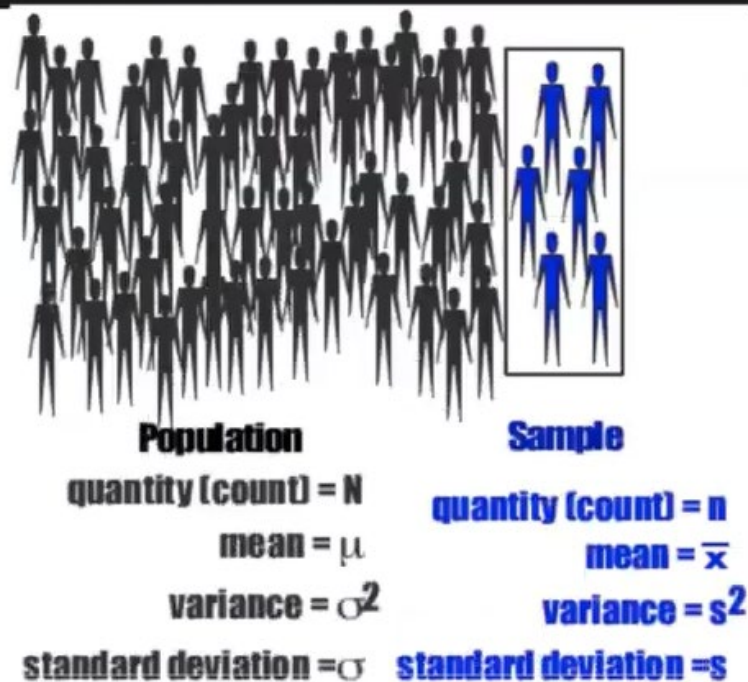
-> Finite

-> Infinite

- Sample - A portion, or part, of the population of interest.



# Population vs. Sample



## Variance:

Variance is a measure of dispersion in a data set.

It is measured by first finding the Deviation of each element in a data set from the mean, and then by squaring it. Variance is an average of all squared deviations.

The below figure shows that On an avg how far point is distributed from the mean ( $\bar{x}$ )



- Here(left side) variance is high because, from the mean( $\bar{x}$ ), the points are distributed at a longer distance as compared to the right side, where the distance is a bit smaller.

**Q. Consider a list of random integers 3,3,3,5,6,1. We will now calculate the variance using the numpy library.**

```
import numpy as np
results = [3,3,3,5,6,1]
np.var(results)
```

```
2.5833333333333335
```

The variance of the random data set is 2.58.

The sample variance,  $s^2$ , is the arithmetic mean of the squared deviations from the sample mean:

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

### Standard deviation

- Standard deviation tells us about the concentration of data around the mean of the data set.
- Standard deviation ( $S$ ) is the square root of the variance.

**Formula to calculate standard deviation:**

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

- The sample standard deviation has the advantage of being in the same units as the original variable ( $x$ ).

To measure standard deviation, use the inbuilt function “std” from numpy, as shown below:

```
np.std(results)
```

```
1.6072751268321592
```

## Facts about Standard Deviation:

- If the standard deviation is small, the data has little spread (i.e., the majority of points fall very near the mean).
- If standard deviation = 0, there is no spread. This only happens when all data items are the same value.
- The standard deviation is significantly affected by outliers and skewed distributions.

## Sample Vs. Population

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad \text{Vs.} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

*Population Mean*                      *Sample Mean*

## The variance of Population and Sample

· The variance of a population is:

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

Population Mean (points to  $\mu$ )  
Population Size (points to  $N$ )

➤ The variance of a sample is:

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

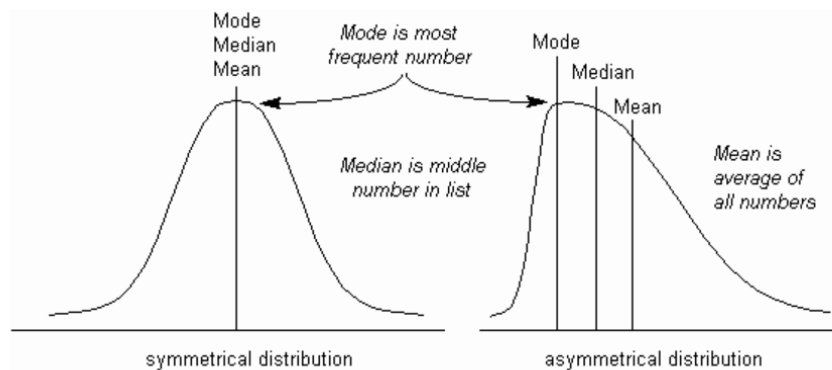
Sample Mean (points to  $\bar{x}$ )

Note! the denominator is sample size (n) minus one !

**Note:** In the sample variance formula, the denominator has n-1 instead of n, where n is the number of observations in the sample. This use of 'n-1' is the Bessel's correction method. The reason behind using this method is, it corrects the bias in the estimation of the population variance.

	Population	Sample
Size	N	n
Mean	$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

## Symmetrical and asymmetrical distribution



*# Using Numpy & Scipy to calculate Mean, Mode, and Median*

```
import numpy as np
```

```
from scipy import stats
```

```
dataset = [5,5,2,3,4,6,18]
```

*#mean value*

```
mean = np.median(dataset)
```

*#median value*

```
median = np.median(dataset)
```

*#mode value*

```
mode = stats.mode(dataset)
```

```
std = np.std(dataset)
```

```
vr = np.var(dataset)
```

```
print("Mean:", mean)
```

```
print("Median:", median)
```

```
print("Mode:", mode)
```

```
print("STD", std)
```

```
print("Var:", vr)
```

Outputs:

Mean: 5.0

Median: 5.0

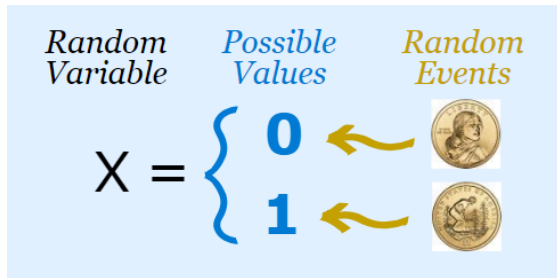
Mode: ModeResult(mode=array([5]), count=array([2]))

STD 4.997958767010258

Var: 24.9795918367347

## Random Variables

A random variable is a set of all the possible values from a random experiment.



A random variable is a variable whose possible values are outcomes of a random phenomenon. Random variables can be discrete or continuous. Discrete variables can only take specific values, while continuous random variables can take any value (within a range).

### Discrete Random Variable

- The discrete random variable is one that may take on only a countable number of distinct values.
- If the random variable can take only the finite number of distinct values, then it must be a discrete random variable.

Examples of discrete random variables :

- \* Number of children in the family.
- \* The Friday night attendance at a Multiplex.
- \* The number of patients in the doctor's surgery.
- \* The number of defective light bulbs in the box.
- The probability distribution of the discrete variable is the list of the probabilities associated with every possible value.

Suppose a random variable  $X$  may take  $k$  different values, with a probability that  $X = x_i$  defined to be  $P(X = x_i) = p_i$ . The probabilities  $p_i$  must satisfy the following:

- 1:  $0 \leq p_i \leq 1$  for each  $i$
- 2:  $p_1 + p_2 + \dots + p_k = 1$ .

### Continuous Random Variables:

It is the one that takes an infinite number of possible values. Continuous random variables usually are the measurements. For example, the height of a person, the weight of a machine, the amount of sugar in tea, etc. A continuous random variable is defined throughout values and is represented by the area under a curve.

## Set

### Set Definition

- A **set** is a well-defined collection of objects.
- A set that contains zero elements is called a **null set (empty set)**.
- Let A and B be two sets. Then A is said to be a subset of B (or B is a superset of A) if every element of A belongs to B.
- A set may be defined by mentioning its elements written in brackets.

For example, if Set X consists of the numbers 1, 2, 3, and 8, we may say  $X = \{1, 2, 3, 8\}$ .

- An empty set is denoted by  $\{ \}$ , which is called a null set.

### Operations of Set

- The **union** of two sets is defined as the set of elements that are present in one or both sets. Thus, if A is  $\{1, 2\}$  and B is  $\{2, 3, 4\}$ , the union of sets A and B is:

$$A \cup B = \{1, 2, 3, 4\}$$

- The common elements that are present in both sets are known as the intersection of two sets. Thus, if A is  $\{1, 2, 5\}$  and Y is  $\{2, 3, 4, 5\}$ , the intersection of sets A and B is:

$$A \cap B = \{2, 5\}$$

- The **complement** of an event is defined as it is the set of all the elements but not in the event.
- Thus, if the sample space is  $\{0, 2, 3, 4, 8, 9\}$ , and X is  $\{2, 3, 4\}$ , then the complement of set X is

$$X' = \{0, 8, 9\}$$

## Examples

1. The set of consonants.

If B is the set of consonants, then B could be described as

$$B = \{b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x\}$$

2. Set  $X = \{1, 2, 3\}$  and Set  $Y = \{3, 2, 1\}$ . Is Set X equal to Set Y?

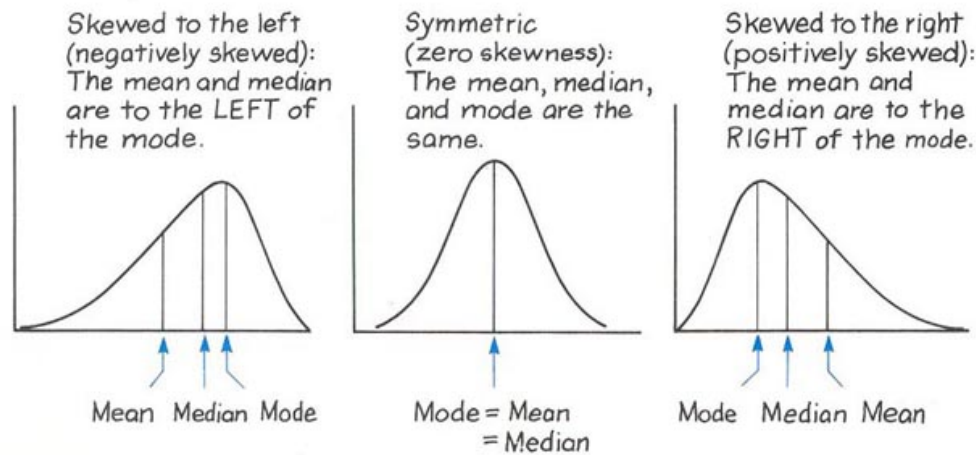
Yes. Two sets are equal only if they have identical elements.

3. Set  $X = \{1, 2, 4\}$  and Set  $Y = \{1, 2, 3, 5, 6\}$ . Is Set X a subset of Set Y?

Set X would be a subset of set Y if *every* element from Set X were also in Set Y.



**Skewness:** Skewness is defined as a measure of the dataset's symmetry. A perfectly symmetrical data set will have zero skewness.



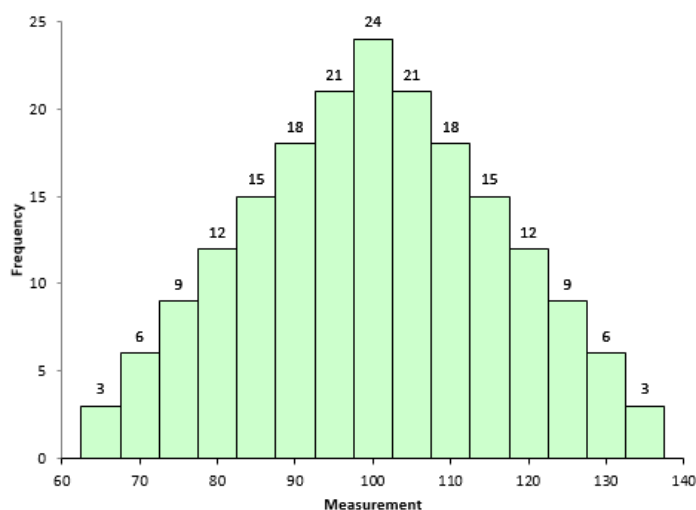
The skewness is defined as:

$$a_3 = \frac{\sum (X_i - \bar{X})^3}{ns^3}$$

Where  $X_i$  is the  $i^{\text{th}}$  X value,  $n$  is the sample size,  $\bar{x}$  is the average, and  $s$  is the sample standard deviation.

$$Skewness = \frac{n}{(n-1)(n-2)} \sum \frac{(X_i - \bar{X})^3}{s^3} = \frac{n}{s^3(n-1)(n-2)} (S_{above} - S_{below})$$

This sample size formula is used here.



The figure above is for an asymmetrical data set. This data set was created by generating the data from 65 to 135 in 5 number of steps with the number of each value, as shown in Figure above.

**The above figure shows Symmetrical Data set with Skewness equals to 0**

For example, there are three 65's, six 70's, and nine 75's, etc.

The set of symmetrical data has a skewness equal to 0, where Each X value is subtracted from the average. So, if a collection of data is symmetrical, for each point that is a distance “d” above the average, there will be a point that is a distance “-d” below the average.

Consider 65 value and 135 value. The average of the data in the above figure is 100.

when  $X = 65$

$$\frac{(X_i - \bar{X})^3}{s^3} = \frac{(65 - 100)^3}{s^3} = \frac{(-35)^3}{s^3} = \frac{-4278}{s^3}$$

For  $x = 135$  then:

$$\frac{(X_i - \bar{X})^3}{s^3} = \frac{(135 - 100)^3}{s^3} = \frac{(35)^3}{s^3} = \frac{4278}{s^3}$$

So, the -4278 value and the value of +4278 even out at 0. So, a Symmetrical data set will have 0 skewness.

To explore +ve & -ve values of skewness, let's define the following terms:

$$S_{\text{above}} = |\sum (X_i - \bar{X})^3| \text{ if } X_i \text{ is above the average}$$

$$S_{\text{below}} = |\sum (X_i - \bar{X})^3| \text{ if } X_i \text{ is below the average}$$

So, when  $X_i$  is above the average,  $S_{\text{above}}$  is the “size” of the deviations from average. Likewise, when  $X_i$  is below the average,  $S_{\text{below}}$  can be viewed as the “size” of the deviations from average.

Then the skewness becomes:

$$Skewness = \frac{n}{(n-1)(n-2)} \sum \frac{(X_i - \bar{X})^3}{s^3} = \frac{n}{s^3(n-1)(n-2)} (S_{\text{above}} - S_{\text{below}})$$

The skewness will be positive if  $S_{\text{above}}$  is larger than  $S_{\text{below}}$ . This means that the right-hand tail will be longer than the left-hand tail. Figure 2 is an example of this. Skewness for this dataset is 0.514. Positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail.

**Fig 2: A dataset with Positive Skewness**

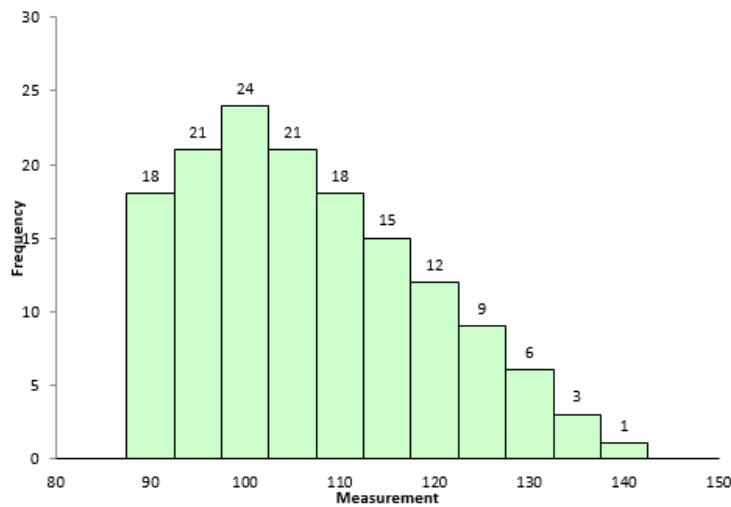
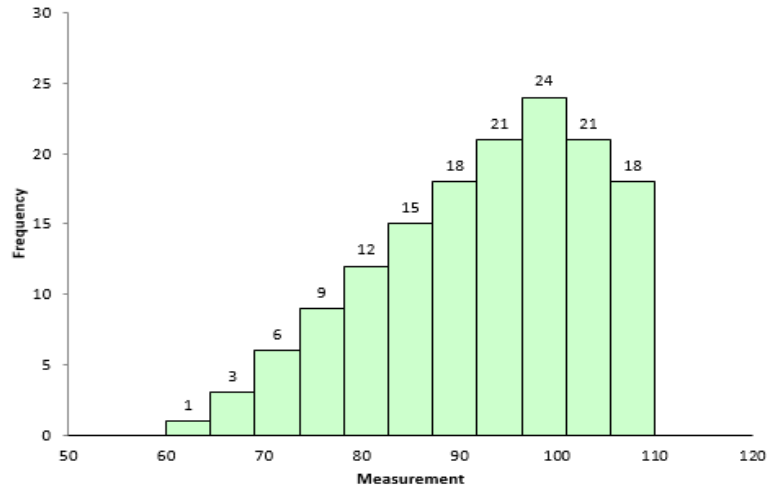


Figure 3 is an example of the dataset with negative skewness. It is the mirror image, necessarily of Figure 2. The skewness is -0.514. In this case,  $S_{\text{above}}$  is smaller than  $S_{\text{below}}$ . The left-hand tail will typically be longer than the right-hand tail.

**Figure 3: Negative Skewness with Dataset**



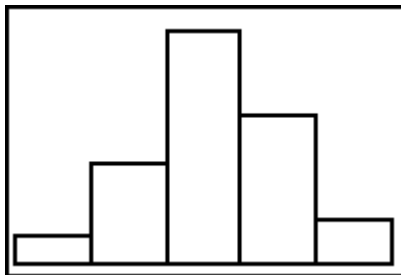
So, when is the skewness too much?

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and -0.5 or between 0.5 and 1, the data is moderately skewed.
- If the skewness is greater than 1 or less than -1, the data is highly skewed.

**Example:** Heights of men in a college

Height (inches)	Class Mark, $x$	Frequency, $f$
59.5–62.5	61	5
62.5–65.5	64	18
65.5–68.5	67	42
68.5–71.5	70	27
71.5–74.5	73	8

In the below histogram the data are skewed towards left:



To know how the dataset is highly skewed compared to other data sets, we need to compute the skewness.

$$n = 4 + 19 + 42 + 27 + 8 = 100$$

$$\bar{x} = (61 \times 5 + 64 \times 18 + 67 \times 42 + 70 \times 27 + 73 \times 8) \div 100$$

$$\bar{x} = 9305 + 1152 + 2814 + 1890 + 584 \div 100$$

$$\bar{x} = 6745 \div 100 = 67.45$$

Now, with the mean in hand, we can compute the skewness.

Class Marks, $x$	Frequency, $f$	$x \cdot f$	$(x - \bar{x})$	$(x - \bar{x})^2 f$	$(x - \bar{x})^3 f$
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15

Class Marks, $x$	Frequency, $f$	$x*f$	$(x-\bar{x})$	$(x-\bar{x})^2f$	$(x-\bar{x})^3f$
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
$\Sigma$		6745	n/a	852.75	-269.33
$\bar{x}, m_2, m_3$		67.45	n/a	8.5275	-2.6933

Finally, the skewness is

$$g_1 = m_3 / m_2^{3/2} = -2.6933 / 08.5275^{3/2}$$

$$= -0.1082$$

## Probability Density Function

A Probability Distribution is a mathematical function through which the probability of occurrence of different possible outcomes in an experiment can be calculated.

If the probability of an event is higher, it's more likely that the event will occur.

### For Example:

While tossing a fair (unbiased) coin, there could be a chance of the possibility of two outcomes ("heads" and "tails"), which are equally probable.

The probability of getting a head or a tail is 50 % or 0.5.

### Types of the probability distribution

There are many different types of probability distribution. Some of them which we will be covering in this blog are listed below:

- Normal Distribution
- Bernoulli's Distribution
- Binomial Distribution
- Uniform Distribution
- Student's T Distribution
- Poisson Distribution

### Expected value

- $E(C) = C$ 
  - The Expected Value of a Constant is only a value of a constant.
- $E(X + C) = E(X) + C$
- $E(CX) = cE(X)$ 
  - We can "pull" a constant out of an expected value expression as a part of a sum with a random variable X.

## Binomial Distribution

The binomial distribution is used when there is more than one outcome of a trial. These outcomes are labeled as “Success” and “Failure.”

Here, the probability of both outcomes is the same for all the trials.

Each trial is independent. The parameters of a binomial distribution are p and n, where p is the probability of success in the individual trial, and n is the total number of trials.

### Binomial distribution:

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n}$$

$$= \frac{N!}{n! (N-n)!} p^n (1-p)^{N-n}$$

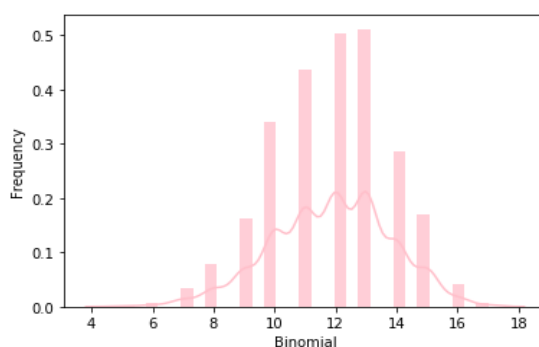
where  $\binom{N}{n}$  is a binomial coefficient, and p and q refers to success and failure, respectively.

$P_p(n|N)$  gives the discrete probability distribution, obtaining n successes out of N trials.

Python binomial distribution tells us the probability how often there will be a success in ‘n’ independent experiments. Such experiments are yes-no questions. One example may be tossing a coin.

```
import seaborn
from scipy.stats import binom
data=binom.rvs(n=17,p=0.7,loc=0,size=1010)
ax=seaborn.distplot(data,kde=True,color='pink',hist_kws={"linewidth": 22,'alpha':0.77})
ax.set(xlabel='Binomial',ylabel='Frequency')
```

```
ut[68]: [Text(0, 0.5, 'Frequency'), Text(0.5, 0, 'Binomial')]
```





## How to Work a Binomial Distribution Formula:

### Examples

**Question:** Hospital records show that of patients suffering from a specific disease, '75%' die of it. What is the probability that of six randomly selected Patients, four will recover?

**Solution:** This is a **binomial** distribution because the reason is that there are only two outcomes (the patient dies, or does not).

Let  $X$  = number who recover.

Here,  $n=6$  and  $x=4$ .

$p=0.25$  (success, i.e., they live),  $q=0.75$  (failure, i.e., they die).

The probability that four will recover:

$$P(X) = C_x^n p^x q^{n-x} = C_4^6 (0.25)^4 (0.75)^2 = 15 \times 2.1973 \times 10^{-3} = 0.0329595$$

**Question:** A die is rolled 3 times. What is the probability of

(a) No fives turning up?

(b) 1 five?

(c) 3 fives?

**Solution:**

There are only two possible outcomes (we get a 5, or we do not).

Now,  $n=3$  for each part.

Let  $X$  = number of fives appearing.

(a) When,  $x = 0$ .

$$P(X = 0) = C_x^n p^x q^{n-x} = C_0^3 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^3 = \frac{125}{216} = 0.5787$$

(b) When,  $x = 1$ .

$$P(X = 1) = C_x^n p^x q^{n-x} = C_1^3 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^2 = \frac{75}{216} = 0.34722$$

(c) When,  $x = 3$ .

$$P(X = 3) = C_x^n p^x q^{n-x} = C_3^3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^0 = \frac{1}{216} = 4.6296 \times 10^{-3}$$

**Question:** In the old days, there was a probability of '0.8' of success in any attempt to make a telephone call. (This often depended on the importance of a person making the call, or the operator's curiosity!)

Calculate the probability of having seven successes in just ten attempts.

**Solution:**

Probability of success  $p=0.8$ , so probability of failure  $q = 0.2$ .

$X$ = success in getting through.

Probability of 7 successes in 10 attempts:

Probability =  $P(X=7)$

$$= C_7^{10} (0.8)^7 (0.2)^{10-7}$$

$$= 0.20133$$

**Question:** A (blindfolded) marksman finds that on the average, he hits the target '4' times out of '5'. If he fires '4' shots, what is a probability of

(a) more than '2' hits?

(b) at least '3' misses?

**Solution:**

Here,  $n=4$ ,  $p=0.8$ , ' $q$ ' = 0.2.

Let  $x$  be the number of hits.

Let  $x_0$  = no hits,  $x_1$  = one hit,  $x_2$  = two hits, etc.

$$\begin{aligned}
 \text{(a) } P(X) &= P(x_3) + P(x_4) \\
 &= C_3^4 (0.8)^3 (0.2)^1 + C_4^4 (0.8)^4 (0.2)^0 \\
 &= 4(0.8)^3 (0.2) + (0.8)^4 \\
 &= 0.8192
 \end{aligned}$$

(b) 3 misses mean 1 hit, and 4 misses mean 0 hits.

$$\begin{aligned}
 P(X) &= P(x_1) + P(x_0) \\
 &= C_1^4 (0.8)^1 (0.2)^3 + C_0^4 (0.8)^0 (0.2)^4 \\
 &= 4(0.8)^1 (0.2)^3 + (0.2)^4 \\
 &= 0.0272
 \end{aligned}$$

**Question:** A manufacturer of metal pistons finds that on the average, '12%' of his pistons get rejected because either they are oversize or they are undersize. What is a probability that a batch of ten pistons will contain?

(a) no more than '2' rejects? (b) at least '2' rejects?

**Solution:** Let  $X$ = number of rejected pistons

(In this case, "success" means rejection!)

Here,  $n=10$ ,  $p=0.12$ ,  $q=0.88$ .

(a) No rejects. That is when  $x=0$ :

$$P(X) = C_x^n p^x q^{n-x} = C_0^{10} (0.12)^0 (0.88)^{10} = 0.2785$$

One reject. That is when  $x=1$ :

$$P(X) = C_1^{10} (0.12)^1 (0.88)^9 = 0.37977$$

Two rejects. That is when  $x=2$ :

$$P(X) = C_2^{10} (0.12)^2 (0.88)^8 = 0.23304$$

So the probability is:

$$\begin{aligned} P(X \leq 2) \\ &= 0.2785 + 0.37977 + 0.23304 \\ &= 0.89131 \end{aligned}$$

(b) We could work out every case for  $X = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ , but it is much easier to solve as follows:

Probability of at least 2 rejects

$$\begin{aligned} &= 1 - P(X \leq 1) \\ &= 1 - (P(x=0) + P(x=1)) \\ &= 1 - (0.2785 + 0.37977) \\ &= 0.34173 \end{aligned}$$

### Question:

A company drills 9 wild-cat oil exploration wells, each with an estimated probability of success of 0.1. What is the probability that all nine wells fail?

### Solution:

Let's do 20,000 trials of the model, and count the number that generates zero positive results.

```
sum(np.random.binomial(9, 0.1, 20000) == 0)/20000.
```

```
0.3918
```

**Note:** Suppose a binomial experiment consists of  $n$  trials and results in  $x$  successes. If the probability of success on an individual trial is ' $P$ ', binomial probability:

$$b(x; n, P) = {}^nC_x * P^x * (1 - P)^{n-x}$$

**Question:**

1. Generate a random number between '0' and '1'. If that number is 0.5, then it will count it as heads, otherwise tails. Do this n times using a list comprehension of python. This should happen within the function **run\_binom**.
2. Repeat this a specified number of times (the input variable trials specify the number of **trials**). We will perform 1,000 trials.

**Solution:**

```
# Import libraries
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Input variables

# Number of trials
trials = 1000

# Number of independent experiments in each trial
n = 10

# Probability of success for each experiment
p = 0.5

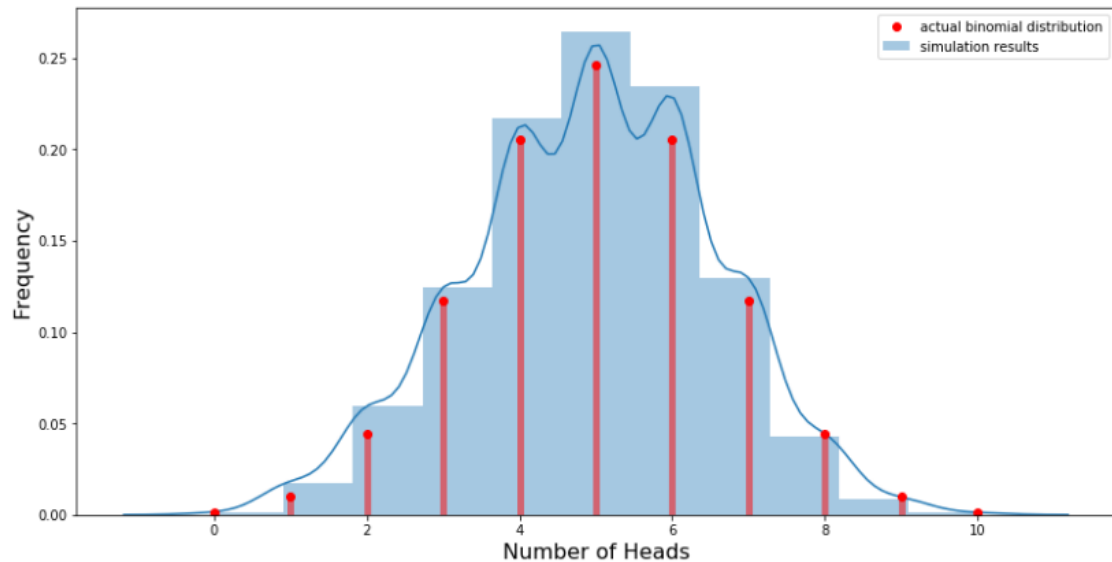
# Function that runs our coin toss trials
def run_binom(trials, n, p):
    head = []
    for i in range(trials):
        toss = [np.random.random() for i in range(n)]
        head.append(len([i for i in toss if i >= 0.5]))
    return head

# Run the function
heads = run_binom(trials, n, p)
```

*# Plot and save the results as the histogram*

```
fig, ax = plt.subplots(figsize=(14,7))
ax = sns.distplot(heads, bins=11, label='simulation results')

ax.set_xlabel("Number of Heads", fontsize=16)
ax.set_ylabel("Frequency", fontsize=16)
```



*# Probability of getting 5 heads*

```
runs = 10000
prob_5 = sum([1 for i in np.random.binomial(n, p, size = runs) if i==5])/runs
print('The probability of 5 heads is: ' + str(prob_5))
```

The probability of 5 heads is: 0.2035

## Normal Distribution (Gaussian Distribution)

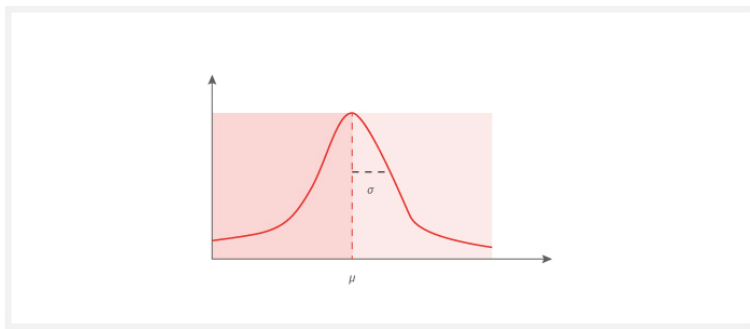
Normal Distribution is one of the most common continuous probability distribution. This type of distribution is important in statistics and is often used to represent random variables whose distribution is not known.

This type of distribution is symmetric, and its mean, median, and mode are equal.

Mathematically, Gaussian Distribution is represented as:

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

Where N stands for Normal, symbol  $\sim$  for distribution, whereas symbol  $\mu$  stands for mean and  $\sigma^2$  stands for the variance.



In the above image, we can view the highest point is located at the mean  $\mu$ , and the spread of the graph can be observed by the standard deviation ' $\sigma$ '.

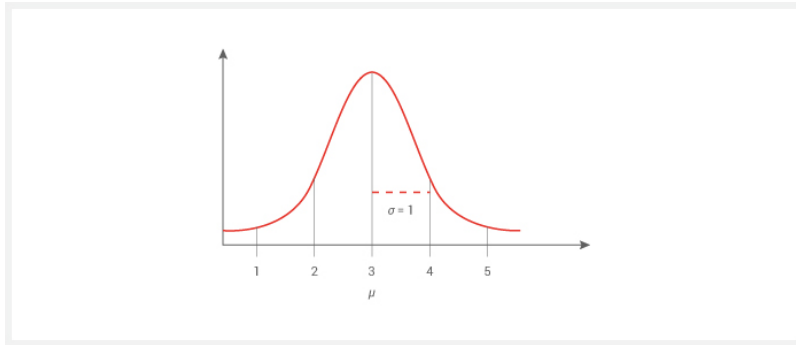
Let us understand this with the easiest example where we can have the random variable X with distribution:

$$X = \{1, 2, 3, 4, 5\}$$

When we take the mean and the standard deviation of the above data set, we get  $\text{mean}(\mu) = 3$  and  $\text{standard deviation}(\sigma) = 1$ .

When we plot it, we get a few distributions like this mentioned below:



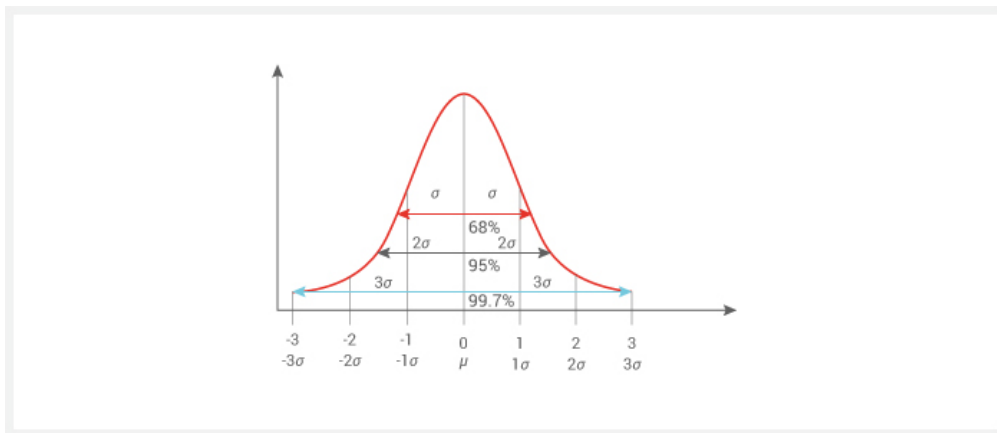


This Bell curve specifies **the Gaussian distribution**.

## 1.1 Empirical Formula

The empirical rule states that for the Normal Distribution, nearly all of the data will fall within three range of standard deviations of a mean. The empirical rule can be understood through the following:

- 68% of the data falls within the 1st standard deviation from the mean.
- 95% fall within two standard deviations.
- 99.7% fall within three standard deviations.



- Approximately 68% of the data falls within one standard deviation of a mean. In mathematical notation, this is represented as  $\mu \pm 1\sigma$
- About 95% of the data falls within 2 standard deviations of the mean (i.e., between the mean  $- 2$  times the standard deviation, & mean  $+ 2$  times the standard deviation). The mathematical notation for this is:  $\mu \pm 2\sigma$
- Approximately 99.7% of the data lies in 3 standard deviation of a mean (i.e., between the mean  $- 3$  times the standard deviation and the mean  $+ 3$  times the standard deviation).
- The following notation is used to represent:  $\mu \pm 3\sigma$

The Empirical Rule is often used to forecast when obtaining the right data is difficult or impossible to get.

## 1.2 Standard Normal Distribution

Understanding Standardization in the context of statistics. Every distribution can be standardized. Let say if the mean and variance of a variable are  $\mu$  and  $\sigma^2$ , respectively.

Standardization is the process of transforming the distribution to one with a mean of 0 and a standard deviation of 1.

i.e.,  $\sim(\mu, \sigma^2) \rightarrow \sim(0, 1)$

When a Normal Distribution is standardized, a result is called a Standard Normal Distribution.

i.e.,  $N(\mu, \sigma^2) \rightarrow N(0, 1)$

We use the following formula for standardization:

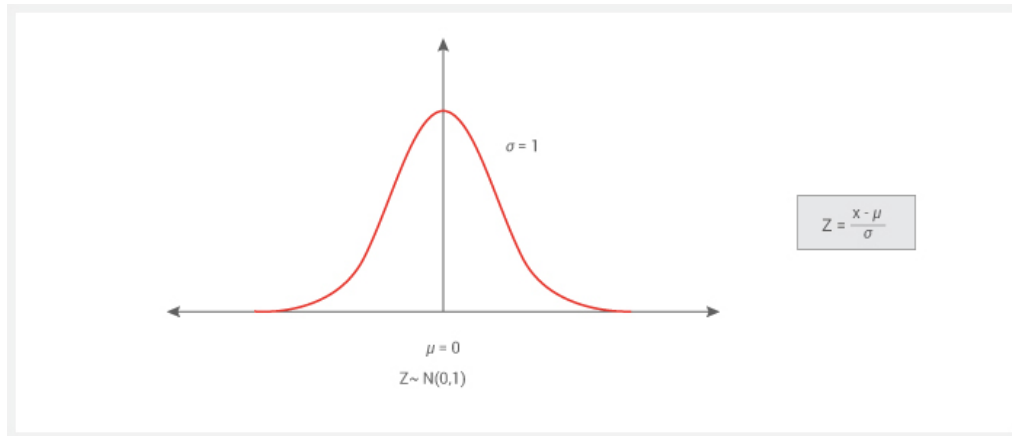
$$Z = \frac{x - \mu}{\sigma}$$

Z - score

Where  $x$  is a data element,  $\mu$  is mean & ' $\sigma$ ' is the standard deviation, and  $Z$  is used to denote standardization, and  $Z$  is known as the z-score.

With the help of  $Z$  scores, we can come to know how far a value is from the mean. When you standardize a random variable, its  $\mu$  becomes 0, and its standard deviation becomes 1.

If the  $Z$  score of  $x$  is 0, then the value of  $x$  is equal to the mean.

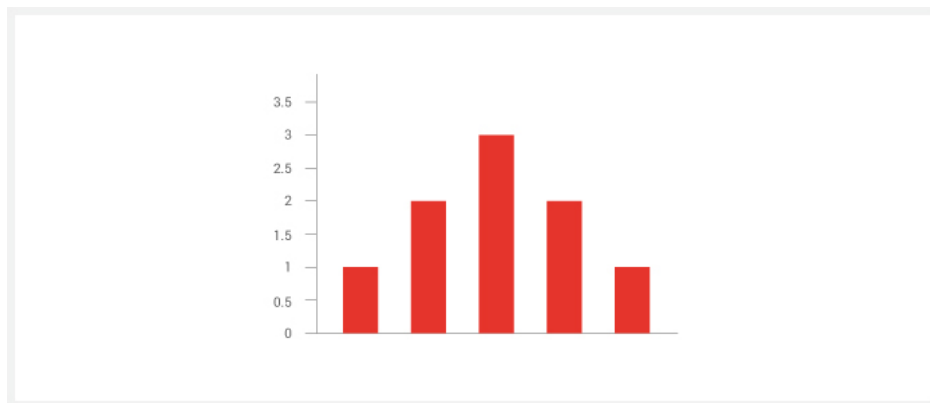


Let us understand the steps in Standardization with the help of a simple example.

Suppose we have a dataset with elements

$$X = \{1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 5\}$$

And uniformly distributed as:



We get mean as 3, variance = 1.49 & standard deviation as 1.22 i.e.,  $N \sim (3, 1.49)$ .

Now we will subtract the mean from every the data points, that is,  $x - \mu$ .

We will get a new data set mentioned below:

$$X1 = \{-2, -1, -1, 0, 0, 1, 1, 1, 2, 2\}$$

$\mu$  as 0, but the variance and std dev still as 1.49 and 1.22 respectively

i.e.,  $N \sim (0, 1.49)$

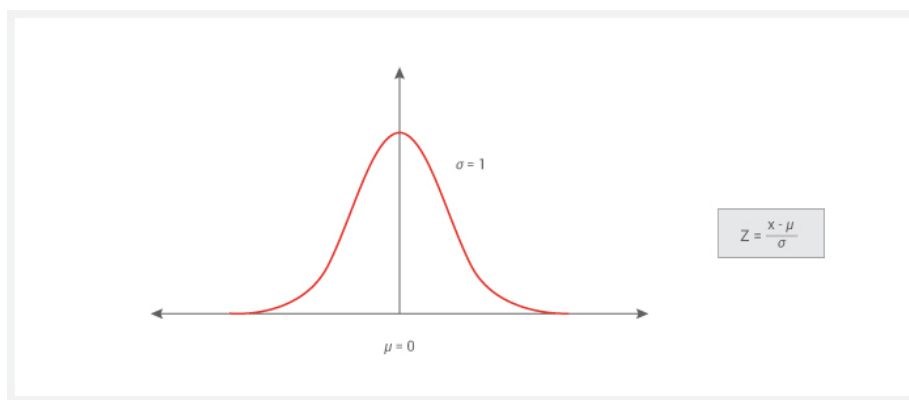
So in the next step of standardization, dividing all data points by the standard deviation, i.e.,  $(x - \mu) / \sigma$

Dividing each datapoint by 1.22(standard deviation) we get a new data set as :

$X_2 = \{-1.6, -0.82, 0, 0.82, 0, 0.82, 0.82, 0 \text{ and } 1.63.\}$

Now if we calculate the mean as 1 i.e.,  $N \sim (0, 1)$

Plotting it on a graph :

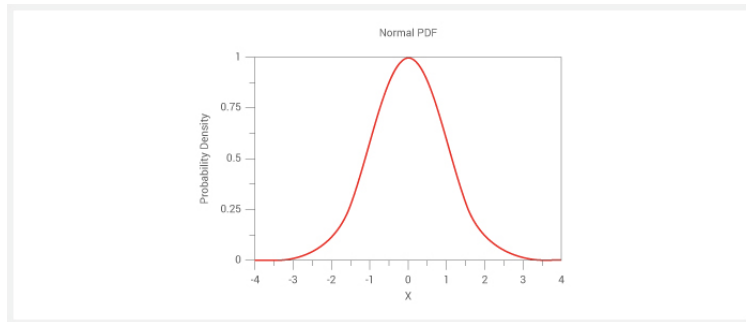


Using this standardized normal distribution makes inferences & predictions much easier.

### 1.3 Probability Density Function and Mass Function

Probability density function and Probability mass function defines a Probability Distribution for a random variable.

If we know the variance and the mean of the dataset, then we can calculate the PDF and PMF. PDF and PMF tell how well data is distributed around mean and standard deviation within a particular curve.

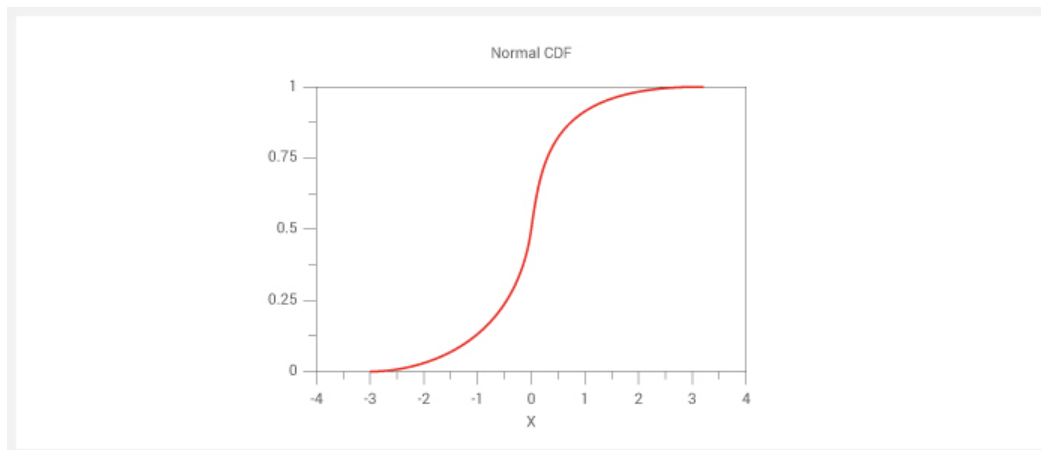


## 1.4 Cumulative Density Function

The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables.

The cumulative frequency is the sum of the frequencies. Cumulative frequency starts at the frequency of the 1st brand, and then we add the 2nd, then 3rd, and so on until it finishes 100%.

For all kinds of random variables (discrete, continuous, and mixed), CDF can be defined.



## Examples of Normal Distribution:

**Question:** An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed. What is a probability that an acme light bulb will last at most 365 days?

**Solution:** Given a mean score of 300 Days & A standard deviation of fifty days, we want to find the cumulative probability that bulb Life is higher than equal to 365 days. Thus, we know the following:

- The value of the standard random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is  $P(X \leq 365)$  equal to 0.90. That means 90% of the chance that a light bulb will burn-out within 365 days.

**Question:** Suppose that the scores on an IQ test are normally distributed(bell-shaped). If the test has a mean is 100 and a standard deviation( $\sigma^2$ ) of 10, what is the probability that a particular person who takes the test will score between 90 and 110.?

**Solution:** Here, we came to know the probability that the test score comes between 90 and 110. The "trick" to solving this problem is the following:

$$\begin{aligned} P(90 < X < 110) \\ = P(X < 110) - P(X < 90) \end{aligned}$$

- Here, the value of the normal random variable is 110, the mean is 100, and the standard deviation = 10. Where  $P(X < 110)$  is 0.84.
- To compute  $P(X < 90)$ , The value of the normal random variable is 90, the mean is 100, and the standard deviation = 10.  $P(X < 90)$  is 0.16.

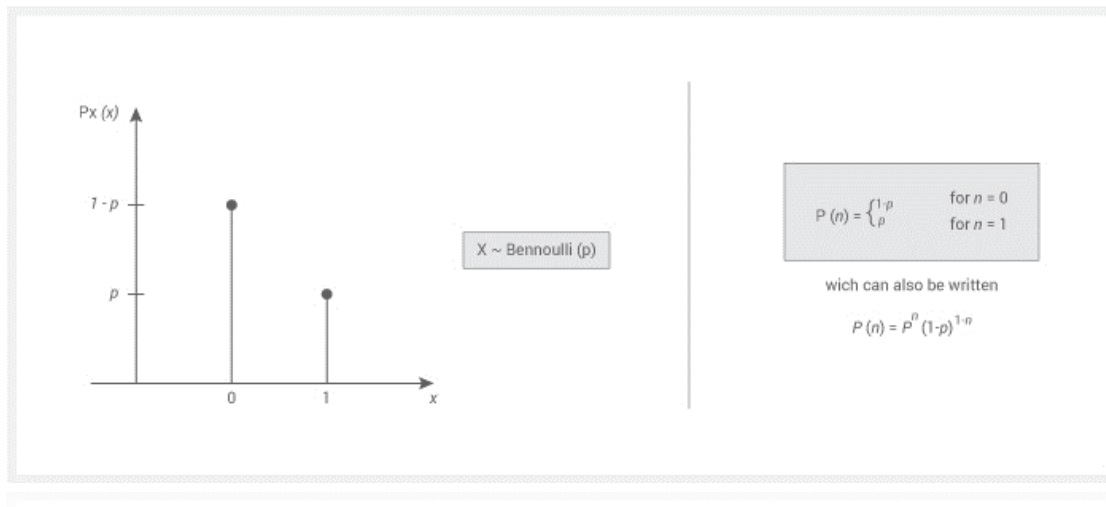
**About 68% of the test scores will lie between 90 and 110.**

### **Bernoulli Distribution**

Bernoulli distribution is a discrete probability distribution of a random variable that has only two outcomes., namely 1 (success) and 0 (failure). where  $n = 1$  occurs with probability  $p$  and  $n = 0$  (usually called a "failure") occurs with probability  $q = 1 - p$ ,

where  $0 < p < 1$ .

Therefore, the probability density function(PDF) & the graph for Bernoulli's Distribution is shown in the figure below:



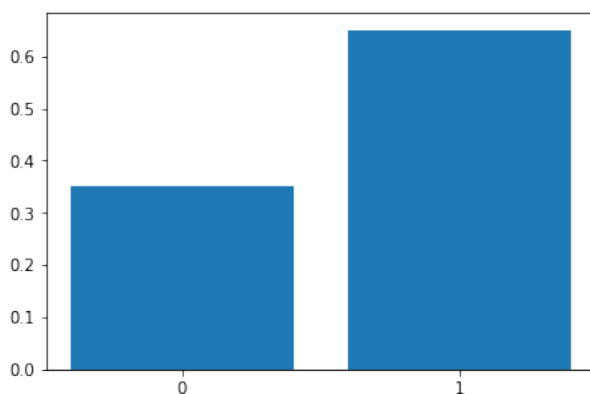
In the above diagram, 1 refers to 'success' & 0 refers to the failure. The head and tail distribution in tossing a coin is an example of Bernoulli's Distribution with  $p = q = \frac{1}{2}$ .

For example, probability (p) of scoring a goal in the last 10 minutes is 0.35 (success); the probability of not scoring a goal in the previous 10 minutes (failure) is  $1 - p = 0.65$ .

Plotting Bernoulli distribution with probability for  $p = 0.65$

```
pyplot.bar(['0', '1'], [0.35, 0.65])
```

<BarContainer object of 2 artists>



## Uniform Distribution

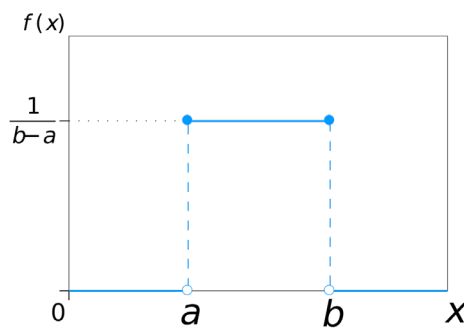
The probability distribution function of the continuous uniform distribution:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$



Since the area under the curve must be equal to 1, and the length of the interval determines the height of the curve, the following figure shows a uniform distribution (a,b).

Note that since the area needs to be 1. The height is set to  $1/(b-a)$ .



#### EXAMPLE:

- If we roll a die(numbered 1 to 8), then the probability of getting 1 is one out of 8.
- Similarly, the probability of getting 2 to 6 is  $1/6$ . There is an equal chance to get each of 8 results (outcomes).

#### Poisson distribution:

**Poisson Distribution** can be used to find the probability of several events in a time period.

#### Conditions for Poisson Distribution:

- Here the events can occur independently.
- An event can occur any number of times.
- The rate of occurrence is constant; i.e., the rate does not change based on time.

#### Example:

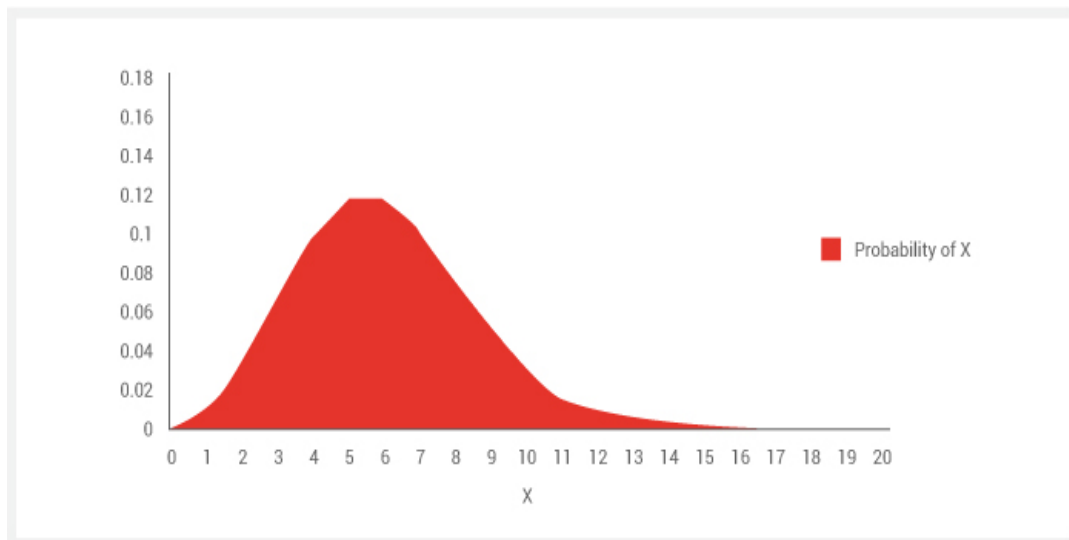
A specific fast-food restaurant gets an average of three visitors to the drive per minute. This is just an average. However, the actual amount can vary.

Suppose we conduct the Poisson experiment, where the average number of success in a given region is  $\mu$ .

Then, the Poisson probability is:

$$P(x; \mu) = \frac{e^{-\mu} (\mu^x)}{x!}$$

The graph of Poisson Distribution:



### Examples:

**Question:** A life insurance salesman sells on the average '3' life insurance policies per week. Use Poisson's law to calculate the probability

- In a given week he will sell how many policies
- In a given week, he will sell '2' or more policies but not more than five policies.
- Assuming that per week, there are '5' working days, what is the probability that on a given day, he will sell one policy?

**Solution:** Here,  $\mu = 3$

(a) "Some policies" means "1 or more policies". We can work this out by computing to find 1 - "zero policies" probability:

$$P(X > 0) = 1 - P(x_0)$$

$$\text{Now } P(X) = \frac{e^{-\mu} \mu^x}{x!} \text{ so } P(x_0) = \frac{e^{-3} 3^0}{0!} = 4.9787 \times 10^{-2}$$

Therefore, the probability of 1 or more than 1 policies:

$$P(X \geq 0)$$

$$= 1 - 4.9787 \times 10^{-2} \text{ because of } 1 - P(x_0)$$

$$= 0.95021$$

(b) The probability of selling 2 or more policies but less than 5 policies are:

$$P(2 \leq X < 5)$$

$$= P(x_2) + P(x_3) + P(x_4)$$

$$= \frac{e^{-3} 3^2}{2!} + \frac{e^{-3} 3^3}{3!} + \frac{e^{-3} 3^4}{4!}$$

$$= 0.61611$$

(c) The average number of policies sold per day is  $3/5 = 0.6$ .

On a given day,

$$P(X) = \frac{e^{-0.6} (0.6)^1}{1!} = 0.32929$$

Q. Aluminium alloy sheets(20 sheets) were examined for surface flaws. The frequency of sheets, along with the number of flaws, is as follows:

**Number of flaws    Frequency**

'0'	'4'
'1'	'3'
'2'	'5'
'3'	'2'
'4'	'4'
'5'	'1'
'6'	'1'

When chosen at random, what is the probability of finding a sheet that contains three or more surface flaws?

Solution: The total number of flaws:

The average number of flaws for twenty sheets is given by:

$$(0 \times 4) + (1 \times 3) + (2 \times 5) + (3 \times 2) + (4 \times 3) + (5 \times 1) + (6 \times 1) = 44$$

An average number of flaws for twenty sheets is given by:

$$\mu = 44/20 = 2.3$$

The required probability is:

$$\text{Probability} = P(X \geq 3)$$

$$= 1 - (P(x=0) + P(x=1) + P(x=2))$$

$$= 1 - \left( \frac{e^{-2.3} 2.3^0}{0!} + \frac{e^{-2.3} 2.3^1}{1!} + \frac{e^{-2.3} 2.3^2}{2!} \right)$$

$$= 0.40396$$

### Question:

If electrical power failures occur according to the Poisson distribution with an average of '3' failures every twenty weeks, calculate the probability of more than 1 failure during a particular week.

Solution:

Per week, the average number of failures  $\mu = 3/20 = 0.15$

We need to include the probabilities for zero failures plus "1 failure".

$$P(x_0) + P(x_1) = \frac{e^{-0.15} 0.15^0}{0!} + \frac{e^{-0.15} 0.15^1}{1!} = 0.98981$$

### Z Stats

(a) It allows us to calculate the **score probability**, which occurs within our normal distribution.

(b) It enables us to compare two **scores** of different normal distributions.

The basic z score formula is:

$$z = (x - \mu) / \sigma$$

When you have multiple samples and their sample means (the standard error), you will use the following z-score formula:

$$z = (x - \mu) / (\sigma / \sqrt{n})$$

### Calculating z-scores

#### Question.

The grades on a physics midterm at Covington are roughly symmetric with  $\mu = 72$  and  $\sigma = 2.0$ .

- Stephanie scored 74 on the exam.
- Find the z-score for Stephanie's exam grade. Round to two decimal places.

#### Solution:

z-score is defined as the number of the standard deviations a particular point is away from the mean.

$$z = (74 - 72) / 2.0$$

$$z \approx 1.00$$

The z-score is 1.00. In other words, Stephanie's score was 1.00 standard deviation above the mean.

#### Question:

The grades on a math midterm at Gardner Bullis are roughly symmetric with  $\mu = 76$  and  $\sigma = 4.5$ .

- Daniel scored 64 on the exam.
- Find the z-score for Daniel's exam grade. Round to two decimal places.

#### Solution:

z-score is defined as the number of the standard deviations a particular point is away from the mean.

$$z = (64 - 76) / 4.5$$

$$z \approx -2.67$$

The z-score is -2.67. In other words, Daniel's score was 2.67 standard deviations below the mean.

**Question:** Molly earned a score of 940 on a national achievement test. The mean test score was 850, along with the standard deviation of 100. What proportion of students had a higher score than Molly?

(Assume that test scores are normally distributed.)

(A) 0.10

(B) 0.18

- (C) 1.50
- (D) 5.82
- (E) 2.90

**Solution:** The correct answer is B. We assume that test scores are normally distributed. Given an assumption of normality, It involves the following steps:

- We transform Molly's test score into the z-score using the z-score transformation equation.

$$z = (X - \mu) / \sigma = (940 - 850) / 100 = 0.90$$

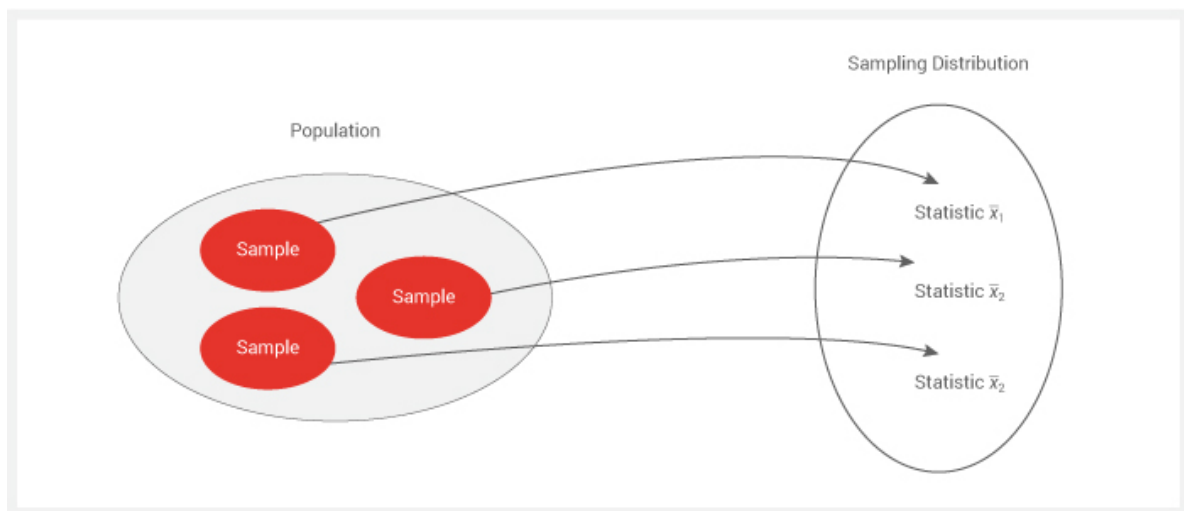
- Then, using the standard normal distribution table, we find the cumulative probability associated with the z-score. In this case, we find  $P(Z < 0.90) = 0.8159$ .
- Therefore, the  $P(z > 0.90)$

$$= 1 - P(z < 0.90)$$

$$= 1 - 0.8159 = 0.1841.$$

Thus, we estimate that 18.41% of the students tested had a higher score than Molly.

## Central Limit Theorem



After fetching different samples, which are enough in numbers, we can then calculate the mean of each sample and then plot the various distributions.

Also, if we take the average of the sample mean, then the result will be equal to the actual population mean & the standard deviation equals  $\sigma/\sqrt{n}$ .

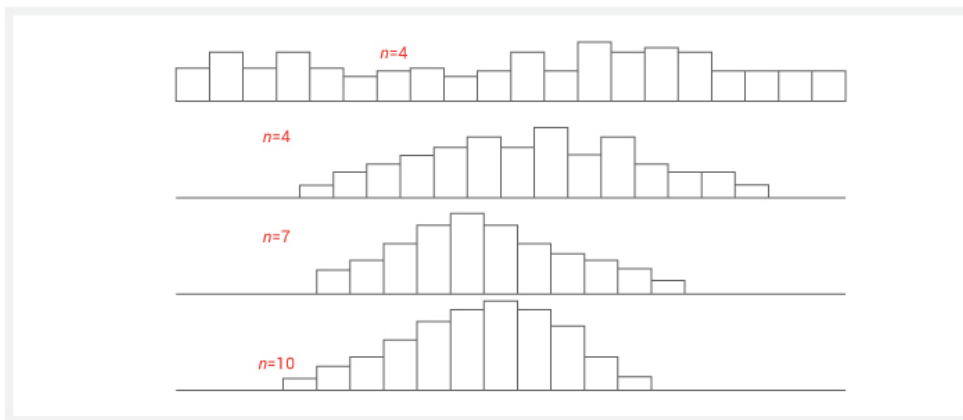
Where,

‘ $\sigma$ ’ is the population of std deviation

$n$  = the sample size(i.e., # of observations in our sample)

Important points while applying the Central Limit Theorem:

- The distribution of the original(population) dataset does not matter. It could be normal, uniform, binomial, etc.
- The distribution of the sample means would be Normal Distribution
- Larger the number of samples taken from the population, the closer to a Normal Distribution the sample means will be.



- The samples extracted should be more significant than 30 observations.
- The sample mean average extracted will be approximately equal to the mean of the population, and its variance would be similar to the original variance, which is divided by the sample size, i.e., ‘ $n$ ’.

## Mathematical Explanation of Central limit Theorem

The central limit theorem states that the mean ( $\bar{X}$ ) follows approximately the Normal distribution with mean  $\mu$  and standard deviation  $\frac{\sigma}{\sqrt{n}}$ , where  $\mu$  and  $\sigma$  are the mean and standard deviation of the population.

To summarize:  $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

To transform  $\bar{X}$  into  $z$  we use:  $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

**Example:** Let  $X$  be a random variable along with  $\mu = 10$  and  $\sigma = 4$ . A sample of size 100. Find the probability that the sample mean ‘ $\mu$ ’ of 100 number of observations is not more than (less than) 9.

We write

$$P(\bar{X} < 9) = P(z < \frac{9-10}{\frac{4}{\sqrt{100}}}) = P(z < -2.5) = 0.0062$$

Similarly, the central limit theorem states that the sum T follows approximately a normal distribution,

$$T \sim N(n\mu, \sqrt{n}\sigma)$$

Where  $\mu$  and  $\sigma$  are the mean and standard deviation of a population. To transform T into z, we use:

$$z = \frac{T-n\mu}{\sqrt{n}\sigma}$$

**Example:** Let X be a random variable along with ' $\mu$ ' = 10 and  $\sigma$  = 4. A sample of size 100 is taken from the population. Find the probability that the sum of these 100 number of observations is less than 900.

We write

$$P(T < 900)$$

$$= P(z < \frac{900-100(10)}{\sqrt{100}(4)})$$

$$= P(z < -2.5)$$

$$= 0.0062 \text{ (from the table of standard normal probabilities)}$$

## Applications of the central limit theorem

**Question:** A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be carried through the elevator. Experience has shown that the weight of some boxes of this type of cargo follows a distribution with mean  $\mu$  = 205 pounds & the standard deviation ' $\sigma$ ' = 15 pounds. Based on this information, what is a probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

**Solution:** We are given  $n = 49$ ,  $\mu = 205$ ,  $\sigma = 15$ .

The elevator can transport up to 9,800 Pounds. Therefore, these 49 boxes will be safely transported if they weigh in total, not more than (less than) 9800 pounds.

The probability that the total weight of 49 boxes is not more than (less than) 9800 pounds is

$$P(T < 9800)$$

$$= P(z < \frac{9800-49(205)}{\sqrt{49}15})$$

$$= P(z < -2.33)$$



$$= 1 - 0.9901$$

$$= 0.0099.$$

**Question:** It is known that the number of tickets purchased by a student standing in line at the ticket counter to buy the tickets for the Football match of U.C.L.A. against USC follows a distribution that has a mean  $\mu = 2.4$  & the standard deviation ' $\sigma$ ' = 2.0.

Suppose 100 eager students are standing in line to purchase the tickets. If only 250 tickets remain unbooked, what is a Probability that all 100 students will be able to buy the tickets they desire?

**Solution:** We are given that  $\mu = 2.4$ ,  $\sigma = 2$ ,  $n = 100$ .

There are 250 tickets available, so a hundred students will be able to purchase the tickets they want if all together ask for not more than 250 tickets. Probability for that is

$$\begin{aligned} &P(T < 250) \\ &= P\left(z < \frac{250 - 100(2.4)}{\sqrt{100} \cdot 2}\right) \\ &= P(z < 0.5) \\ &= 0.6915. \end{aligned}$$

**Question:** Suppose that you have a sample of 100 values from a population with mean  $\mu = 500$  and with standard deviation  $\sigma = 80$ .

- What is the probability that the  $\mu$  will be in the interval (490, 510)?
- Give the interval that covers the average 95 percent of the distribution of the sample mean.

**Solution:** We are given  $\mu = 500$ ,  $\sigma = 80$ ,  $n = 100$ .

$$\begin{aligned} \text{a. } &P(490 < \bar{x} < 510) \\ &= P\left(\frac{490 - 500}{\frac{80}{\sqrt{100}}} < z < \frac{510 - 500}{\frac{80}{\sqrt{100}}}\right) \\ &= P(-1.25 < z < 1.25) \\ &= 0.8944 - (1 - 0.8944) \\ &= 0.7888. \end{aligned}$$

$$\begin{aligned} \text{b. } &\pm 1.96 = \frac{\bar{x} - 500}{\frac{80}{\sqrt{100}}} \\ &\bar{x} = 484.32, \bar{x} = 515.68. \end{aligned}$$

Therefore  $P(484.32 < \bar{x} < 515.68)$

= 0.95.

**Question:** The amount of regular unleaded gasoline purchased every week at a gas station near UCLA follows the normal distribution with mean 50,000 gallons and a standard deviation of 10,000 gallons. The starting supply of gasoline is 74,000 Gallons, and there is a scheduled weekly delivery of 47000 gallons.

- Find the probability that, after 11 weeks, a supply of gasoline will be below 20000 gallons.
- How much should the weekly delivery be so that after 11 weeks, the probability that the supply is only 0.5% below 20000 gallons is?

Solution:

Given: ' $\mu$ ' = 50000, ' $\sigma$ ' = 10000,  $n$  = 11. The starting supply is 74000 gallons. 47000 gallons is the weekly delivery. Therefore, the total supply for the eleven weeks is  $74000 + 11 \times 47000 = 591000$  Gallons.

- The supply will be below 20000 gallons if the gasoline purchased in these 11 weeks is more than  $591000 - 20000 = 571000$  Gallons

Therefore, we need to find

$$P(T > 571000) = P\left(z > \frac{571000 - 11(50000)}{\sqrt{11}10000}\right)$$

$$= P(z > 0.63)$$

$$= 1 - 0.7357$$

$$= 0.2643.$$

- Let A be the unknown scheduled delivery.

Now, the total gasoline purchased must be higher than  $74000 + 11 \times A - 20000$ .

We need this with probability 0.5% or Probability  $P(T > 74000 + 11A - 20000) = 0.005$ .

The z value is 2.57566.

$$\text{So, } 2.575 = \frac{74000 + 11A - 20000 - 11(50000)}{\sqrt{11}10000}$$

$$A = 52854.88$$

The weekly delivery must be 52854.8 gallons.

