Parul® University

Data Visualization and Data Analytics

Prof. Khushbu Chauhan, Assistant Professor
Information Technology

CHAPTER 2

Descriptive Statistics

Outlines

- Descriptive statistics: Population and sample
- Types of Data
- Measurement levels
- Representation of categorical variables
- Measures of central tendency (Mean, Median, Mode)
- Skewness
- Variancce, Standard deviation
- Confident of variation , Covariance, Correlation
- Histogram Analysis
- Distribution & its types, Central limit theorem

Introduction

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

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of two basic categories of measures: measures of central tendency and measures of variability (or spread).
- Measures of central tendency describe the center of a data set.
- Measures of variability or spread describe the dispersion of data within the set.

Population

- A population is the entire group that you want to draw conclusions about. A sample is a specific group that you will collect data. The size of the sample is always less than the total size of the population. In research, a population doesn't always refer to people
- It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a parameter. For example, All people living in India indicate the population of India.

- There are different types of population. They are:
- Finite Population
- Infinite Population
- Existent Population
- Hypothetical Population

Sample

- It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic. Sampling is the process of selecting the sample from the population. For example, some people living in India is the sample of the population.
- Basically, there are two types of sampling. They are:
- Probability sampling
- Non-probability sampling

Sampling Techniques

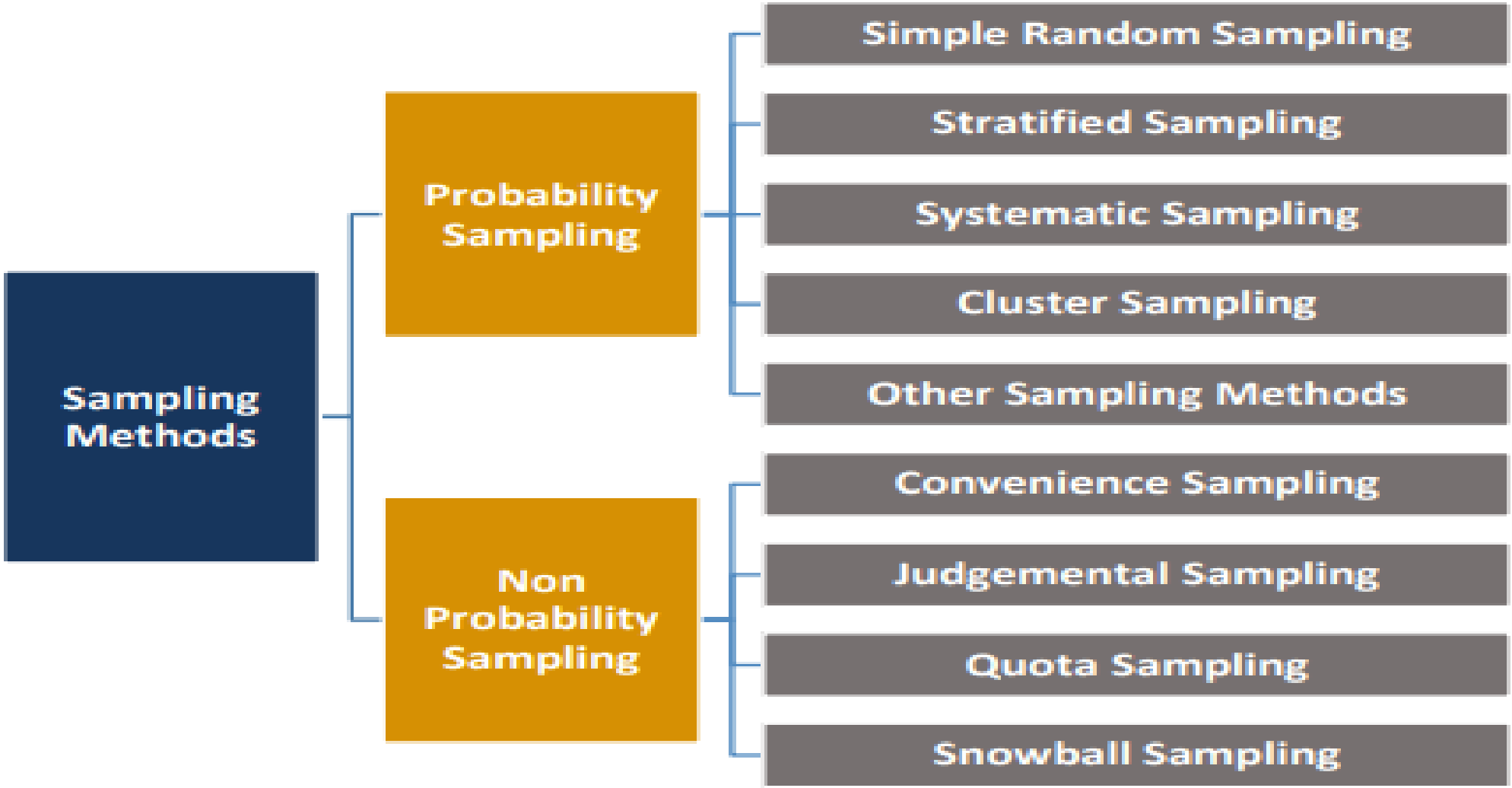


Figure 2: Types of Sampling

Population and sample example

- All the people who have the ID proofs is the population and a group of people who only have voter id with them is the sample.
- All the students in the class are population whereas the top 10 students in the class are the sample.
- All the members of the parliament are population and the female candidates present there is the sample.

Comparison	Population	Sample
Meaning	Collection of all the units or elements that possess common characteristics	A subgroup of the members of the population
Includes	Each and every element of a group	Only includes a handful of units of population
Characteristics	Parameter	Statistic
Data Collection	Complete enumeration or census	Sampling or sample survey
Focus on	Identification of the characteristics	Making inferences about the population

Measurement Levels

- **Levels of measurement,** also called scales of measurement, tell you how precisely variables are recorded. In scientific research, a variable is anything that can take on different values across your data set (e.g., height or test scores).
- There are 4 levels of measurement:
- **Nominal:** the data can only be categorized
- **Ordinal:** the data can be categorized and ranked
- **Interval:** the data can be categorized, ranked, and evenly spaced
- **Ratio:** the data can be categorized, ranked, evenly spaced, and has a natural zero.

Categorical variables

- In statistics, a categorical variable (also called a qualitative variable) is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property.
- A categorical variable that can take on exactly two values is termed a *binary variable* or a dichotomous variable; an important special case is the Bernoulli variable.

Representation of categorical variable

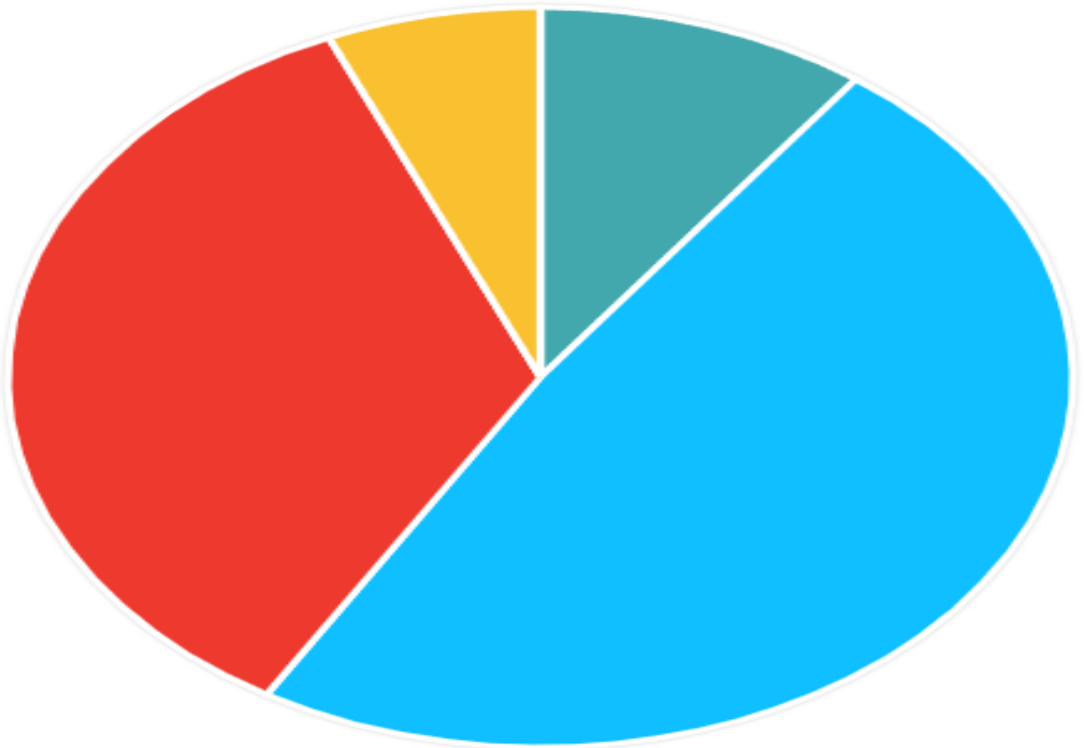
- Frequency tables, pie charts, and bar charts are the most appropriate graphical displays for categorical variables

Frequency Table

A table containing the counts of how often each category occurs.

Diagnosis	Count	Percent
Depression	40835	48.5%
Anxiety	29388	34.9%
OCD	5465	6.5%
Abuse	8513	10.1%
Total	84201	100.0%

Pie Chart of Diagnosis

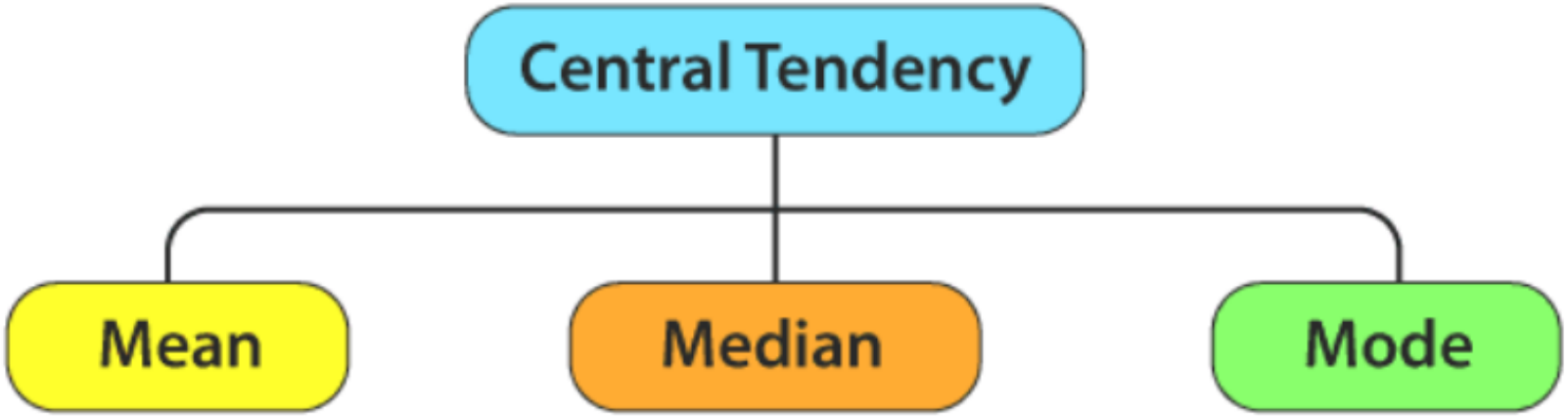


Category

Depression (48.5%)
Anxiety (34.9%)
OCD (6.5%)
Abuse (10.1%)

Measures 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.



Mean

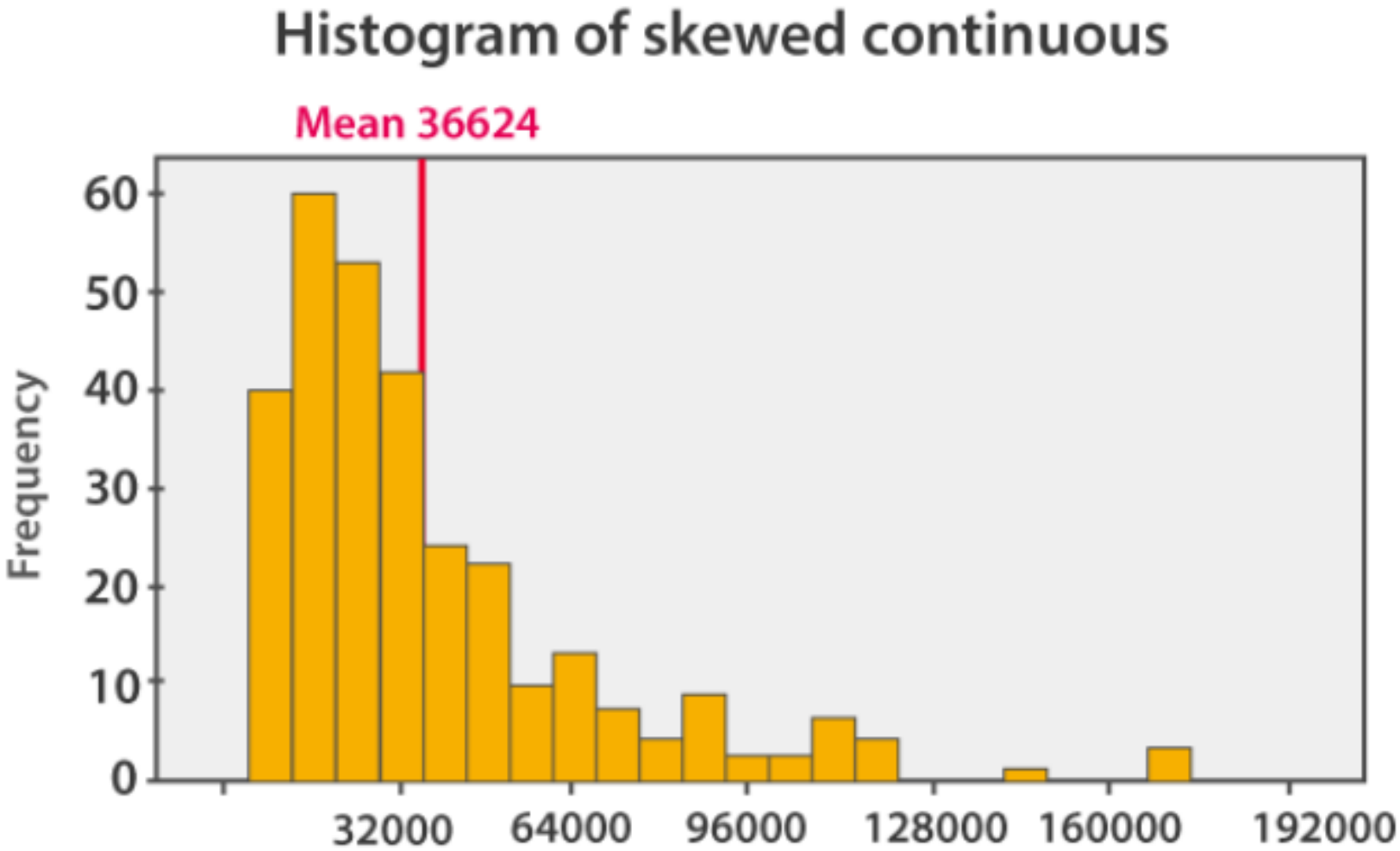
- The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered the arithmetic mean.
- It is observed that if all the values in the dataset are the same, then all geometric, arithmetic, and harmonic mean values are the same. If there is variability in the data, then the mean value differs. Calculating the mean value is completely easy. The formula to calculate the mean value is given as

$$\frac{x_1+x_2+..+x_n}{n}$$

$$\exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln a_i\right)$$

Mean example

	A	B
1	Observation	Cost per order
2	x1	$2,700.00
3	x2	$19,250.00
4	x3	$15,937.50
5	x4	$18,150.00
93	x92	$74,375.00
94	x93	$72,250.00
95	x94	$6,562.50
96	Sum of cost/order	$2,471,760.00
97	Number of observations	94
98		
99	Mean cost/order	$26,295.32
100		
101	Excel AVERAGE function	$26,295.32


Histogram of skewed continuous
Mean 36624
Frequency
60
50
40
30
20
10
0
32000
64000
96000
128000
160000
192000

Median

- Median is the middle value of the dataset in which the dataset is arranged in ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.
- Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2

Median odd
23
21
18
16
15
13
12
10
9
7
6
5
2

- Now, consider another example with an even number of observations that are arranged in descending order – 40, 38, 35, 33, 32, 30, 29, 27, 26, 24, 23, 22, 19, and 17.
- the two middle values obtained are 27 and 29.
- Now, find out the mean value for these two numbers.
- i.e.,(27+29)/2 =28
- Therefore, the median for the given data distribution is 28.

Mode

- The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.
- Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5

Mode
5
5
5
4
4
3
2
2
1

Measures of variability

- **Variability** describes how far apart data points lie from each other and from the center of a distribution. Along with measures of central tendency, measures of variability give you descriptive statistics that summarize your data.
- Variability is also referred to as spread, scatter, or dispersion. It is most commonly measured with the following:

- **Range:** the difference between the highest and lowest values
- **Interquartile range:** the range of the middle half of a distribution
- **Standard deviation:** average distance from the mean
- **Variance:** average of squared distances from the mean
- **Skewness:** Skewness measures the degree and direction of asymmetry
- **Coefficient of variation:** Measure the relative dispersion of data points in a data series around the mean
- **Covariance and correlation:** Measure the joint variability of two random variables and the relationship between them.

Range

- The range tells you the spread of your data from the lowest to the highest value in the distribution. It's the easiest measure of variability to calculate.

Data (minutes)	72	110	134	190	238	287	305	324

The highest value (H) is **324** and the lowest (L) is **72**.

$$R = H - L$$

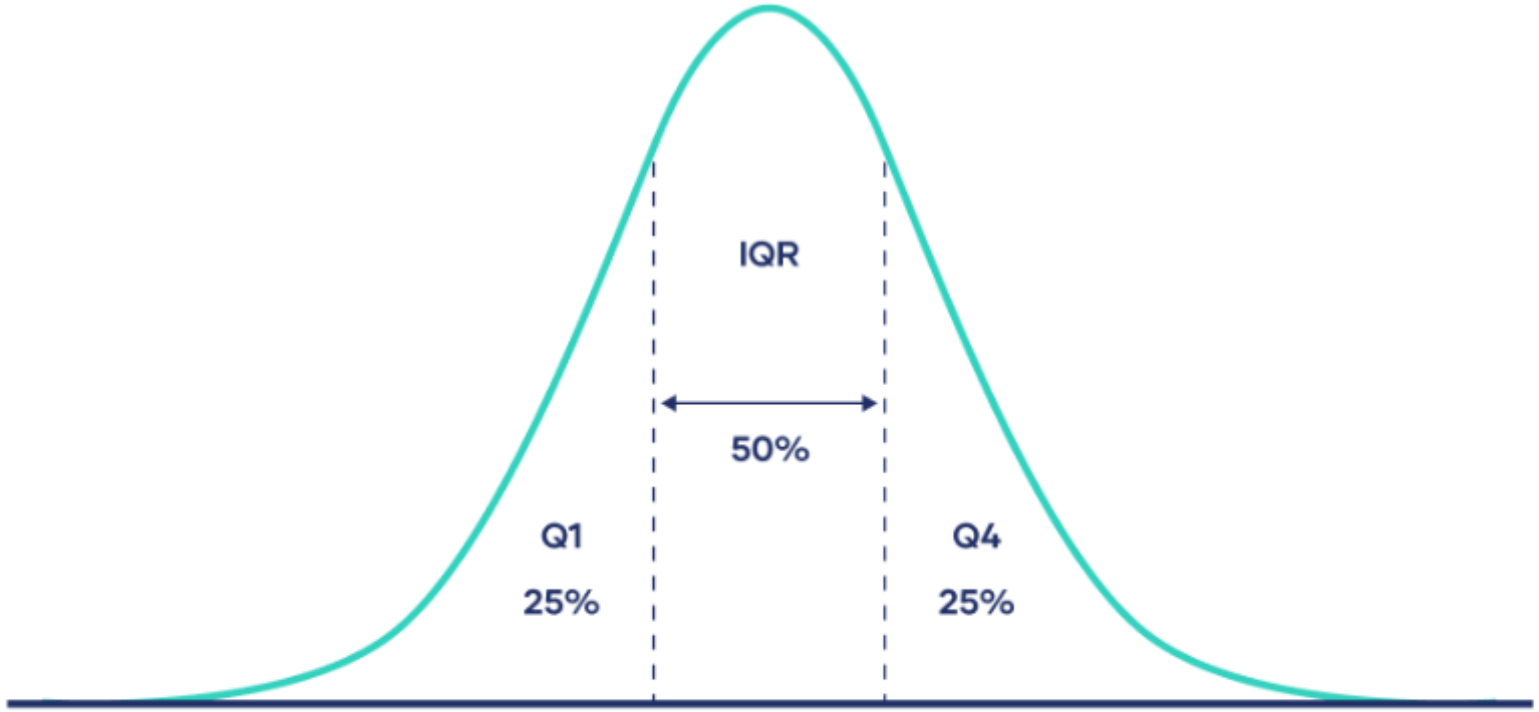
$$R = 324 - 72 = \mathbf{252}$$

The range of your data is **252 minutes**.

Interquartile range

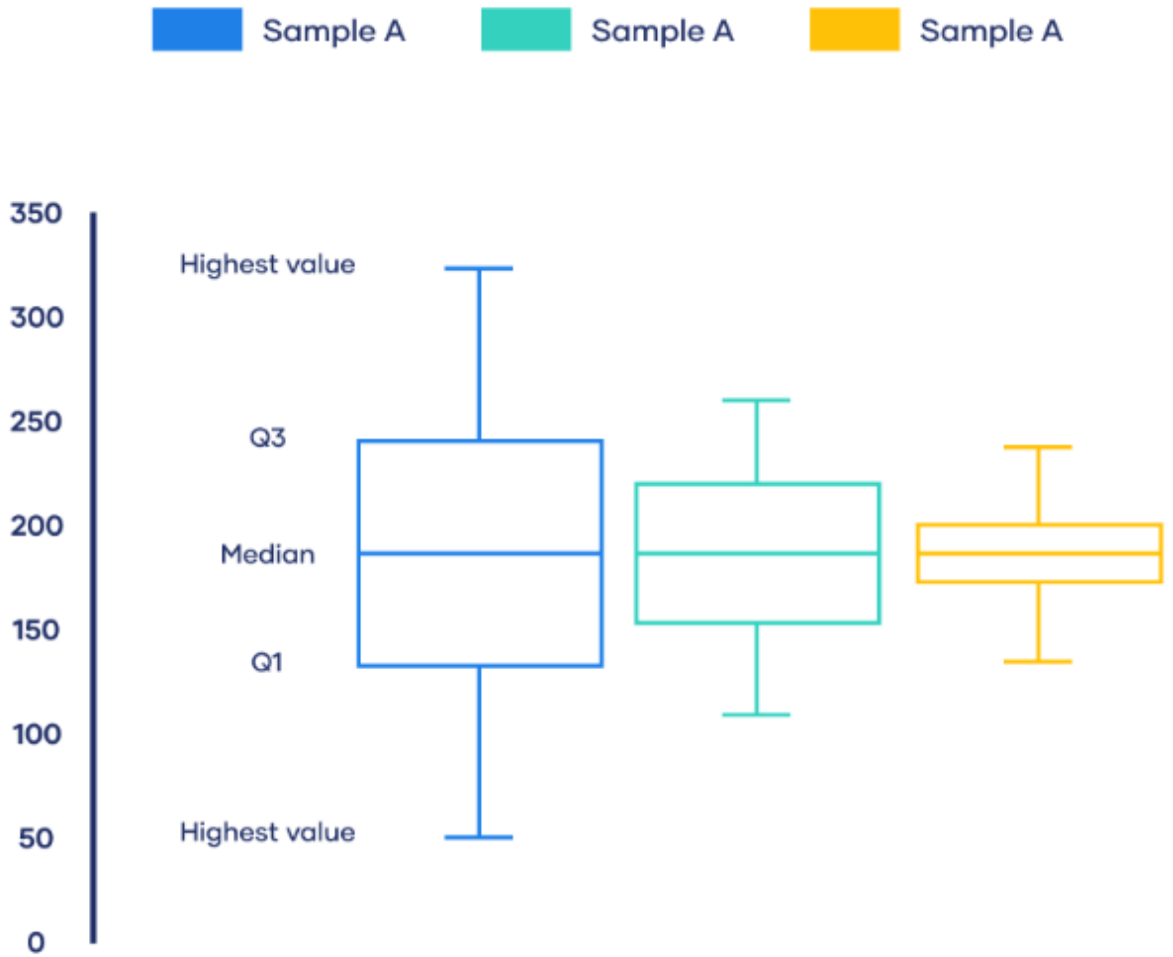
- The interquartile range gives you the spread of the middle of your distribution.
- For any distribution that's ordered from low to high, the interquartile range contains half of the values. While the first quartile (Q1) contains the first 25% of values, the fourth quartile (Q4) contains the last 25% of values.

Interquartile range on a normal distribution



The interquartile range is the third quartile (Q3) minus the first quartile (Q1). This gives us the range of the middle half of a data set.

Average phone use per day in minutes



Standard Deviation

- The standard deviation is the square root of the variance. For a population, the standard deviation is computed as

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

and for samples, it is

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

Variance

- commonly used measure of dispersion is the variance, whose computation depends on all the data.
- The larger the variance, the more the data are spread out from the mean and the more variability one can expect in the observations.
- The formula used for calculating the variance is different for populations and samples.
- The formula for the variance
- of a population is: where
- Xi: the value of the ith item
- N: number of items in the population
- μ is the population means.

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

- Essentially, the variance is the average of the squared deviations of the observations from the mean.
- A significant difference exists between the formulas for computing the variance of a population and that of a sample. The variance of a sample is calculated using the formula

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

- where n is the number of items in the sample and $\bar{x}$ is the sample mean.

Skewness

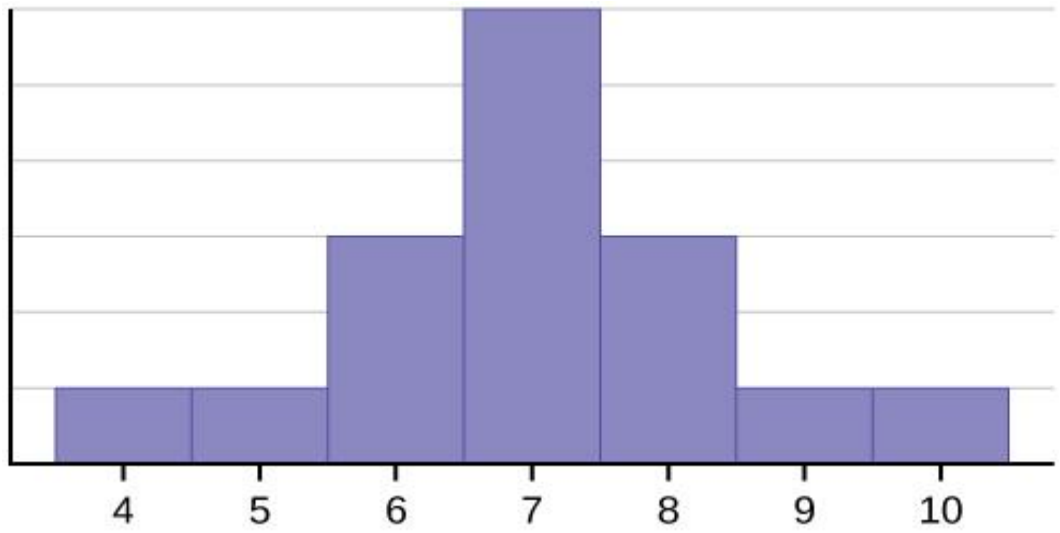
- Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

- Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.

- A normal distribution has a skew of zero, while a lognormal distribution, for example, would exhibit some degree of right-skew.

Skewness and the Mean, Median, Mode

onsider the following data set.
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10



The mean is 6.3, the median is 6.5, and the mode is 7.

Notice that the mean is less than the median, and they are both less than the mode.

If mean>median, positive/right skew
Else, it is a negative/left skew

Therefore, it is **skewed to the right**.

Coefficient of variation

- The coefficient of variation (CV) is the ratio of the standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around the mean. It is generally expressed as a percentage. Without units, it allows for comparison between distributions of values whose scales of measurement are not comparable.

- When we are presented with estimated values, the CV relates the standard deviation of the estimate to the value of this estimate. The lower the value of the coefficient of variation, the more precise the estimate.

Covariance

- In mathematics and statistics, covariance is a measure of the relationship between two random variables.
- The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables.
- However, the metric does not assess the dependency between variables.
- Positive covariance: Indicates that two variables tend to move in the same direction.
- Negative covariance: Reveals that two variables tend to move in inverse directions.

Calculating covariance

- $Cov\,(X,Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n-1}$
- Where:
  - X_i – the values of the X-variable
  - Y_j – the values of the Y-variable
  - $\overline{X}$ – the mean (average) of the X-variable
  - $\overline{Y}$ – the mean (average) of the Y-variable
  - n – the number of data points

Correlation

- A correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other.
- The correlation coefficient is a value that indicates the strength of the relationship between variables. The interpretations of the values are:

  - -1: Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).
  - 0: No correlation. The variables do not have a relationship with each other.
  - 1: Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).

Calculating correlation

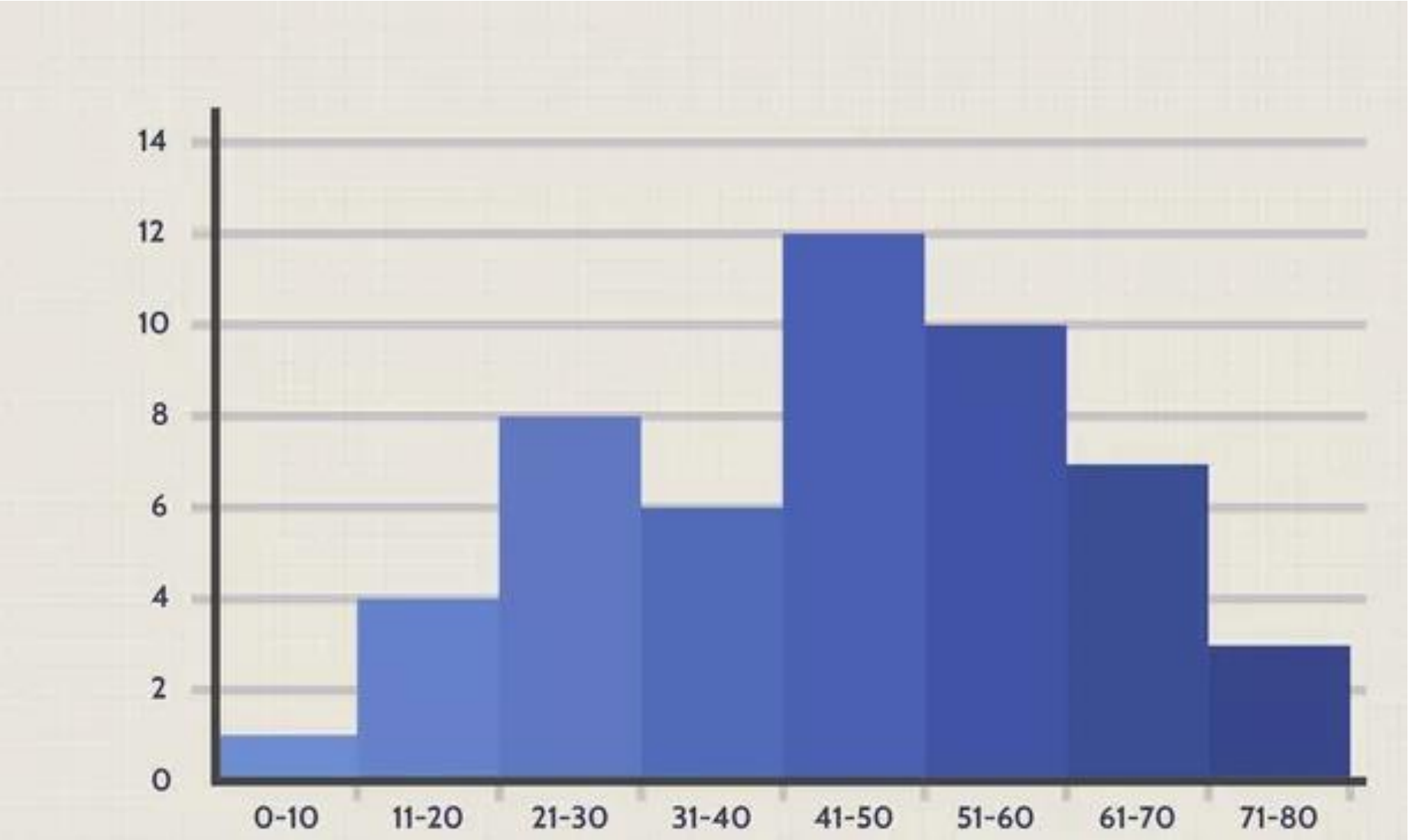
- $r_{xy} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$
- Where:
  - r_{xy} – the correlation coefficient of the linear relationship between the variables x and y
  - X_i – the values of the x-variable in a sample
  - $\overline{X}$ – the mean of the values of the x-variable
  - Y_i – the values of the y-variable in a sample
  - $\overline{Y}$ – the mean of the values of the y-variable

Histogram

- A histogram is a graphical representation that divides a set of data points into user-specified ranges. The histogram, which resembles a bar graph in appearance, condenses a data series into an easily interpreted visual by grouping many data points into logical ranges or bins.
- For example, a census focusing on a country's demographics may use a histogram to illustrate how many individuals are between the ages of 0 and 10, 11 and 20, 21 and 30, 31 and 40, 41 and 50, and so on. This histogram might be like the one below.

- The analyst can modify histograms in a variety of ways. The first step is to adjust the distance between buckets. In the above example, there are 5 buckets with a ten-point spacing. This may be modified to, say, 10 buckets with a 5 interval instead.
- Another thing to think about is how to define the y-axis. The simplest basic label is the frequency of occurrences seen in the data, however, the percentage of total or % of total might also be used.

Histogram Analysis



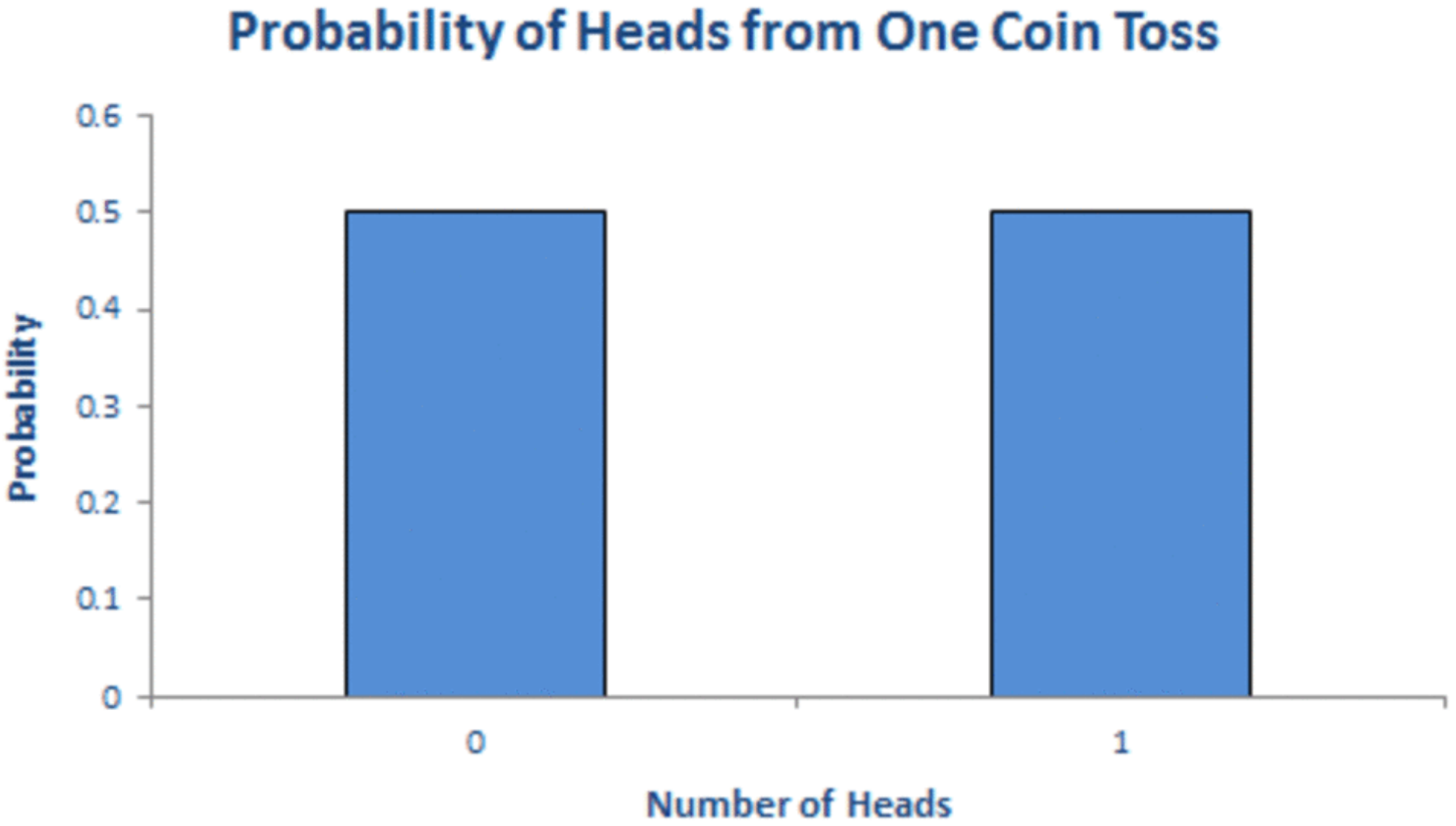
Distribution

- A distribution in statistics is a function that shows the possible values for a variable and how often they occur.

- Types:
  - Bernoulli Distribution
  - Uniform Distribution
  - Binomial Distribution
  - Normal Distribution
  - Poisson Distribution
  - Exponential Distribution

Bernoulli Distribution

- A Bernoulli distribution has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, and the value 0 with failure.
- For example, distribution of toss of a coin.
- Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.
- Note that the probabilities of success and failure need not be equally likely.


Probability of Heads from One Coin Toss
Probability
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
Number of Heads

- The expected value of a random variable X from a Bernoulli distribution is found as follows:
- $E(X) = 1 \times p + 0 \times (1 - p) = p$

- The variance of a random variable from a Bernoulli distribution is:
- $V(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p)$

Uniform Distribution

- Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.
- A variable X is said to be uniformly distributed if the density function is:

$$f(x) = \frac{1}{b-a} \qquad \forall \; -\infty < a \leq x \leq b < \infty$$

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

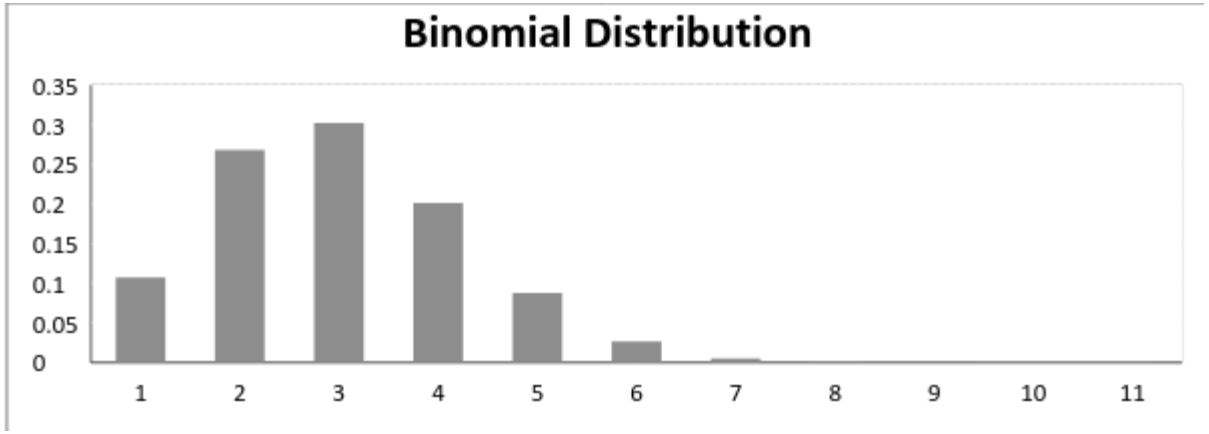
Mean -> $E(X) = \frac{(a+b)}{2}$

Variance -> $V(X) = \frac{(b-a)^2}{12}$

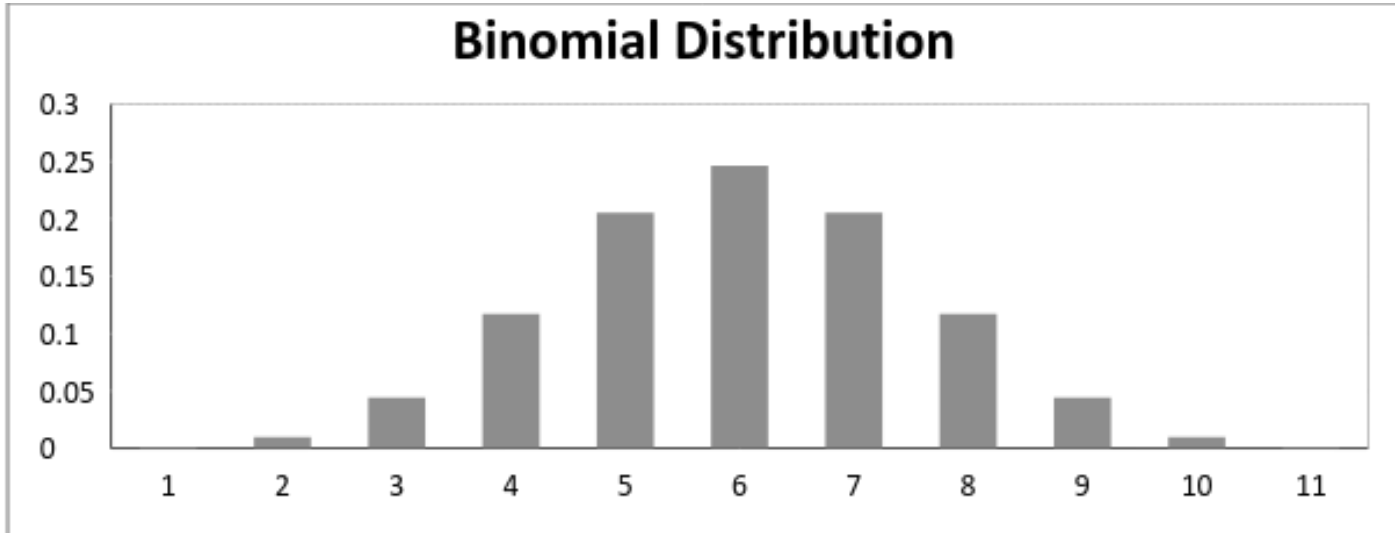
Binomial Distribution

- The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!\ x!} p^x q^{n-x}$$



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



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

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

Mean $\mu = n \times p$

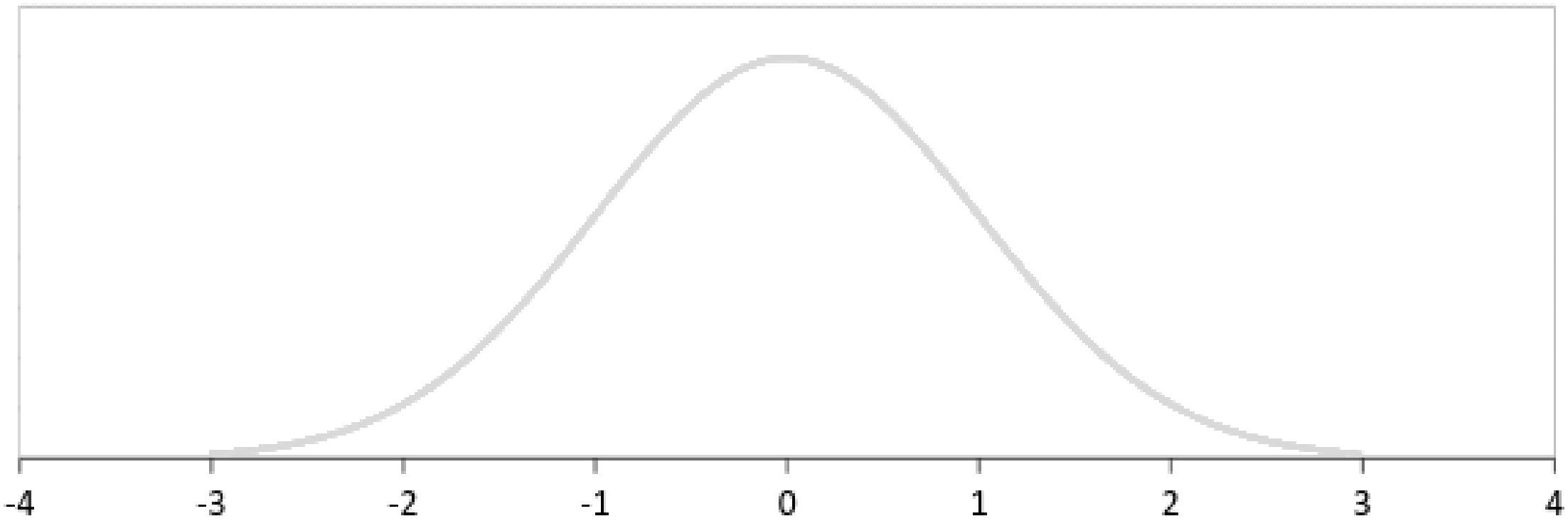
Variance -> $Var(X) = n \times p \times q$

Normal Distribution

- Any distribution is considered normal if it possesses the following characteristics:

- The distribution's mean, median, and mode all coincide.
- The distribution curve is bell-shaped and symmetrical about the line $x = \mu$.
- The entire area under the curve is equal to one.
- Half of the values are to the left of the center, while the other half are to the right.

Standard Normal Distribution



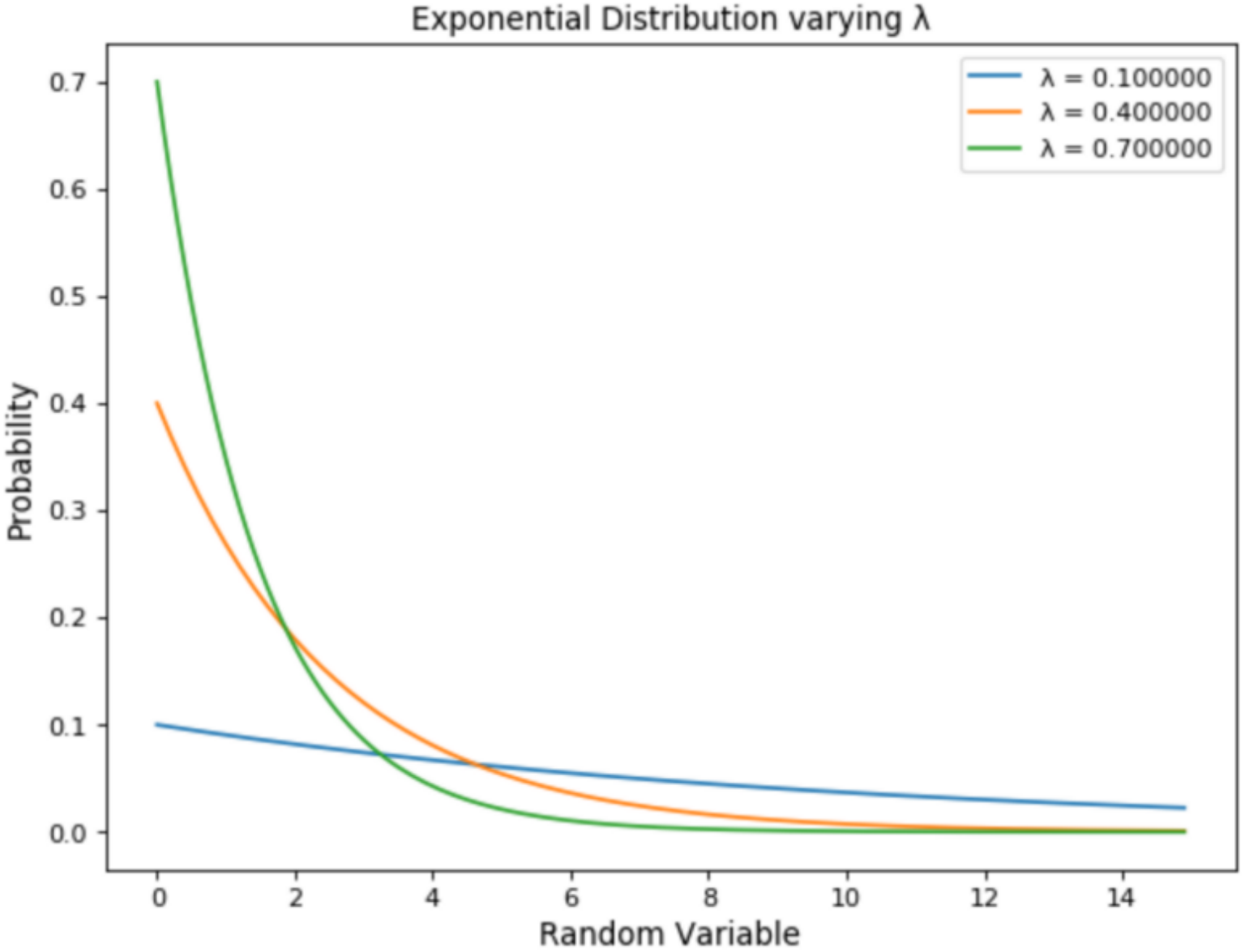
Poisson Distribution

- A distribution is called Poisson distribution when the following assumptions are valid:
  - Any successful event should not influence the outcome of another successful event.
  - The probability of success over a short interval must equal the probability of success over a longer interval.
  - The probability of success in an interval approaches zero as the interval becomes smaller.
- Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:
  - λ is the rate at which an event occurs,
  - t is the length of a time interval,
  - And X is the number of events in that time interval.
  - Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Exponential Distribution

- The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. We will now mathematically define the exponential distribution and derive its mean and expected value. Then we will develop the intuition for the distribution and discuss several interesting properties that it has.
- The Exponential Distribution is modeled using the following formula

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$


Exponential Distribution varying λ
λ = 0.100000
λ = 0.400000
λ = 0.700000
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
12
14
Random Variable

Central limit theorem

- The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution.
- The Central Limit Theorem requires that the average of your sample means equal the population mean. To put it another way, sum up the means from all of your samples, determine the average, and that average will be your real population mean. Similarly, if you take the average of all the standard deviations in your sample, you'll get the population standard deviation. It's a fairly handy phenomenon that can help precisely anticipate population features.

$$\lim_{n \to \infty} \mathbb{P} \frac{\left(\overline{X}_n - \mu \right)}{\sigma} \leq z = \Phi(z)$$

Where:

X_n is an IID sequence,

$\Phi(z) = \mathbb{P}(Z \leq z) =$

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in