Introduction to Statistics and Probability

Agenda

1. Introduction to Statistics
2. Data Science and Statistics
3. Type of Statistics
4. Descriptive Statistics
    a) MEASURES OF CENTRAL TENDENCY
    b) MEASURES OF VARIABILITY
    c) SHAPE
    d) DISTRIBUTION
    e) COVARIANCE AND CORRELATION
5. Recap

Introduction to Statistics

What is Statistics?

- Statistics is one of the popularly known disciplines that is mainly focused on data collection, data organization, data analysis, data interpretation, and data visualization.

Introduction to Statistics

Who uses Statistics?

- Earlier, statistics was practiced by statisticians, economists, business owners to calculate and represent relevant data in their field.
- Nowadays, statistics has taken a pivotal role in various fields like data science, machine learning, data analyst role, business intelligence analyst role, computer science role, and much more..

Data Science and Statistics

Why does Statistics matter in Data Science?

The importance of statistics for data science and statistics for data analytics is immense.

- Description and quantification of data
- Data identification and conversion of data patterns into usable format
- To collect, analyze, evaluate, and conclude the results for data using mathematical models
- Organize data while spotting the trends.
- Contributes to probability distribution and estimation
- Enhance the data visualization and reduce the assumptions.

Data Science and Statistics

Statistics in Relation With Machine Learning:

- Machine learning is like a puzzle, and the most important piece is statistics.
- To use machine learning to solve real-life problems, you need to know statistics well
- If you want to understand machine learning deeply, you should learn how statistics is the basis for things like prediction and sorting data.
- It helps us learn from information and make sense of data that doesn't have clear labels.

Data Science and Statistics

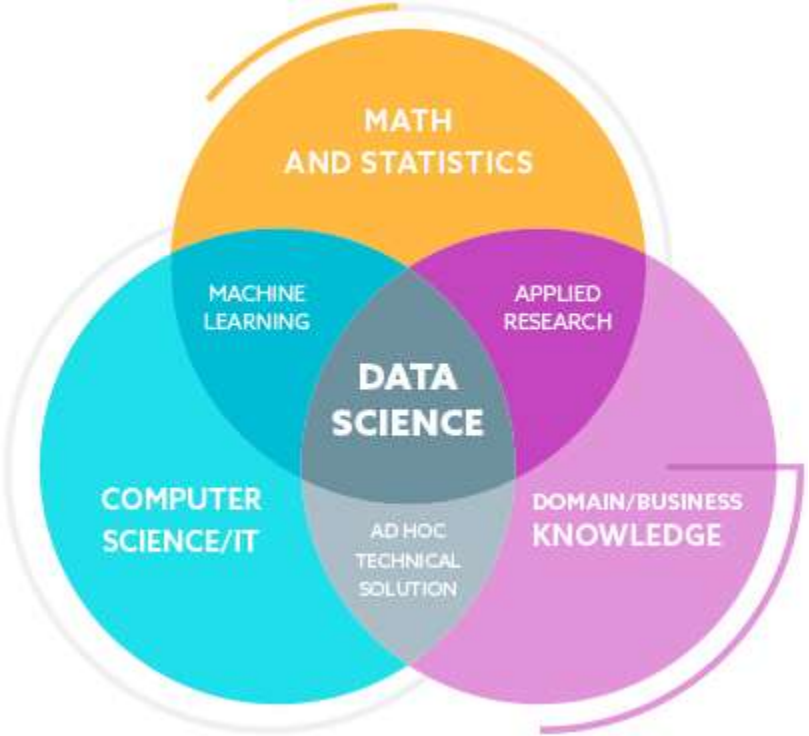
But Why Statistics?

Each and every organization aspires to be data-driven. This explains why the demand for data scientists and analysts is rising so quickly.

- Statistics play a vital role in the medical industry as they help determine the effectiveness of drugs before prescribing them.
- Netflix, for example, uses the number of movies browsed in different genres to recommend new movies based on individual preferences.

Data Science and Statistics

Always remember that Data Science is a multi- disciplinary science that consists of several sciences together.



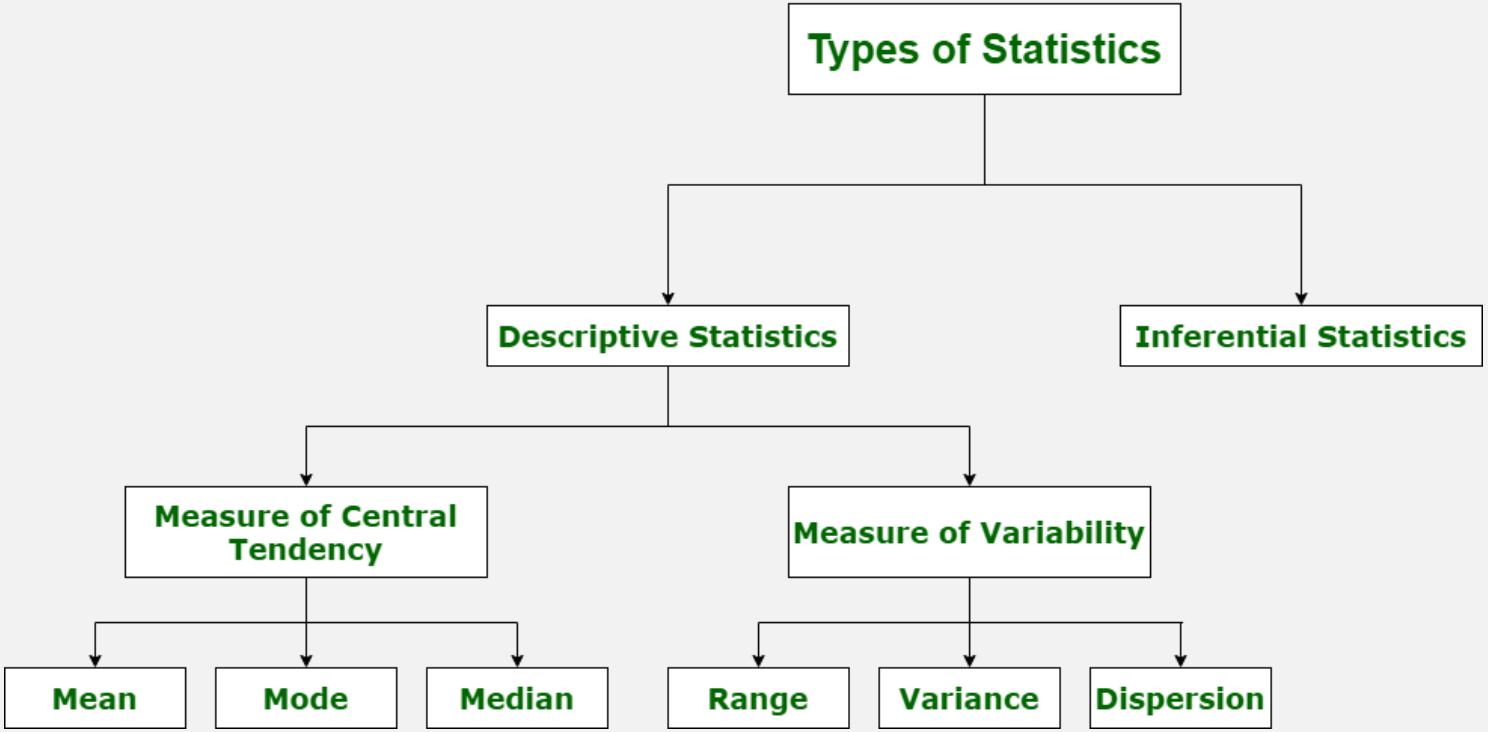
Data Science and Statistics

Popular Terms Used in Statistics

Population	Sample	Parameter
Statistics	Variable	Probability Distribution

Types of Statistics

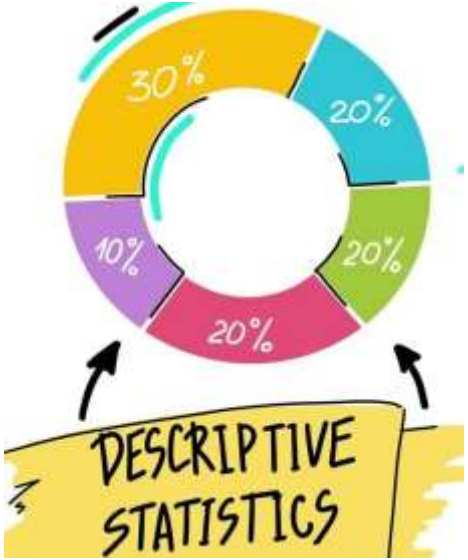
There are 2 main types of statistics:



Types of Statistics

Descriptive Statistics allows us to analyze and summarize data and organize the same in the form of numbers graph, bar plots, histogram, pie chart, etc.

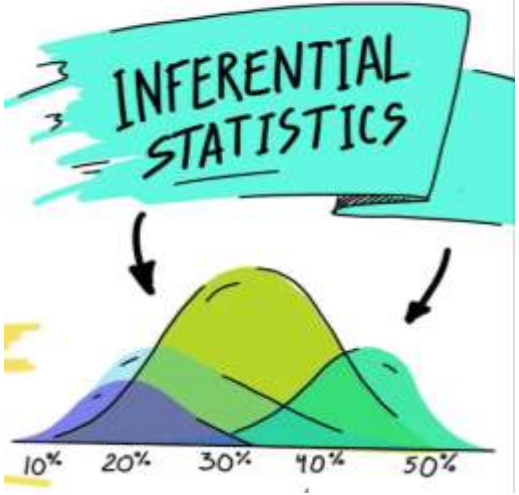
- Descriptive statistics is simply a process to describe our existing data.
- Concepts like standard deviation, central tendency are widely used around the world when it comes to learning descriptive statistics.



Types of Statistics

Inferential Statistics makes inferences and predictions about the population based on a sample of data taken from the population.

- It is simply used to analyze, interpret results, and draw conclusions.
- Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject the null hypothesis..



How to Make Sense of the Current Data..?

How to Make Sense of the Current Data..?

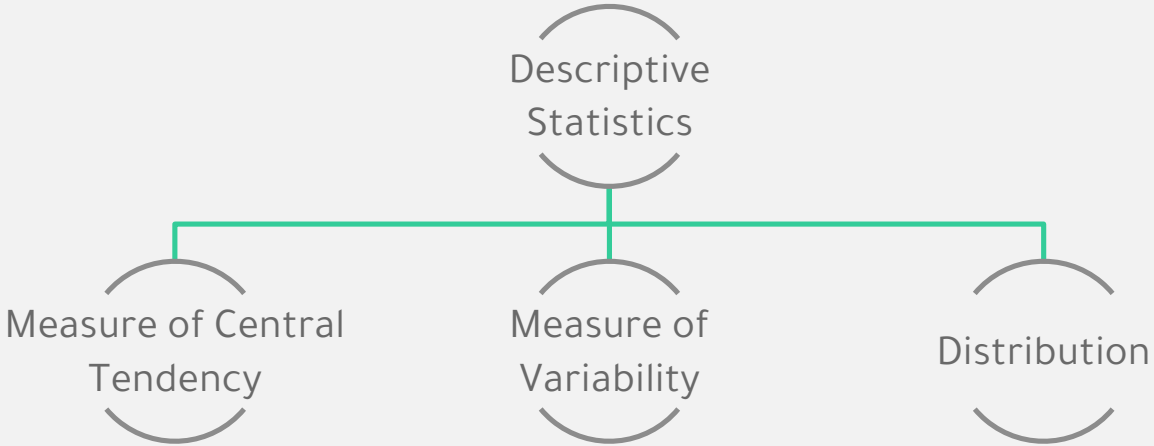
Data is essentially nothing but a collection of observations that are present in our company system.

With help of descriptive statistics, we can collect, organize, categorize, sample, visualize the data to make informed decisions for the company.

We can also use inferential statistics to predict outcomes. Generally, you conduct surveys to collect samples of data, and based on that we predict the findings for the entire population of that location.

Descriptive Statistics

Descriptive Statistics



Descriptive Statistics

MEASURE OF CENTRAL TENDENCY

Measure of central tendency is also known as summary statistics that are used to represent the **center point** or a particular value of a data set or sample set.

In statistics, there are three common measures of central tendency that are:

1. **Mean**
2. **Median**
3. **Mode**

Descriptive Statistics

MEASURE OF CENTRAL TENDENCY

1. **Mean:**

   It is the measure of the average of all values in a sample set. The mean of the data set is calculated using the formula:

$$Mean = \mu = \sum {}^{x}/_{n}$$

Descriptive Statistics

MEASURE OF CENTRAL TENDENCY

1. **Mean:**

   It is the measure of the average of all values in a sample set. The mean of the data set is calculated using the formula:

$$Mean = \mu = \sum {}^{x}/_{n}$$

   Example: Calculate the mean for each of the below columns.

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santra	19	4
i 20	15	4

Descriptive Statistics

MEASURE OF CENTRAL TENDENCY

2. **Median:**

It is the measure of the central value of a sample set. In these, the data set is ordered from lowest to highest value and then finds the exact middle. The formula used to calculate the median of the data set is, suppose we are given 'n' terms in s data set:

If n is even:

$$Median = [({}^{n}/_{2})^{th}\ term + ({}^{n+1}/_{2})^{th}\ term]/2$$

If n is odd:

$$Median = (n+1)/2$$

Descriptive Statistics

MEASURE OF CENTRAL TENDENCY

2. **Median:**

It is the measure of the central value of a sample set. In these, the data set is ordered from lowest to highest value and then finds the exact middle. The formula used to calculate the median of the data set is, suppose we are given 'n' terms in s data set:

If n is even:

$$Median = [(^{n}/_{2})^{th}\ term + (^{n+1}/_{2})^{th}\ term]/2$$

If n is odd:

$$Median = (n+1)/2$$

Example: Calculate the Median for each of the below columns.

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santra	19	4
i 20	15	4

Descriptive Statistics

MEASURE OF CENTRAL TENDENCY

3. **Mode:**

   It is the value most frequently arrived in the sample set. The value repeated most of the time in the central set is actually mode. The mode of the data set is calculated using the formula:

$$Mode = Term\ with\ Highest\ Frequency$$

Descriptive Statistics

MEASURE OF CENTRAL TENDENCY

3. **Mode:**

It is the value most frequently arrived in the sample set. The value repeated most of the time in the central set is actually mode. The mode of the data set is calculated using the formula:

$$Mode = Term\ with\ Highest\ Frequency$$

Example: Calculate the mean for each of the below columns.

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santra	19	4
i 20	15	4

Descriptive Statistics

MEASURE OF VARIABILITY

The measure of Variability is also known as the measure of dispersion and is used to describe variability in a sample or population.

In statistics, there are three common measures of central tendency that are:

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

Descriptive Statistics

MEASURE OF VARIABILITY

1. **Range:**

   It is a given measure of how to spread apart values in a sample set or data set. The formula of the range in statistics can simply be given by the difference between the highest and lowest values.

   Range = Maximum value – Minimum value

Descriptive Statistics

MEASURE OF VARIABILITY

1. **Range:**

   It is a given measure of how to spread apart values in a sample set or data set. The formula of the range in statistics can simply be given by the difference between the highest and lowest values.

   $$\textbf{\textit{Range = Maximum value – Minimum value}}$$

   Example: Find the range of given observations:

   $$\{32, 41, 28, 54, 35, 26, 23, 33, 38, 40\}$$

Descriptive Statistics

MEASURE OF VARIABILITY

2. **Variance:**

   In probability theory and statistics, variance measures a data set's spread or dispersion. It is calculated by averaging the squared deviations from the mean.

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

Descriptive Statistics

MEASURE OF VARIABILITY

2. **Variance:**

In probability theory and statistics, variance measures a data set's spread or dispersion. It is calculated by averaging the squared deviations from the mean.

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

```
def variance(data):
        # Number of observations
        n = len(data)
        # Mean of the data
        mean = sum(data) / n
        # Square deviations
        deviations = [(x - mean) ** 2 for x in data]
        # Variance
        variance = sum(deviations) / n
        return variance

variance([4, 8, 6, 5, 3, 2, 8, 9, 2, 5])
```

Descriptive Statistics

MEASURE OF VARIABILITY

3. **Standard Deviation:**

   The standard deviation is simply the square root of the variance and measures the extent to which data varies from its mean.

$$\textit{Standard Deviation} = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} (x_i - \mu)^2 \Big/ N}$$

Descriptive Statistics

MEASURE OF VARIABILITY

3. **Standard Deviation:**

   The standard deviation is simply the square root of the variance and measures the extent to which data varies from its mean.

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

```
def variance(data):
    # Number of observations
    n = len(data)
    # Mean of the data
    mean = sum(data) / n
    # Square deviations
    deviations = [(x - mean) ** 2 for x in data]
    # Variance
    variance = sum(deviations) / n
    return variance

def stdev(data):
   var = variance(data)
   std_dev = math.sqrt(var)
   return std_dev

stdev([4, 8, 6, 5, 3, 2, 8, 9, 2, 5])
```

Descriptive Statistics

SHAPE

Measures of shape describe the distribution (or pattern) of the data within a dataset.

The distribution shape of quantitative data can be described as there is a logical order to the values, and the 'low' and 'high' end values on the x-axis of the histogram are able to be identified.

The distribution shape of a qualitative data cannot be described as the data are not numeric.

1. **Symmetrical Distribution**
2. **Asymmetrical Distribution**

Descriptive Statistics

SHAPE

Shape Metrics

Skewness essentially is a commonly used measure in descriptive statistics that characterizes the asymmetry of a data distribution, while **kurtosis** determines the heaviness of the distribution tails.

- Skewness is a measure of asymmetry in a distribution. Types of skewness include:
    1. **Positive skewness:** The tail of the distribution extends more towards the right, indicating a longer right tail and a majority of lower values**.**
    2. **Negative skewness:** The tail of the distribution extends more towards the left, indicating a longer left tail and a majority of higher values.
- Kurtosis measures the peakedness or flatness of a distribution's shape. It describes the tails of the distribution compared to the normal distribution.

Descriptive Statistics

SHAPE

1. **Symmetrical Distribution**

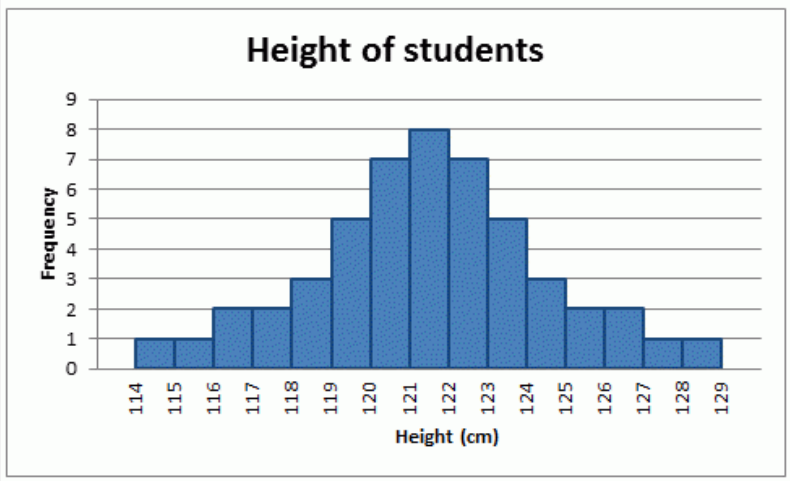
   - In a symmetrical distribution the two sides of the distribution are a mirror image of each other.
   - A normal distribution is a true symmetric distribution of observed values.
   - When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape. This is why this distribution is also known as a 'normal curve' or 'bell curve'.

Descriptive Statistics

SHAPE

1. **Symmetrical Distribution**

   - In a symmetrical distribution the two sides of the distribution are a mirror image of each other.
   - A normal distribution is a true symmetric distribution of observed values.
   - When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape. This is why this distribution is also known as a 'normal curve' or 'bell curve'.



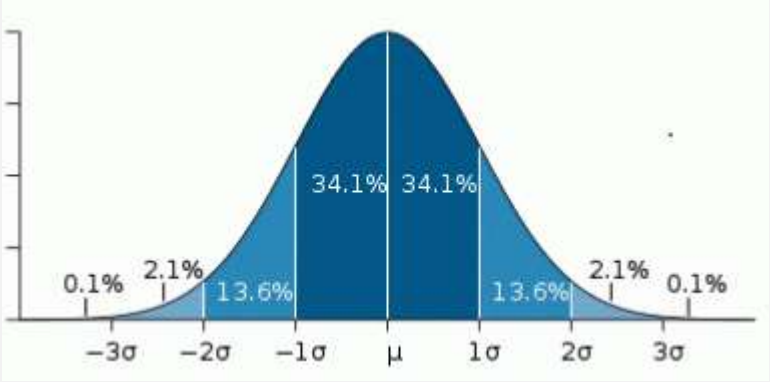
Descriptive Statistics

SHAPE

1. **Symmetrical Distribution**

Key features of the normal distribution:

- symmetrical shape
- mode, median and mean are the same and are together in the center of the curve
- most of the data are clustered around the center, while the more extreme values on either side of the center become less rare as the distance from the center increases

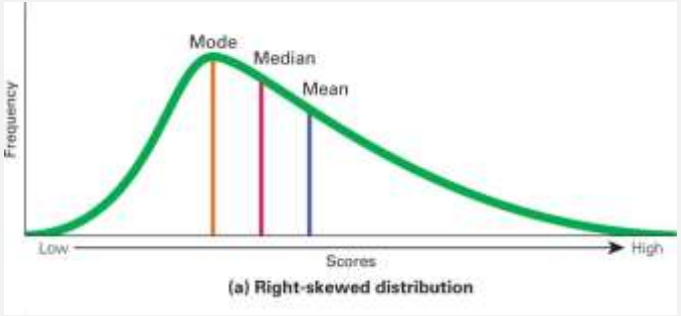


Descriptive Statistics

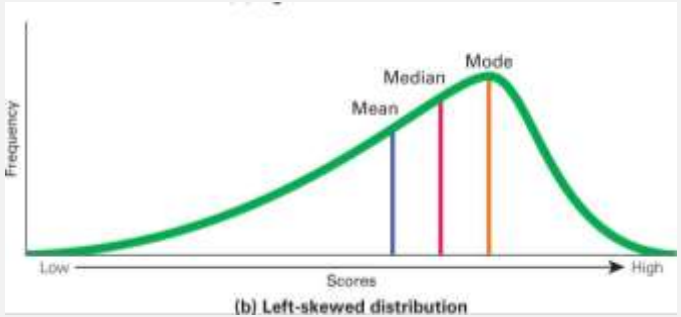
SHAPE

2. **Asymmetrical Distribution**

- In an asymmetrical distribution the two sides will not be mirror images of each other.
- Skewness is the tendency for the values to be more frequent around the high or low ends of the x-axis.
- When a histogram is constructed for skewed data it is possible to identify skewness by looking at the shape of the distribution.



(a) Right-skewed distribution



(b) Left-skewed distribution

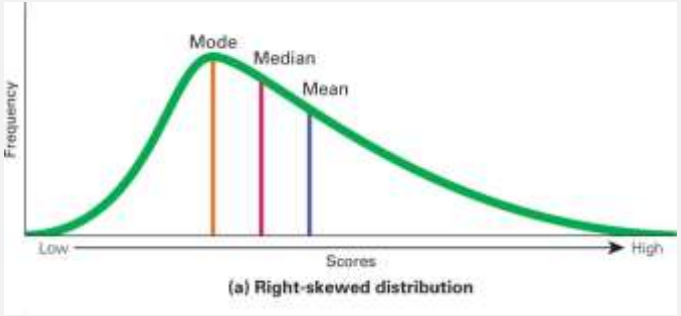
Descriptive Statistics

SHAPE

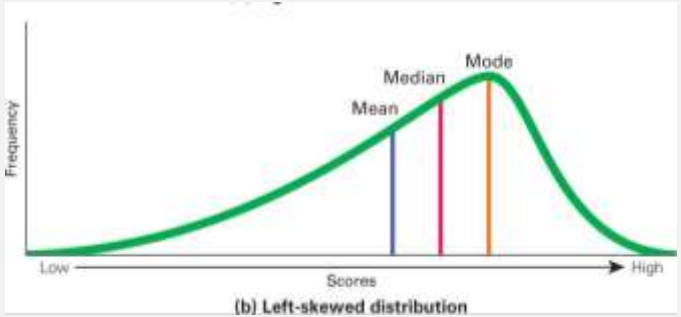
2. **Asymmetrical Distribution**

Key features asymmetrical distribution:

- asymmetrical shape
- mean and median have different values and do not all lie at the center of the curve
- there can be more than one mode
- the distribution of the data tends towards the high or low end of the dataset



(a) Right-skewed distribution



(b) Left-skewed distribution

Descriptive Statistics

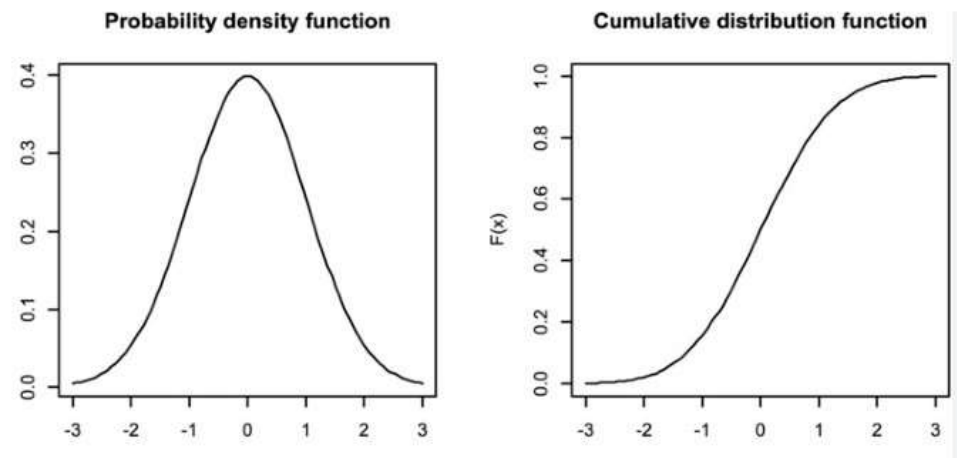
DISTRIBUTION

A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.

A sample of data will form a distribution, and by far the most well-known distribution is the Gaussian distribution, often called the Normal distribution.

Distributions are often described in terms of their density or density functions.

Density functions are functions that describe how the proportion of data or likelihood of the proportion of observations change over the range of the distribution.



Descriptive Statistics

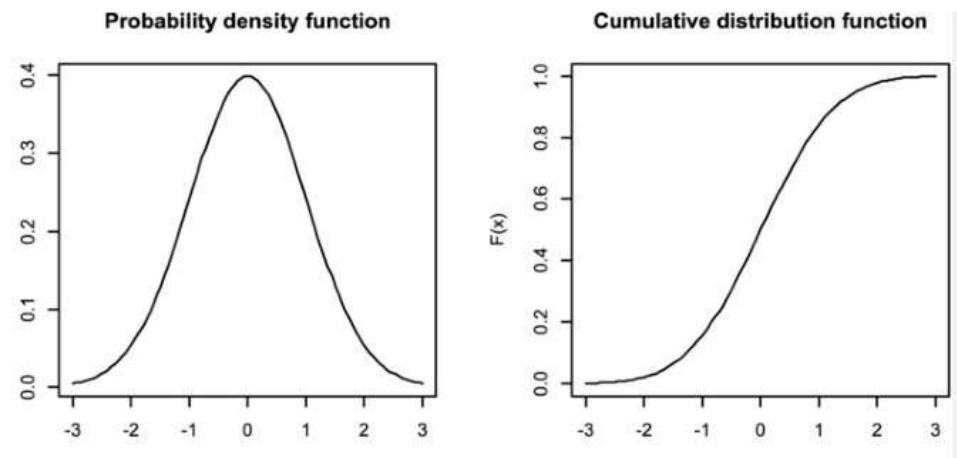
DISTRIBUTION

A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.

A sample of data will form a distribution, and by far the most well-known distribution is the Gaussian distribution, often called the Normal distribution.

Distributions are often described in terms of their density or density functions.

Density functions are functions that describe how the proportion of data or likelihood of the proportion of observations change over the range of the distribution.



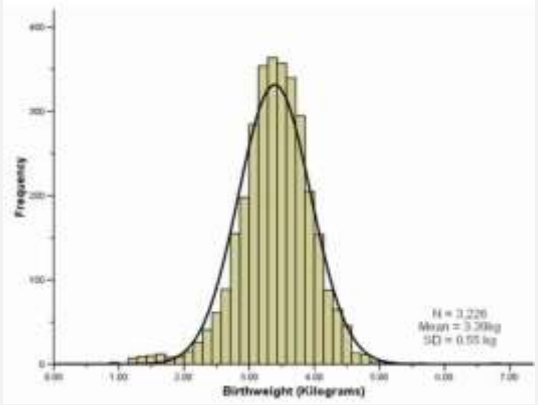
Next, let's look at the Gaussian distribution and two other distributions related to the Gaussian that you will encounter when using statistical methods.

Descriptive Statistics

DISTRIBUTIONS

1. **Gaussian Distribution**

   - Data from many fields of study surprisingly can be described using a Gaussian distribution, so much so that the distribution is often called the "normal" distribution because it is so common.
   - A Gaussian distribution can be described using two parameters:
     - ✓ **mean**: Denoted with the Greek lowercase letter mu, is the expected value of the distribution.
     - ✓ **variance**: Denoted with the Greek lowercase letter sigma raised to the second power (because the units of the variable are squared), describes the spread of observation from the mean.

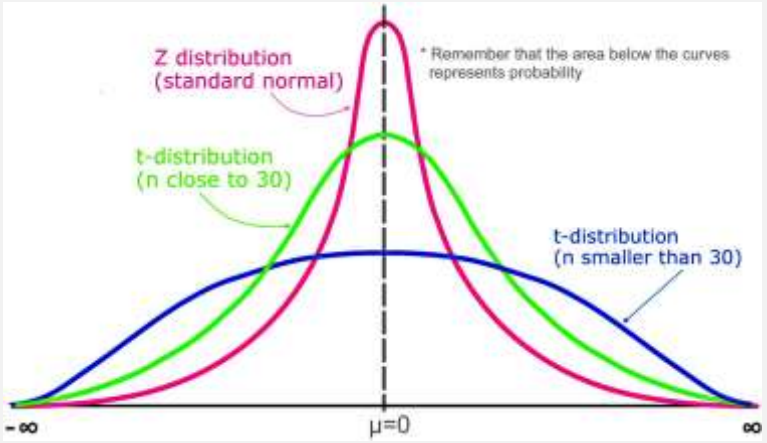


Descriptive Statistics

DISTRIBUTIONS

2. **Student's t-Distribution**

- It is a distribution that arises when attempting to estimate the mean of a normal distribution with different sized samples.
- It is a helpful shortcut when describing uncertainty or error related to estimating population statistics for data drawn from Gaussian distributions when the size of the sample must be taken into account.
- The distribution can be described using a single parameter:
  - ✓ number of degrees of freedom: denoted with the lowercase Greek letter n, denotes the number degrees of freedom.

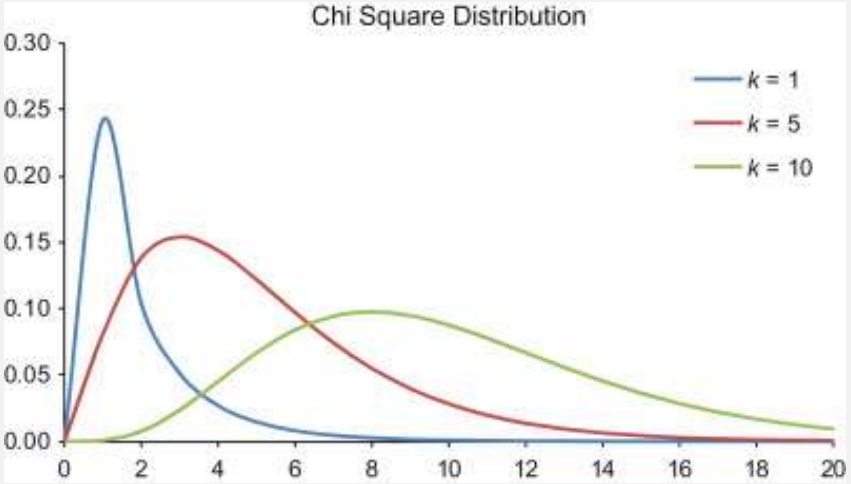


Descriptive Statistics

DISTRIBUTIONS

3. **Chi-Squared Distribution**

- The chi-squared distribution is denoted as the lowercase Greek letter chi χ^2
- Like the Student's t-distribution, the chi-squared distribution is also used in statistical methods on data drawn from a Gaussian distribution to quantify the uncertainty.
- The chi-squared distribution has one parameter:
  - ✓ degrees of freedom, denoted k



Descriptive Statistics

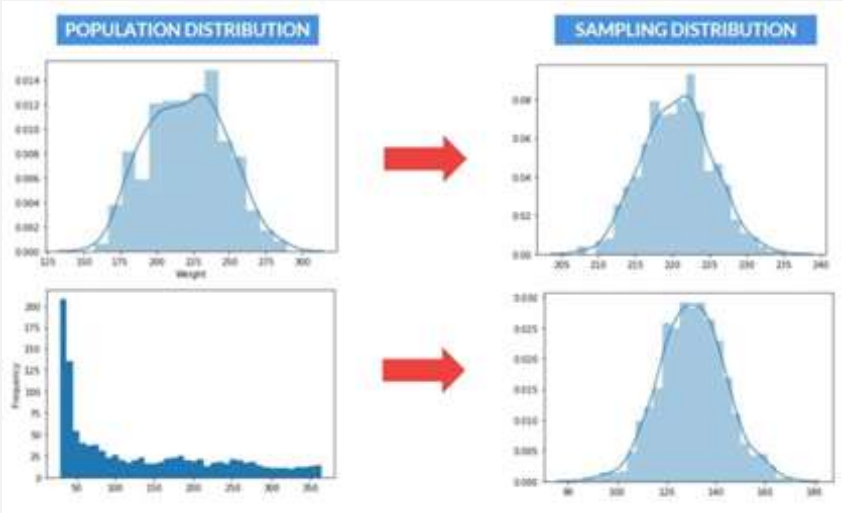
DISTRIBUTIONS

Central Limit Theorem

- The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

> *" The central limit theorem says that the sampling distribution of the mean will always be **normally distributed**, as long as the sample size is large enough "*

- Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.
- By convention, we consider a sample size of 30 to be "sufficiently large."



Descriptive Statistics

COVARIANCE

Covariance is a measure of the joint variability of two variables. In other words, covariance measures how one variable varies with respect to the other variable.

- Covariance can be either positive or negative.
- ***Example***, cov = 5.0, This means that the covariance agrees with us that the relation between x and y is positive. But what about the strength of the relationship?
- Covariance does not give a proper indication of the strength of the relationship. ***Is the value 5.0 strong, or is 50.0 strong, or 500000.0 is strong?***

$$Cov(X,Y) = {}^{1}/_{n} \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

Descriptive Statistics

CORRELATION

The correlation coefficient is a simple descriptive statistic that measures the **strength** of the linear relationship between two interval- or ratio-scale variables.

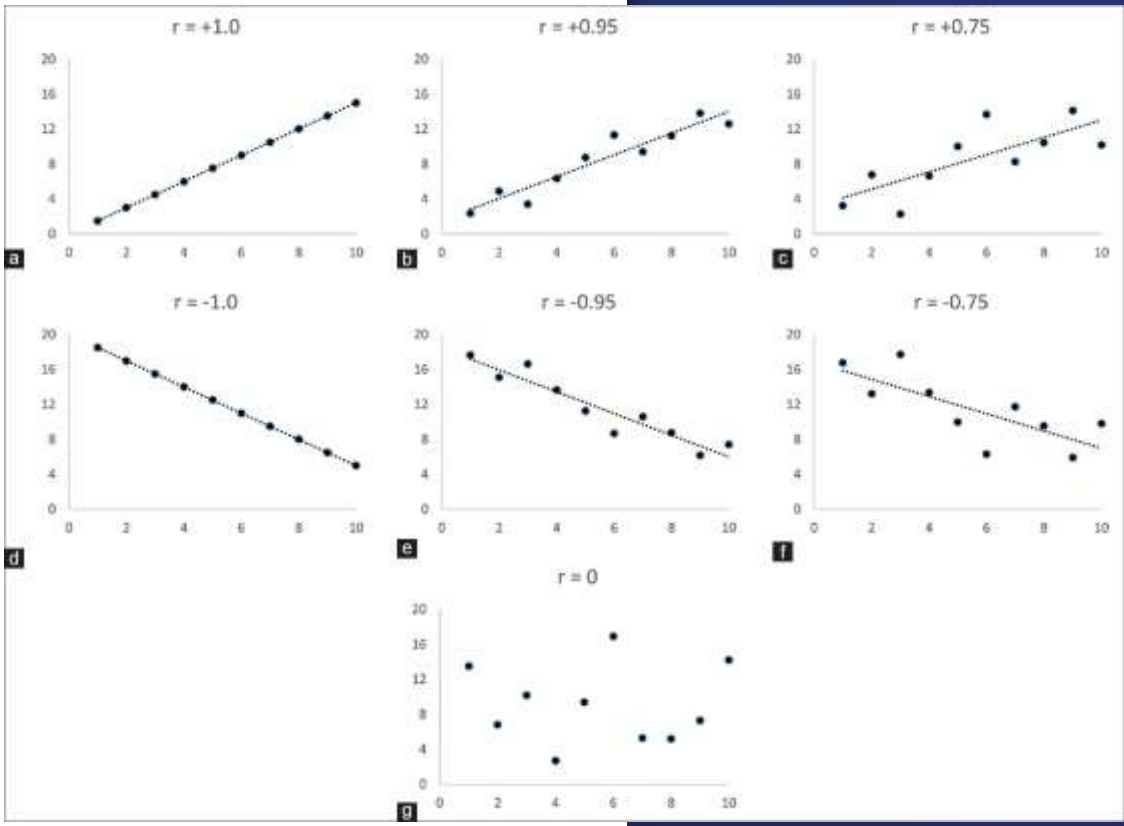
- The values here range from -1 to 1.
- The sample correlation coefficient, r, is also tested for statistical significance.
- Correlation is symmetrical, the correlation between A and B is the same as the correlation between B and A.

Descriptive Statistics

CORRELATION

Correlations of >+0.5 or <-0.5, respectively, are often regarded as a strong positive or strong negative association between two variables.

- ✓ **Perfectly Positive Correlation:** When correlation value is exactly 1.
- ✓ **Positive Correlation**: When correlation value falls between 0 to 1.
- ✓ **No Correlation**: When correlation value is 0.
- ✓ **Negative Correlation**: When correlation value falls between -1 to 0.
- ✓ **Perfectly Negative Correlation:** When correlation value is exactly -1.



Descriptive Statistics

CORRELATION

Types of Correlation Metrics:

Pearson Correlation	➉ It can be used for nominal variables or continuous variables. ➉ Only captures linear relationships.
Spearman's Rank Correlation	➉ It can be used for ordinal variables or continuous variables. ➉ Captures both linear and non-linear relationships.
Kendall Rank Correlation	➉ It can capture both linear or non-linear relationships between two variables. ➉ It is a robust estimator against outliers and non-normality.

TO BE CONTINUED…