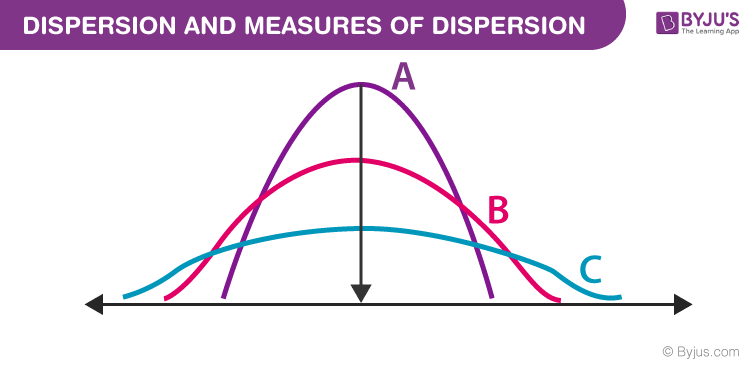
**Descriptive statistics**

**Measure of Dispersion** –



Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

* Absolute Measure of Dispersion

The types of absolute measures of dispersion are:

**Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6

**Variance:** Deduct the mean from each data in the set, square each of them and add each square and finally divide them by the total no of values in the data set to get the variance. Variance (σ2) = ∑(X−μ)2/N

**Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = √σ.

**Quartiles and Quartile Deviation:**The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.

**Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

* Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

1. Co-efficient of Range
2. Co-efficient of Variation
3. Co-efficient of Standard Deviation
4. Co-efficient of Quartile Deviation
5. Co-efficient of Mean Deviation

## Co-efficient of Dispersion

The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages. The dispersion coefficient is also used when two series with different measurement units are compared. It is denoted as C.D.

The common coefficients of dispersion are:

| **C.D. in terms of** | **Coefficient of dispersion** |
| --- | --- |
| Range | C.D. = (X max – X min) ⁄ (X max + X min) |
| Quartile Deviation | C.D. = (Q3 – Q1) ⁄ (Q3 + Q1) |
| Standard Deviation (S.D.) | C.D. = S.D. ⁄ Mean |
| Mean Deviation | C.D. = Mean deviation/Average |

## What are quartiles?

Quartiles are a set of [descriptive statistics](https://www.scribbr.com/statistics/descriptive-statistics/). They summarize the [central tendency](https://www.scribbr.com/statistics/central-tendency/) and [variability](https://www.scribbr.com/statistics/variability/) of a dataset or distribution.

* The**first quartile** (Q1, or the lowest quartile) is the 25th percentile, meaning that 25% of the data falls below the first quartile.
* The **second quartile**(Q2, or the median) is the 50th percentile, meaning that 50% of the data falls below the second quartile.
* The **third quartile** (Q3, or the upper quartile) is the 75th percentile, meaning that 75% of the data falls below the third quartile.

**Step-by-step example**

Imagine you conducted a small study on language development in children 1–6 years old. You’re writing a paper about the study and you want to report the quartiles of the children’s ages.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age (years)** | 1 | 2 | 3 | 4 | 5 | 6 |
| **Frequency** | 2 | 3 | 4 | 1 | 2 | 2 |

Step 1: Count the number of observations in the data set *n*= 2 + 3 + 4 + 1 + 2 + 2 = 14

Step 2: Sort the observations in increasing order

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

Step 3: Find the first quartile *n*\* (1 / 4) = 14 \* (1 / 4) = 3.5  
3.5 is not an integer, so Q1 is the number at position 4.  
1, 1, 2, **2**, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

**Q1 = 2 years**

Step 4: Find the second quartile *n*\* (2 / 4) = 14 \* (2 / 4) = 7  
7 is an integer, so Q2 is the mean of the numbers at positions 7 and 8.  
1, 1, 2, 2, 2, 3, **3**, **3**, 3, 4, 5, 5, 6, 6  
Q2 = (3 + 3) / 2  
**Q2 = 3 years**

Step 5: Find the third quartile *n*\* (3 / 4) = 14 \* (3 / 4) = 10.5  
10.5 is not an integer, so Q3 is the number at position 11.  
1, 1, 2, 2, 2, 3, 3, 3, 3, 4, **5**, 5, 6, 6  
**Q3 = 5 years**

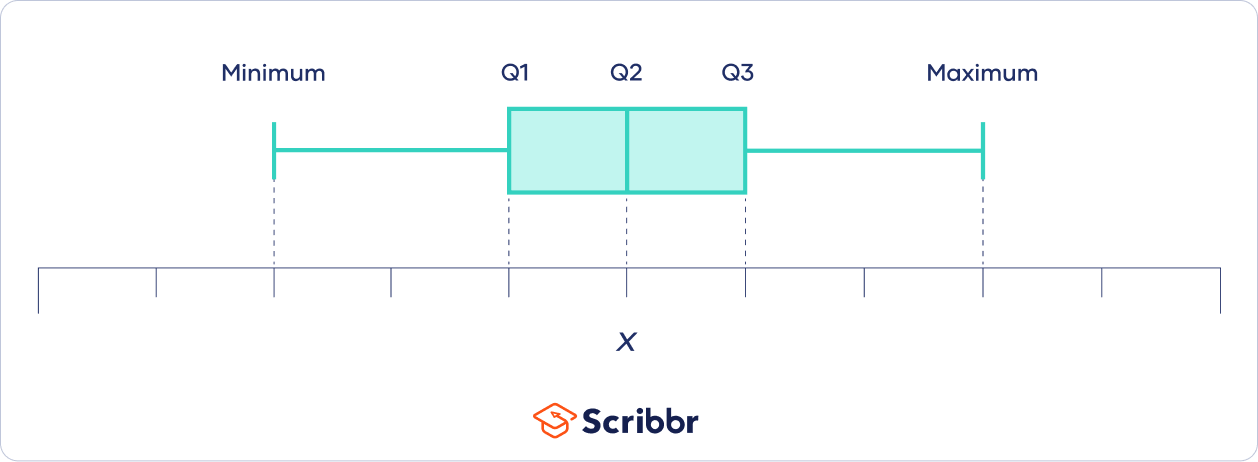
**Visualizing quartiles with boxplots**

Boxplots are helpful visual summaries of a dataset. They’re composed of boxes, which show the quartiles, and whiskers, which show the lowest and highest observations.

To make a boxplot, you first need to calculate the **five-number summary**:

1. The minimum
   * This is the lowest observation. If you order the numbers in your dataset from lowest to highest, the minimum is the first number. The minimum is sometimes called the zeroth quantile.
2. The first quartile
3. The second quartile
4. The third quartile
5. The maximum
   * This is the highest observation. If you order the numbers in your dataset from lowest to highest, the maximum is the last number. The maximum is sometimes called the fourth quantile.

With these five numbers, you can draw a boxplot:



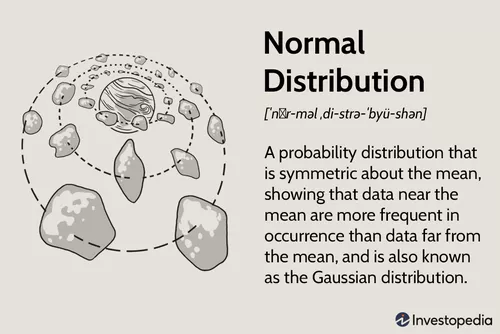
## What are quantiles?

A quartile is a type of quantile.

Quantiles are values that split sorted data or a [probability distribution](https://www.scribbr.com/statistics/probability-distributions/) into equal parts. In general terms, a *q*-quantile divides sorted data into *q*parts. The most commonly used quantiles have special names:

* **Quartiles (4-quantiles)**: Three quartiles split the data into four parts.
* **Deciles (10-quantiles)**: Nine deciles split the data into 10 parts.
* **Percentiles (100-quantiles):**99 percentiles split the data into 100 parts.

**Normal/Gaussian Distribution**

****

### **KEY TAKEAWAYS**

* The normal distribution is the proper term for a probability bell curve.
* In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
* Normal distributions are symmetrical, but not all symmetrical distributions are normal.
* Many naturally-occurring phenomena tend to approximate the normal distribution.
* In finance, most pricing distributions are not, however, perfectly normal.

## Properties of the Normal Distribution

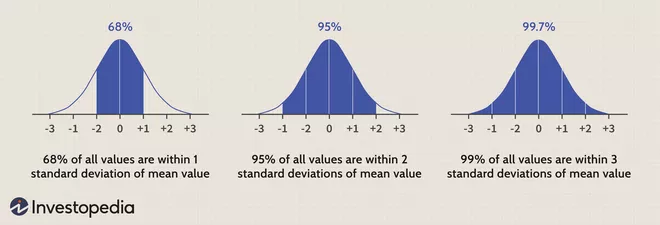
The normal distribution has several key features and properties that define it.

First, its [mean](https://www.investopedia.com/terms/m/mean.asp) (average), [median](https://www.investopedia.com/terms/m/median.asp) (midpoint), and [mode](https://www.investopedia.com/terms/m/mode.asp) (most frequent observation) are all equal to one another. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp).

*All normal distributions can be described by just two parameters: the mean and the standard deviation.*

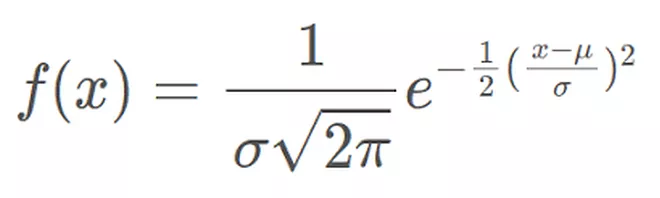
### **The Empirical Rule**

For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% of the observations will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations. This fact is sometimes referred to as the "empirical rule,"

****

The Formula for the Normal Distribution

The normal distribution follows the following formula. Note that only the values of the mean (μ ) and standard deviation (σ) are necessary



Normal Distribution Formula.

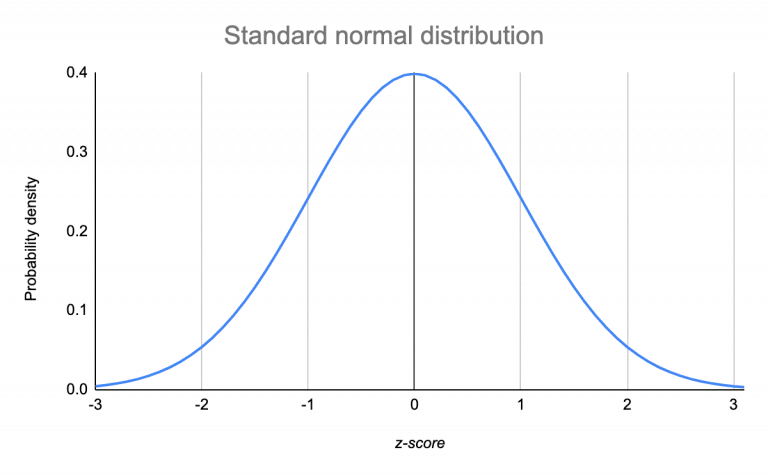
where:

* *x* = value of the variable or data being examined and f(x) the probability function
* μ = the mean
* σ = the standard deviation

# The Standard Normal Distribution

The **standard normal distribution**, also called the **z-distribution**, is a special [normal distribution](https://www.scribbr.com/statistics/normal-distribution/) where the [mean](https://www.scribbr.com/statistics/mean/) is 0 and the [standard deviation](https://www.scribbr.com/statistics/standard-deviation/) is 1.

Any normal distribution can be standardized by converting its values into z-scores. Z-scores tell you how many standard deviations from the mean each value lies.



## Standardizing a normal distribution

When you standardize a normal distribution, the mean becomes 0 and the standard deviation becomes 1.

 A *z*-score is a **standard score** that tells you how many standard deviations away from the mean an individual value (*x*) lies:

* A positive *z*-score means that your *x*-value is greater than the mean.
* A negative *z*-score means that your *x*-value is less than the mean.
* A *z*-score of zero means that your *x*-value is equal to the mean.

Converting a normal distribution into the standard normal distribution allows you to:

* Compare scores on different distributions with different means and standard deviations.
* Normalize scores for statistical decision-making (e.g., grading on a curve).
* Find the probability of observations in a distribution falling above or below a given value.
* Find the probability that a [sample mean](https://www.scribbr.com/methodology/population-vs-sample/) significantly differs from a known population mean.

### How to calculate a z-score

| ***Z*-score formula** | **Explanation** |
| --- | --- |
| z=\dfrac{x-\mu}{\sigma} | * *x* = individual value * μ = mean * σ = standard deviation |

# Normalization vs Standardization

**Normalization or Min-Max Scaling** is used to transform features to be on a similar scale. The new point is calculated as:

X\_new = (X - X\_min)/(X\_max - X\_min)

This scales the range to [0, 1] or sometimes [-1, 1].

Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

**Standardization or Z-Score Normalization**

 This is often called as Z-score.

X\_new = (X - mean)/Std

Standardization can be helpful in cases where the data follows a Gaussian distribution. Standardization does not get affected by outliers because there is no predefined range of transformed features.

**Difference between Normalization and Standardization**

| S.NO. | Normalization | Standardization |
| --- | --- | --- |
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| 6. | This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. | It translates the data to the mean vector of original data to the origin and squishes or expands. |
| 7. | It is useful when we don’t know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |
| 8. | It is a often called as Scaling Normalization | It is a often called as Z-Score Normalization. |

# **Central Limit Theorem (CLT)**

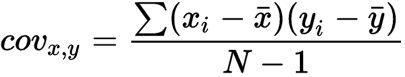
### **KEY TAKEAWAYS**

* The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
* Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
* A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
* A sufficiently large sample size can predict the characteristics of a population more accurately.
* CLT is useful in finance when analyzing a large collection of securities to estimate portfolio distributions and traits for returns, risk, and correlation.

## Covariance

### **KEY TAKEAWAYS**

* Covariance is a statistical tool that is used to determine the relationship between the movements of two random variables.
* When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.
* Covariance is different from the correlation coefficient, a measure of the strength of a correlative relationship.
* Covariance is a significant tool in modern portfolio theory used to ascertain what securities to put in a portfolio.
* Risk and volatility can be reduced in a portfolio by pairing assets that have a negative covariance.



cov\_{x,y} = covariance between variable x and y

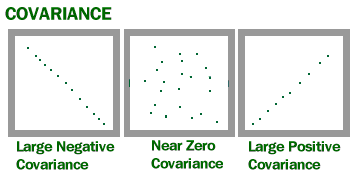
x\_{i} = data value of x

y\_{i} = data value of y

\bar{x}= mean of x

\bar{y}= mean of y

N = number of data values

****

# Pearson Correlation Coefficient

The **Pearson correlation coefficient (r)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

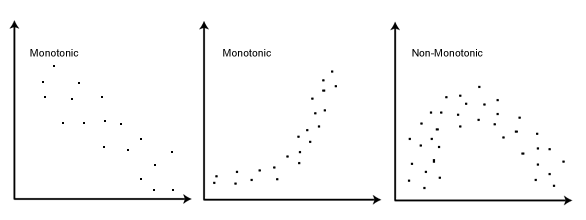
| **Pearson correlation coefficient (*r*)** | **Correlation type** | **Interpretation** | **Example** |
| --- | --- | --- | --- |
| Between 0 and 1 | Positive correlation | When one variable changes, the other variable changes in the **same direction**. | Baby length & weight:  The longer the baby, the heavier their weight. |
| 0 | No correlation | There is **no relationship** between the variables. | Car price & width of windshield wipers:  The price of a car is not related to the width of its windshield wipers. |
| Between 0 and –1 | Negative correlation | When one variable changes, the other variable changes in the **opposite direction**. | Elevation & air pressure:  The higher the elevation, the lower the air pressure. |

## ****Spearman correlation coefficient****

It measures the strength and direction of the association between two ranked variables.

Monotonic relationship

A monotonic relationship is a relationship that does one of the following:

****

**Inferential Statistics**

## Hypothesis Testing

### **KEY TAKEAWAYS**

* Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.
* The test provides evidence concerning the plausibility of the hypothesis, given the data.
* Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analysed.

All analysts use a random population sample to test two different hypotheses: the [null hypothesis](https://www.investopedia.com/terms/n/null_hypothesis.asp) and the alternative hypothesis. The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are [mutually exclusive](https://www.investopedia.com/terms/m/mutuallyexclusive.asp), and only one can be true. However, one of the two hypotheses will always be true.

### **4 Steps of Hypothesis Testing**

All hypotheses are tested using a four-step process:

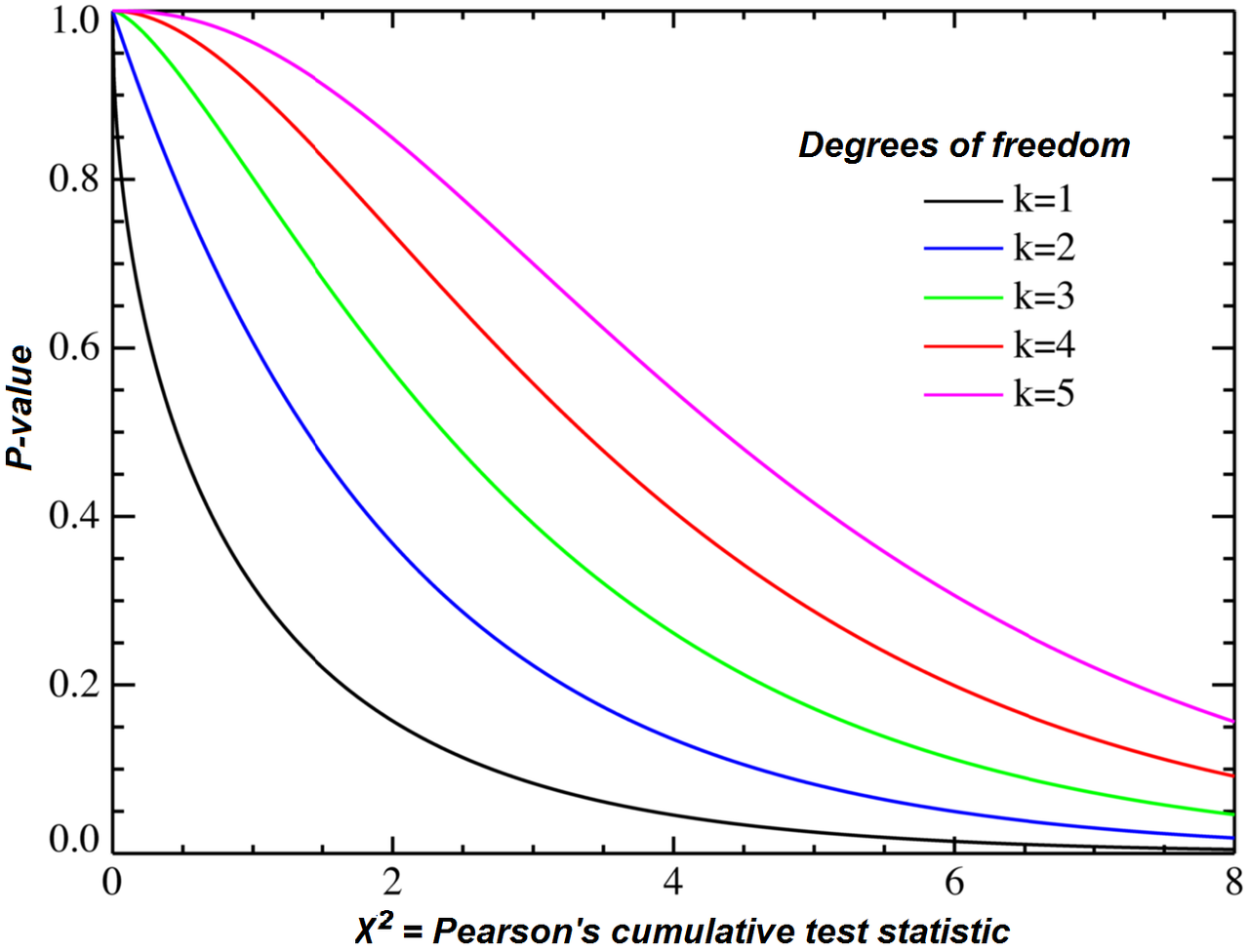
1. The first step is for the analyst to state the two hypotheses so that only one can be right.
2. The next step is to formulate an analysis plan, which outlines how the data will be evaluated.
3. The third step is to carry out the plan and physically analyze the sample data.
4. The fourth and final step is to analyze the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

**p-value**

A[p-value](https://www.scribbr.com/statistics/p-value/), or probability value, is a number describing how likely it is that your data would have occurred under the [null hypothesis](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/) of your [statistical test](https://www.scribbr.com/statistics/statistical-tests/).

# **re** Chi Square

A **chi-squared test** (symbolically represented as **χ2**) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. A chi-square test is a statistical test used **to compare observed results with expected results**. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying. It gives the probability of independent variables.

**Note:** Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.

# An Introduction to T-Tests

A t-test is a [statistical test](https://www.scribbr.com/statistics/statistical-tests/) that is used to compare the [means](https://www.scribbr.com/statistics/mean/) of two groups. It is often used in [hypothesis testing](https://www.scribbr.com/statistics/hypothesis-testing/) to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

## When to use a t-test

A t-test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an[ANOVA test](https://www.scribbr.com/statistics/one-way-anova/) or a post-hoc test.

The t-test is a [parametric test](https://www.scribbr.com/statistics/statistical-tests/#parametric) of difference, meaning that it makes the same assumptions about your data as other parametric tests. The t-test assumes your data:

1. are independent
2. are (approximately) normally distributed.
3. have a similar amount of [variance](https://www.scribbr.com/statistics/variance/) within each group being compared (a.k.a. homogeneity of variance)

If your data do not fit these assumptions, you can try a [nonparametric](https://www.scribbr.com/statistics/statistical-tests/#nonparametric) alternative to the t-test, such as the Wilcoxon Signed-Rank test for data with unequal variances.

## What type of t-test should I use?

### One-sample, two-sample, or paired t-test?

* If the groups come from a single population (e.g. measuring before and after an experimental treatment), perform a **paired t-test**.
* If the groups come from two different populations (e.g. two different species, or people from two separate cities), perform a **two-sample t-test**(a.k.a. **independent t-test**).
* If there is one group being compared against a standard value (e.g. comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample t-test**.

### One-tailed or two-tailed t-test?

* If you only care whether the two populations are different from one another, perform a **two-tailed t-test**.
* If you want to know whether one population mean is greater than or less than the other, perform a **one-tailed t-test.**

### T-test formula



In this formula, t is the t-value, x1 and x2 are the means of the two groups being compared, s2 is the pooled standard error of the two groups, and n1 and n2 are the number of observations in each of the groups.

**Anova Test**

### **KEY TAKEAWAYS**

* Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
* A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
* If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

**when to use Anova test ?**

1. It is only conducted when there is no relationship between the subjects in each sample. this means that subjects in the first group cannot also be in the second group ie independent samples between groups.
2. Groups must have equal sample size.

The Formula for ANOVA is



F = MST/MSE



**where:**

F=ANOVA coefficient



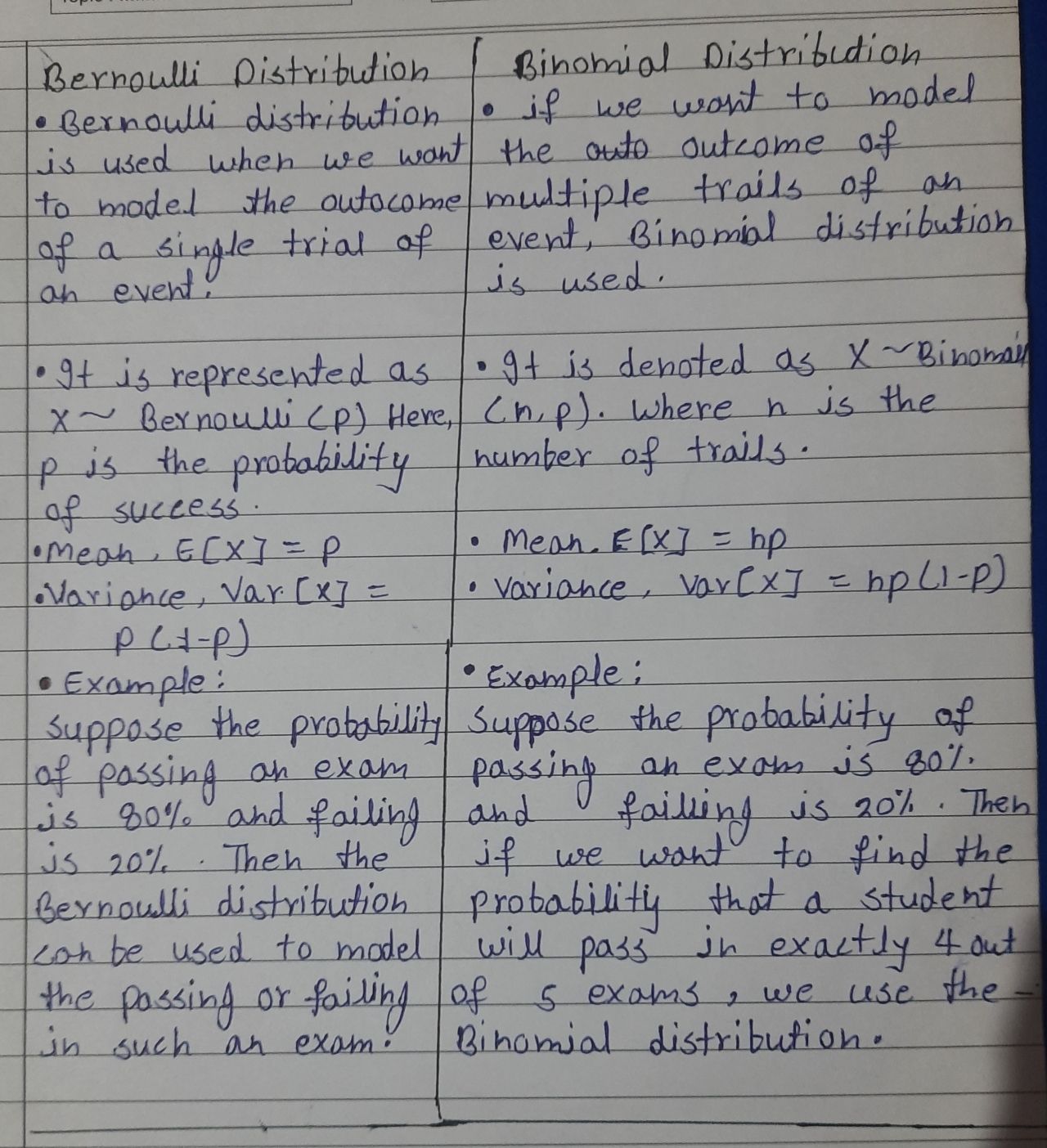
MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

## What is the difference between one-way and two-way ANOVA tests?

This is defined by how many independent variables are included in the ANOVA test. One-way means the analysis of variance has one independent variable. Two-way means the test has two independent variables. An example of this may be the independent variable being a brand of drink (one-way), or independent variables of brand of drink and how many calories it has or whether it’s original or diet.

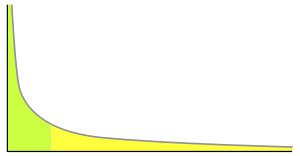
**Difference between Bernoulli and Binomial Distribution**



**Power Law Distribution/Pareto Diatribution**

 A power law distribution has the property that large numbers are rare, but smaller numbers are more common. So it is more common for a person to make a small amount of money versus a large amount of money.

The **Pareto** distribution is a continuous power law distribution that is based on the observations that Pareto made.

****

An example power-law graph that demonstrates ranking of popularity. To the right is the [long tail](https://en.wikipedia.org/wiki/Long_tail), and to the left are the few that dominate (also known as the [80–20 rule](https://en.wikipedia.org/wiki/Pareto_principle)).

The **power law** (also called the scaling law) states that a relative change in one quantity results in a proportional relative change in another. The simplest example of the law in action is a square; if you double the length of a side (say, from 2 to 4 inches) then the area will quadruple (from 4 to 16 inches squared). A power law distribution has the form Y = k Xα, where:

* X and Y are variables of interest,
* α is the law’s exponent,
* k is a constant.

Any inverse relationship like Y = X-1 is also a power law, because a change in one quantity results in a negative change in another.