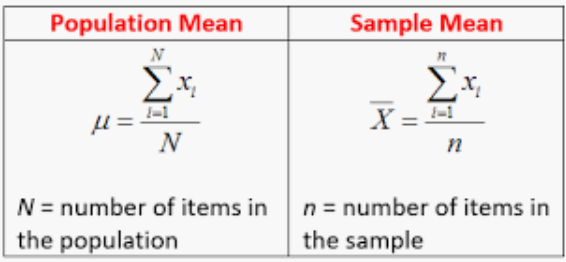
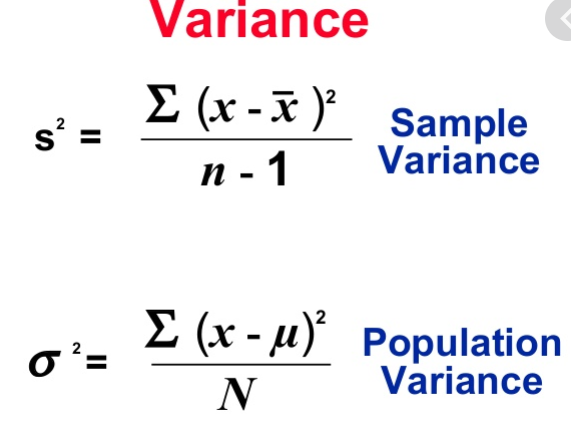
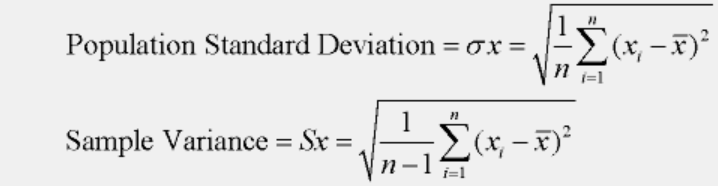
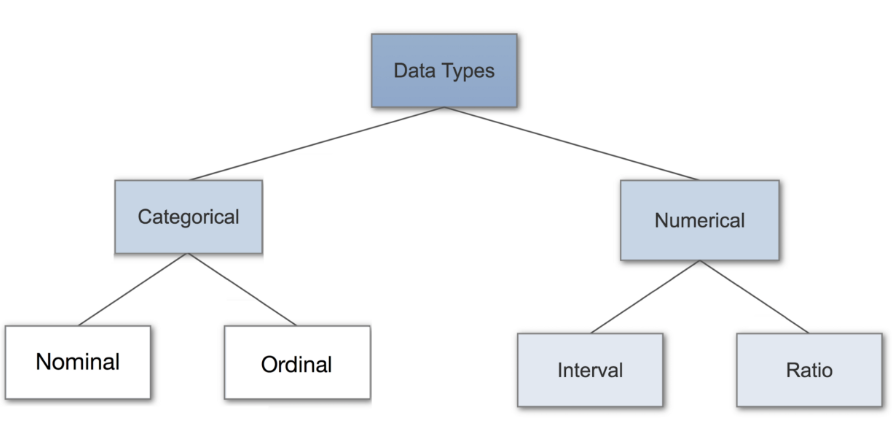
**Population (Parameters) and sample (Statistics) Symbols and formulas:**





**Different types of data:**

****

# 1. Categorical Data

Categorical data represents characteristics. Therefore, it can represent things like a person’s gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don’t have mathematical meaning.

**1.1.** **Nominal Data:** Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as „labels“. Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. Such as sex, languages, etc.

**1.2 Ordinal Data:** Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it’s ordering matters. Such as year, level of educations, etc. Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

**2. Numerical data**

# 2.1. Discrete Data: We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data ****can’t be measured but it can be counted****. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips. You can check by asking the following two questions whether you are dealing with discrete data or not: Can you count it and can it be divided up into smaller and smaller parts?

# 2.2. Continuous Data: Continuous Data represents measurements and therefore their values ****can’t be counted but they can be measured****. An example would be the height of a person, which you can describe by using intervals on the real number line.

**2.3. Interval Data:** Interval values represent **ordered units that have the same difference**. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. The problem with interval values data is that they **don’t have a „true zero“**. That means in regards to our example, that there is no such thing as no temperature. With interval data, we can add and subtract, but we cannot multiply, divide or calculate ratios. Because there is no true zero, a lot of descriptive and inferential statistics can’t be applied.

**2.4. Ratio Data:** Ratio values are also ordered units that have the same difference. Ratio values are**the same as interval values, with the difference that they do have an absolute zero**. Good examples are height, weight, length etc.

# Statistical Methods

# 1. Nominal Data

When you are dealing with nominal data, you collect information through:

**1.1. Frequencies**: The Frequency is the rate at which something occurs over a period of time or within a dataset.

**1.2. Proportion**: You can easily calculate the proportion by dividing the frequency by the total number of events. (e.g how often something happened divided by how often it could happen)

**1.3. Percentage.**

**1.4. Visualisation Methods:** To visualise nominal data you can use a pie chart or a bar chart.

**Note:** In Data Science, you can use one hot encoding, to transform nominal data into a numeric feature.

# 2. Ordinal Data

When you are dealing with ordinal data, you can use the same methods like with nominal data, but you also have access to some additional tools. Therefore you can summarise your ordinal data with frequencies, proportions, percentages. And you can visualise it with pie and bar charts. Additionally, you can use percentiles, median, mode and the interquartile range to summarise your data.

**Note:** In Data Science, you can use one label encoding, to transform ordinal data into a numeric feature.

# 3. Continuous Data

When you are dealing with continuous data, you can use the most methods to describe your data. You can summaries your data using percentiles, median, interquartile range, mean, mode, standard deviation, and range.

**Visualization Methods:** To visualize continuous data, you can use a histogram or a box-plot. With a histogram (we use buckets or bins (range of data) to build histogram), you can check the central tendency, variability, modality, and kurtosis of a distribution. Note that a histogram can’t show you if you have any outliers. This is why we also use box-plots.

**Mean, Median and Mode**

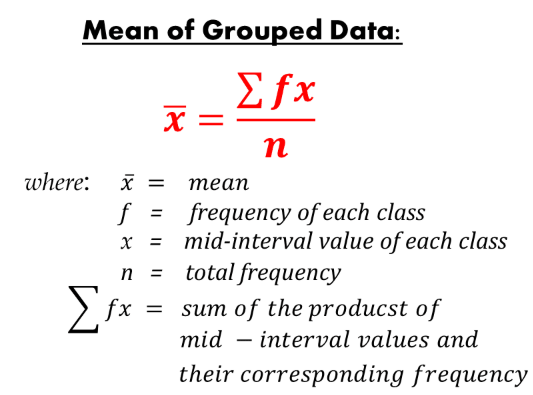
**1. Arithmetic Mean:** it is the average of all observation

**1.1. Trimmed Mean:** the mean can be influenced by extreme observations that are different from the rest of observation (*before jump to the trimmed mean first double check your data and make sure there is no data error or recording error*).

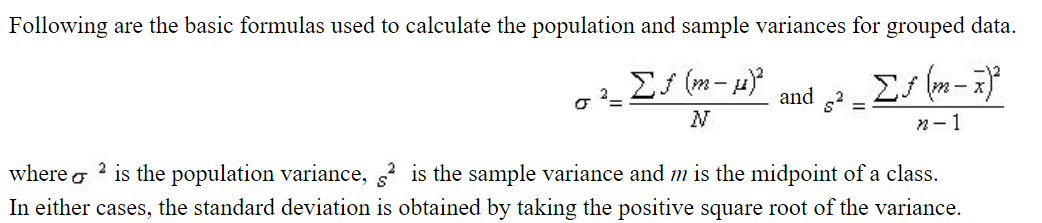
* For calculating the trimmed mean, 1) sort smallest to largest, 2) Remove the same number of observation or % from both ends 3) calculate the new mean.

**1.2. Mean of Group Frequencies:** sometimes you may not have all the original data. And only have a frequency count table or histogram for certain groups or bins of data. You can approximate the mean of your data by using midpoint method (*the higher the count in each bin and the narrower the bin range, the better the approximation*)

* **Midpoint method (upper+lower/2):** since we don’t know each individual data what we can do is the midpoint method. 1) multiply each midpoint by its frequency (weighting process), 2) sum all values from the first step, 3) divide the sum by total number of observations.



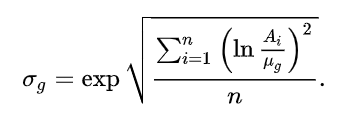
* **Variance and standard deviation of group frequency (*f*):**

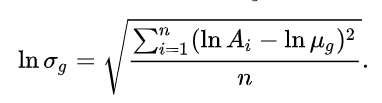
****

**1.3.** **Geometric Means:** as the name suggests, is used to calculate the **geometric mean** of a set of numbers. To recall, the **geometric mean** (or GM) is a type of **mean** that indicates the central tendency of a set of numbers by using the product of their values. It is defined as the nth root of the product of n numbers.  It should be noted that you cannot calculate the geometric mean from the arithmetic mean. In statistics, the geometric mean is well defined only for a positive set of real numbers. Example of using the formula for the geometric mean is to calculate the central frequency f0 of a bandwidth BW= f2–f1.



**Note1:** the geometrical mean mostly uses in Finance and business. But any major that use growth rate can use geometric mean.

**Note 2**: we can use the geometric mean for average growth rate over a period (e.g. compound annual growth rate in finance); using **Natural Logarithm Method**.



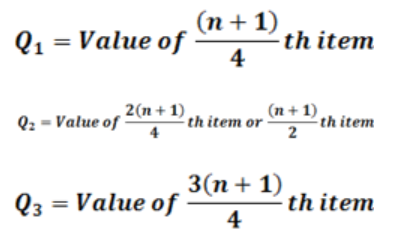
**Note 3:** the geometric mean is suitable for multiplicative process. All values must be positive. And any rate of change over sequential periods should not be used; meaning always keep periods standard (days, weeks, trials, etc.). In Excel you can use **GEOMEAN** formula.

**2. Median:** the middle number of ascending sorted data

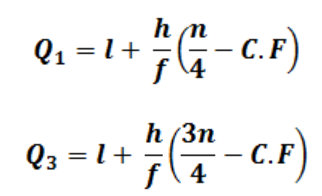
**3. Mode:** the observation that occurs most often in the data. A dataset can have one mode, multiple modes, or none.

**Percentiles & Quartiles:**

## **1. Quartiles:** The values which divide an array (a set of data arranged in ascending or descending order) into four equal parts are called Quartiles. The first, second and third quartiles are denoted by Q1, Q2,Q3 respectively. The first and third quartiles are also called the lower and upper quartiles respectively. The second quartile represents the median, the middle value.



* **Grouped Quartiles:** Q2 is our Median.



*l* = lower class boundary of the class containing the [clip_image038](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0382.png), i.e. the class corresponding to the cumulative frequency in which n/4 or 3n/4 lies

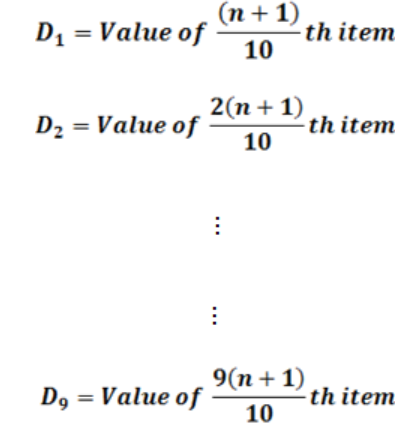
*h* = class interval size of the class containing[clip_image040](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0402.png).

*f*= frequency of the class containing [clip_image038[1]](https://econtutorials.com/wp-content/uploads/2015/07/clip_image03811.png).

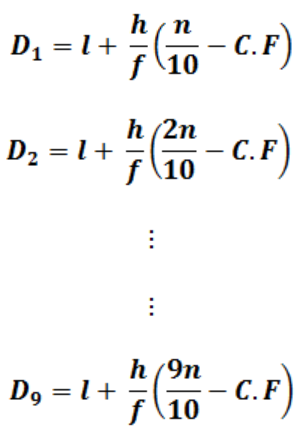
*n*= number of values, or the total frequency.

*C.F* = cumulative frequency of the class preceding the class containing [clip_image038[2]](https://econtutorials.com/wp-content/uploads/2015/07/clip_image03821.png).

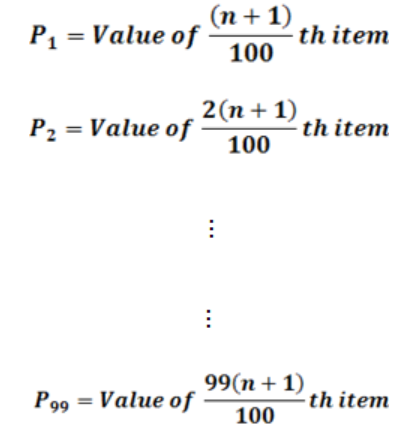
**2. Deciles:** The values which divide an array into ten equal parts are called deciles. The first, second,…… ninth deciles by [clip_image064](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0641.png) respectively. The fifth decile ([clip_image066](https://econtutorials.com/wp-content/uploads/2015/07/clip_image0661.png) corresponds to median. The second, fourth, sixth and eighth deciles which collectively divide the data into five equal parts are called **quintiles**.

****

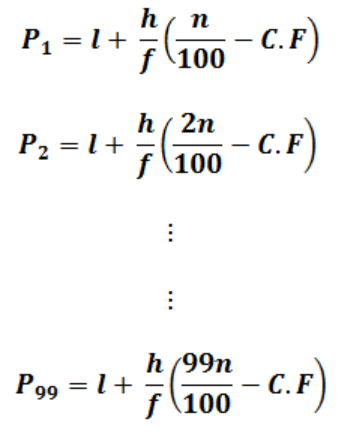
* **Grouped Deciles:**



**3. Percentiles:** The values which divide an array into one hundred equal parts are called percentiles. The first, second,……. Ninety-ninth percentile are denoted by [clip_image130](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1301.png) The 50th percentile ([clip_image132](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1321.png)) corresponds to the median. The 25th percentile [clip_image134](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1342.png) corresponds to the first quartile and the 75th percentile [clip_image136](https://econtutorials.com/wp-content/uploads/2015/07/clip_image1361.png) corresponds to the third quartile.

****

* **Grouped Percentiles:**

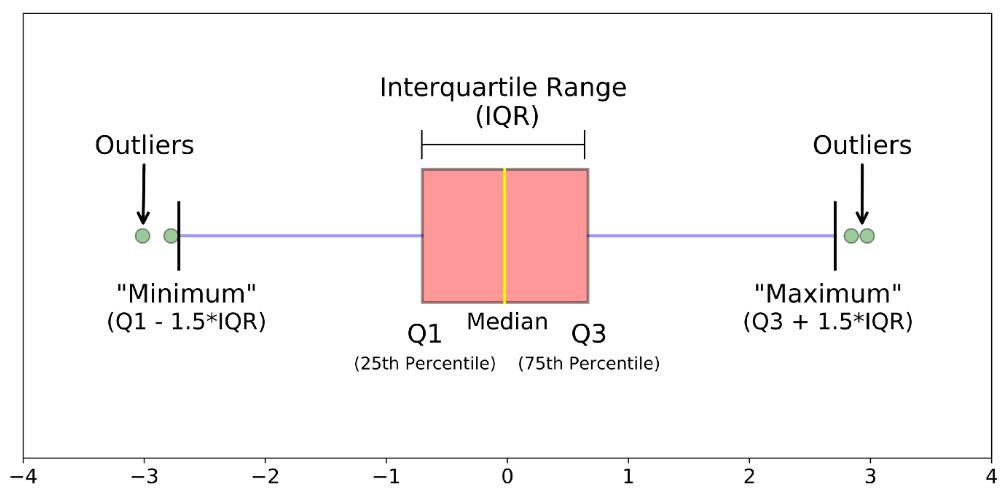
****

**V.imp Note 1:** for finding the location of data based on its percentile you can use the following formula. **LP = (p/100) \* (n+1).** Where **L** is a location in a sorted data. **P** is the percentile you are looking for, and **n** is the total number of observations. In Excel (=PERCENTILE.EXE())

**V.imp Note 2:** for finding the percentile of an observation we can use the following formula **P=(x+o.5y)/n**. where, **x** is the number of observations up to but not including the value of the nth value. **Y** is the count of observations equal to the value of desired observation. **N** is the total number of observations. Then round up to the next whole number.

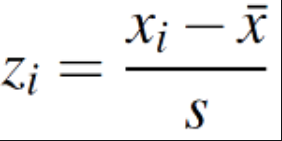
**IQR and Box Plots**

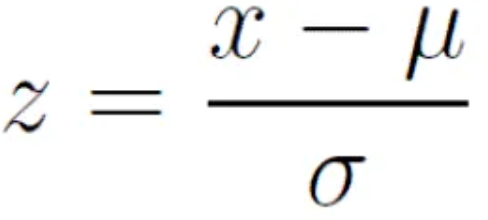
The "interquartile range", abbreviated "IQR", is just the width of the box in the box-and-whisker plot. That is, IQR = Q3 – Q1 . The IQR can be used as a measure of how spread-out the values are. Statistics assumes that your values are clustered around some central value. The IQR tells how spread out the "middle" values are; it can also be used to tell when some of the other values are "too far" from the central value. These "too far away" points are called "outliers", because they "lie outside" the range in which we expect them. The IQR is the length of the box in your box-and-whisker plot. An outlier is any value that lies more than one and a half times the length of the box from either end of the box.

****

**Z-Score**

**Why Z-Score?** Sometimes we may have two sets of data that have same mean but with different distribution, in this case the ”mean” alone can’t help us so we use z-score to understand the variation of our data.

Z-score is a measure of distance from the mean. We set the mean to “ZERO” like a starting point because it is ZERO distance from itself. We measure distance from the mean by the number of STD away.



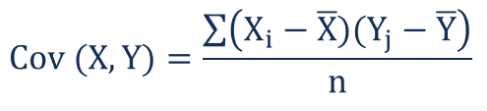
**Coefficient of Variant:** is a relative measure of variability (**Std/Mean\*100**). It measure the Std relative to mean “How large is Std relative to mean?”. Since it’s a percentage, it is unit independent and helps us to compare dataset that have different mean and Std.

**Covariance and Correlation**

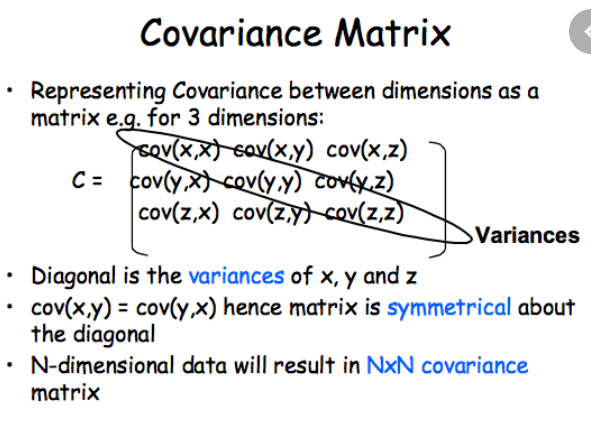
**1. Covariance:** In mathematics and [statistics](https://corporatefinanceinstitute.com/resources/knowledge/basic-statistics-concepts/), covariance is a measure of the relationship between two random variables. It tells us how ‘two variables behave as PAIR?’ The metric evaluates how much – to what extent – 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. Unlike the correlation coefficient, covariance is measured in units. The units are computed by multiplying the units of the two variables. The variance can take any positive or negative values. The values are interpreted as follows:

* **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.

**Note:** covariance only shows the direction of two variables not the strength of their relationship. For interpreting the relationship between two variables using covariance; only look at its ***SIGN*** (+ or -), ***IGNORE*** the ***SIZE*** of the number

****

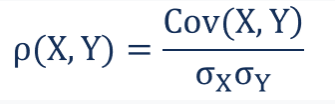
**1.1 Covariance Matrix:**



**Note 1 (V.V.IMP):** the first thing we should do to compare relationship of variables is to run **Matrix of Scatter Plots**. A figure that plots each variable against every other variable. We should search for clear linear relationship between two variables. This would help us to find out what we should expect from our analysis later on.

**Note 2:** Excel computes covariance using Population covariance formula which has a dominator of n instead of n-1 (sample). Hence, we should match the Excel result to proper statistical software by multiplying each cell by (n/n-1).

**2. Correlation:** Covariance provides only the direction of relationship between two variables while correlation provides directions and strength. Correlation always between -1 and +1 and its scale is *independent* of the scale of the variables.



**Note 1:** before going crazy computing correlations look at a scatterplot of your data. What pattern (if any) does it exhibits?

**Note 2:** Correlation is only applicable to LINEAR relationship.

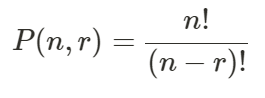
**Note 3:** Correlation is **Not** Causation.

**Note 4:** Correlation strength doesn’t necessarily mean the correlation is statistically significant; related to sample size.

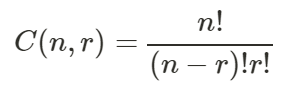
**2.1. Testing the significance of correlation (rule of thumb):** if |p(x,y)| >= 2/sqrt (n) then the relationship exist.

**Introduction to Probability**

**1. Permutation (Individual order):** the number of different ways that a certain number of objects can be *arranged in order* from a larger number of objects. If there are **n** objects, how many different ways can we make ordered list of size **r (Order Matters!)**



**2. Combinations (Unordered group or set):** the number of different ways that a certain number of objects **As a Group** can be selected for larger number of objects. Often said “**n** choose **r**”. in other words, if there are **n** objects, how many different ways can we select **group or sets** of size **r**.



**2.1. Combination-Nearly Normal:** as we increase our group size and then calculate the frequency of each possible combination where 0 =< r =< n our histogram begins to look like a normal (bell) curve. That is where we make jump from discrete random variables like C (n,r) where X=r to contains random variables; histogram to continuous function (bell curve).

* What we think of as the “normal curve” is actually a derivation or byproduct of counting and its associated probabilities
* Using the mean and Std of continuous random variable along with the fact that we know the area under the normal curve (=1), we can find out all kind of interesting stuff.

**3. Probability of Multiple Random Variables:** In majority of cases, we are likely to work with many random variables. For example, given a table of data, such as in excel, each row represents a separate observation or event, and each column represents a separate random variable. Variables may be either discrete, meaning that they take on a finite set of values, or continuous, meaning they take on a real or numerical value. As such, we are interested in the probability across two or more random variables. This is complicated as there are many ways that random variables can interact, which, in turn, impacts their probabilities. This can be simplified by reducing the discussion to just two random variables (X=A, Y=B), although the principles generalize to multiple variables.

* **Joint Probability**: Probability of events *A* and *B ->* P(A and B) = P(A|B) x P(B) = P(B|A) \* P(A)
* **Marginal Probability**: Probability of event X=*A* given variable *Y (*P(X=A) = sum P(X=A, Y=yi) for all y*)*.
* **Conditional Probability**: Probability of event *A* given event *B (P(A and B)/P(B))*.

**Discrete Probability Distribution**

**1. Random Variables:** is a variable takes on ***numerical value*** as a result of a random experiment or measurement; associate a numerical value with each possible outcome. Two things to keep separate in mind: 1) the random variable itself (random). 2) the possible outcomes or values the random variable can take (not random).

**1.1. Discrete Vs. Continuous Random Variable:** a very important distinction needs to be made between discrete and continuous random variables.

* A **Discrete random variable** has a finite number of values or an infinite sequence of values (0,1,2,…) **AND** the difference between the outcomes are meaningful (e.g. Dice). The probability distribution for a random variable X describe how probabilities are assigned to each outcome for the random variables. The discrete probability is always a number between 0 and 1. And the sum of all probabilities should be equal to 1.
* A **Continuous random variable** has a nearly infinite number of outcomes that ***can’t*** be easily counted **AND** the difference between the outcomes are not meaningful (e.g. with average income, the difference between 40,000 and 40,001 is not meaningful).

**1.2. Conditions for Discrete Probabilities:**

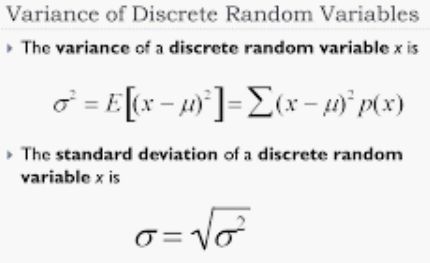
* The probability function P(x) denotes the associated probability for each outcome of the random variables
* 0 ≤ P(x) ≤ 1
* Σ P(x) = 1

**1.3. Probability Distribution:** describes how probabilities are assigned to each outcome for the random variable.

**1.2. Expected Value:** is simply the mean of random variables; the average expected outcome. It does not have to be a value the discrete variable can assume.

**E (X) = µ = ∑ X\*P(X) (a weighted average)**

**1.3. Variance of Discrete Random Variable:** though expected value tells us the mean of random variable; often times we need to know the variability, or how spread out the random variable is from the mean. We can use variance and Std to learn about how **dispersed** it is relative to its mean. This information can be used to calculate other statistics, compare datesets, and inform other conclusion about our data.



**2. The Binomial Distribution:** the **binomial distribution** with parameters ***n*** and ***p*** is the [discrete probability distribution](https://en.wikipedia.org/wiki/Discrete_probability_distribution) of the number of successes in a sequence of *n* [independent](https://en.wikipedia.org/wiki/Statistical_independence) [experiments](https://en.wikipedia.org/wiki/Experiment_(probability_theory)), each asking a [yes–no question](https://en.wikipedia.org/wiki/Yes%E2%80%93no_question)

**2.1. Binomial Experiment**: has below characteristics:

* The process consists of a sequence of ***n*** trials.
* Only two exclusive outcomes are possible in each trial. One outcome is called a “success ([yes](https://en.wikipedia.org/wiki/Yes_and_no)/[true](https://en.wikipedia.org/wiki/Truth_value)/[one](https://en.wikipedia.org/wiki/One))” and the other a “failure ([no](https://en.wikipedia.org/wiki/Yes_and_no)/[false](https://en.wikipedia.org/wiki/False_(logic))/[zero](https://en.wikipedia.org/wiki/Zero))”.
  + “success” and “failure” can be counterintuitive at times if taken literally.
* The probability of success (***p***) does not change from trial to trail. The probability of failure is ***q=1-p*** and is also fixed from trial to trial.
* The trials are independent; the outcomes of previous trials not influence future trials.
* The sum of all possible outcomes should be equal to 1.
* The probability of any given outcome is a combination of both the number of trials and the success rate **Bionmcdf: P(x) = C (n, x) Px (1-P)n-x**

**2.2 Binomial Mean (Expected Value) and Standard Deviation:**

* **µ = n.p**
* 

**2.3. Extra information about Binomial Distribution:** A single success/failure experiment is also called a [Bernoulli trial](https://en.wikipedia.org/wiki/Bernoulli_trial) or Bernoulli experiment and a sequence of outcomes is called a [Bernoulli process](https://en.wikipedia.org/wiki/Bernoulli_process); for a single trial, i.e., *n* = 1, the binomial distribution is a [Bernoulli distribution](https://en.wikipedia.org/wiki/Bernoulli_distribution). The binomial distribution is the basis for the popular [binomial test](https://en.wikipedia.org/wiki/Binomial_test) of [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance). The binomial distribution is frequently used to model the number of successes in a sample of size *n* drawn [with replacement](https://en.wikipedia.org/wiki/With_replacement) from a population of size *N*. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a [hypergeometric distribution](https://en.wikipedia.org/wiki/Hypergeometric_distribution), not a binomial one. However, for *N* much larger than *n*, the binomial distribution remains a good approximation, and is widely used.

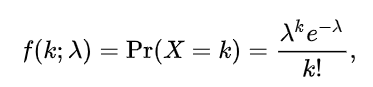
**Note:** In many cases we can use the reverse way which is easier to calculate (1 – the amount we don’t look for). In excel you can use =BINOMDIST (Number of suceesses; Total number of trials; probability of success; True/False). True will calculate Binomcdf and False will calculate Binompdf.

**How to interpret the probability of binomial distribution (v.v.imp):** the probability of a scenario give us the level of error in our hypothesis (assumption) based on random chance variation. If the probability cover more than 95% the result is not by accident and you fail to reject your hypothesis.

**3. Poisson Distribution:**  is a [**discrete probability distribution**](https://en.wikipedia.org/wiki/Discrete_probability_distribution) that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and [**independently**](https://en.wikipedia.org/wiki/Statistical_independence) of the time since the last event.[[1]](https://en.wikipedia.org/wiki/Poisson_distribution#cite_note-Haight1967-1) The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day. If receiving any particular piece of mail does not affect the arrival times of future pieces of mail, i.e., if pieces of mail from a wide range of sources arrive independently of one another, then a reasonable assumption is that the number of pieces of mail received in a day obeys a Poisson distribution.[[2]](https://en.wikipedia.org/wiki/Poisson_distribution#cite_note-Brooks2007-2) Other examples that may follow a Poisson distribution include the number of phone calls received by a call center per hour.

**Landa = (number of occurrences)/interval** ; 



**3.1. Poisson characteristics:**

* Discrete outcomes (x=0,1,2,3, …)
* The number of occurrence in each interval can range from 0 to infinity.
* Describe the distribution of infrequent (rare) events
* Each event is independent of other events.
* Describes discrete events over an interval (time, distance, etc.)
* E(X) is assumed to be constant throughout the Experience

**Note:** In excel you can use =POISSON:DIST (X=K; Mean=Landa; True/False). True will calculate possioncdf and False will calculate possionpdf.