[**https://www.edx.org/course/statistical-thinking-data-science-columbiax-ds101x-0**](https://www.edx.org/course/statistical-thinking-data-science-columbiax-ds101x-0)

**Random Variables:**

are variables whose values are subject to variations due to chance. A random variable can take on a set of possible different values (similarly to other mathematical variables), each with an associated [probability](https://en.wikipedia.org/wiki/Probability). Or ways to map outcomes of random process to numbers i.e. quantifying the outcomes.

\*\* Always generates numerical values \*\*

**Types of Random Variables :**

The mathematical function describing the possible values of a random variable and their associated probabilities is known as a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution). *Random variables* can be [***discrete***](https://en.wikipedia.org/wiki/Discrete_random_variable), that is, taking any of a specified finite or countable list of values, endowed with a [probability mass function](https://en.wikipedia.org/wiki/Probability_mass_function), characteristic of a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution); or [***continuous***](https://en.wikipedia.org/wiki/Continuous_random_variable), taking any numerical value in an interval or collection of intervals, via a [probability density function](https://en.wikipedia.org/wiki/Probability_density_function)

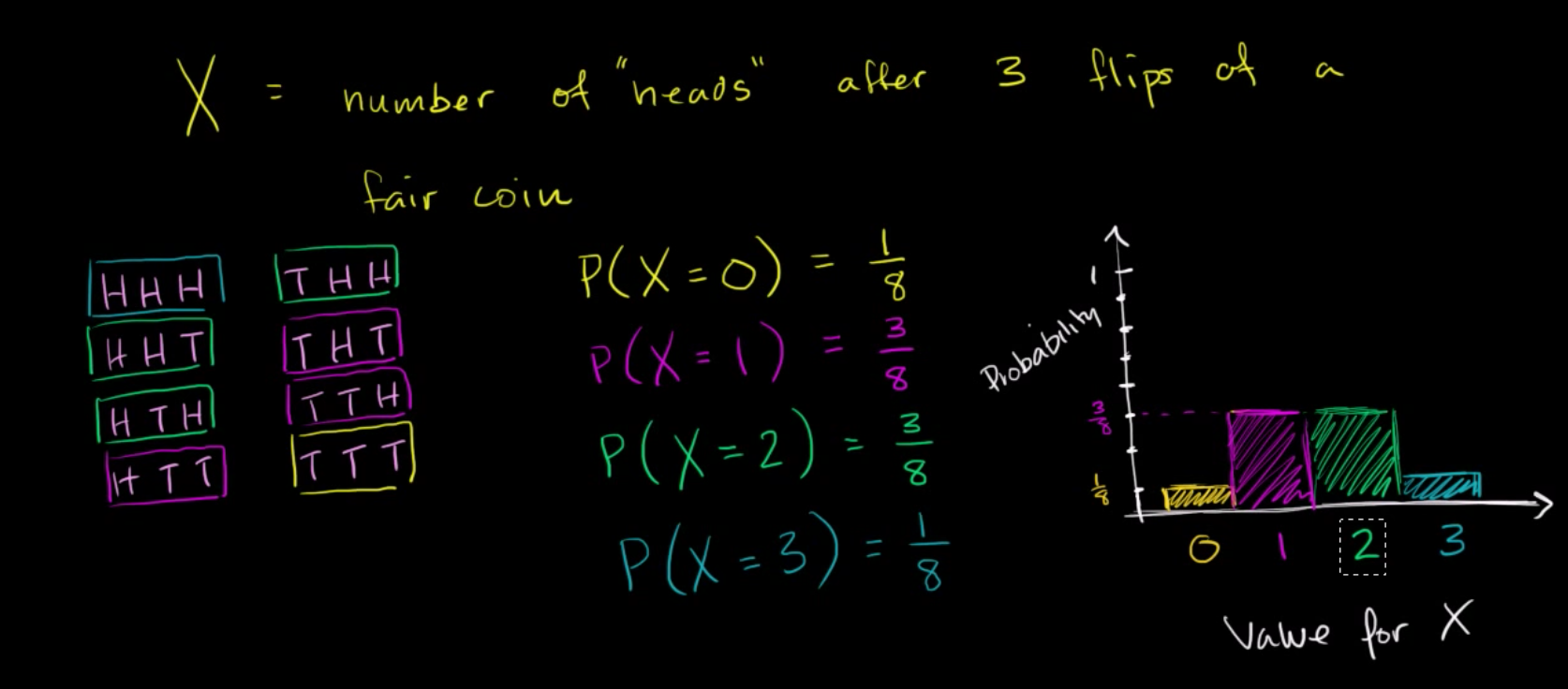
### **Discrete random variable**: ( countably infinite many or finite many)

In an experiment a person may be chosen at random, and one random variable may be the person's height. Mathematically, the random variable is interpreted as a function which maps the person to the person's height. Associated with the random variable is a probability distribution that allows the computation of the probability that the height is in any subset of possible values, such as the probability that the height is between 180 and 190 cm, or the probability that the height is either less than 150 or more than 200 cm.

Another random variable may be the person's number of children; this is a discrete random variable with non-negative integer values. It allows the computation of probabilities for individual integer values – the probability mass function (PMF) – or for sets of values, including infinite sets. For example, the event of interest may be "an even number of children". For both finite and infinite event sets, their probabilities can be found by adding up the PMFs of the elements; that is, the probability of an even number of children is the infinite sum

Can be quantified by integer values or rounded real values to some decimal place

An example of probability distribution of discrete RV (Probability Mass Function) :

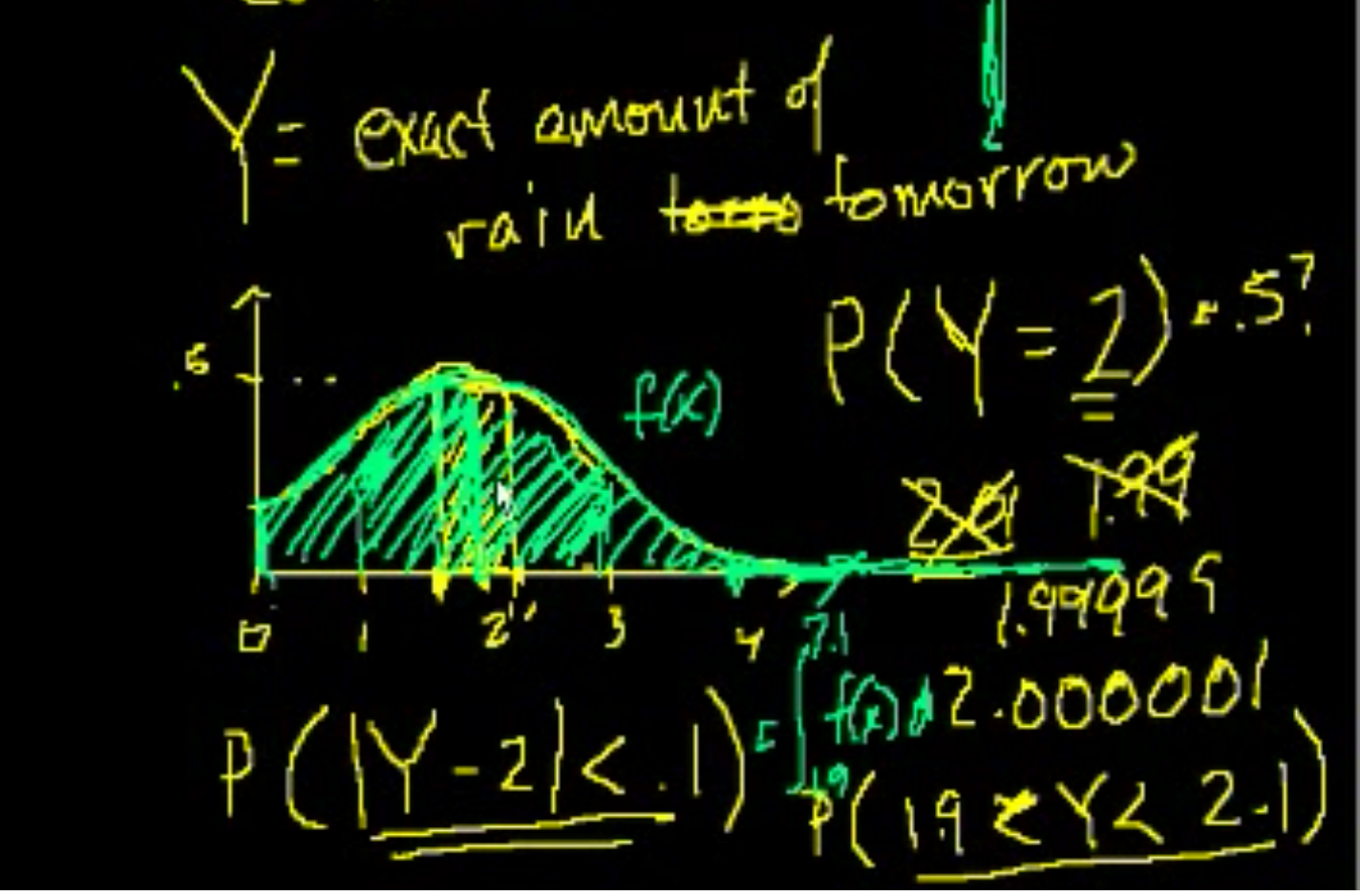


### **Continuous random variable**: (uncountably infinite many)

An example of a continuous random variable would be one based on a spinner that can choose a horizontal direction. This can be done, for example, by mapping a direction to a bearing in degrees clockwise from North. The random variable then takes values which are real numbers from the interval [0, 360), with all parts of the range being "equally likely". In this case, ***X*** = the angle spun. Any real number has probability zero of being selected, but a positive probability can be assigned to any *range* of values. For example, the probability of choosing a number in [0, 180] is 1⁄2. Instead of speaking of a probability mass function, we say that the probability *density* of ***X*** is 1/360. The probability of a subset of [0, 360) can be calculated by multiplying the measure of the set by 1/360. In general, the probability of a set for a given continuous random variable can be calculated by integrating the density over the given set.

Probability for an exact value of continuous RV is always 0 i.e. P( Y=2) =0. We can only calculate probability in interval i.e. P( 1.9 < Y < 2.1) will be either zero or positive probability.

### Can be quantified by real values



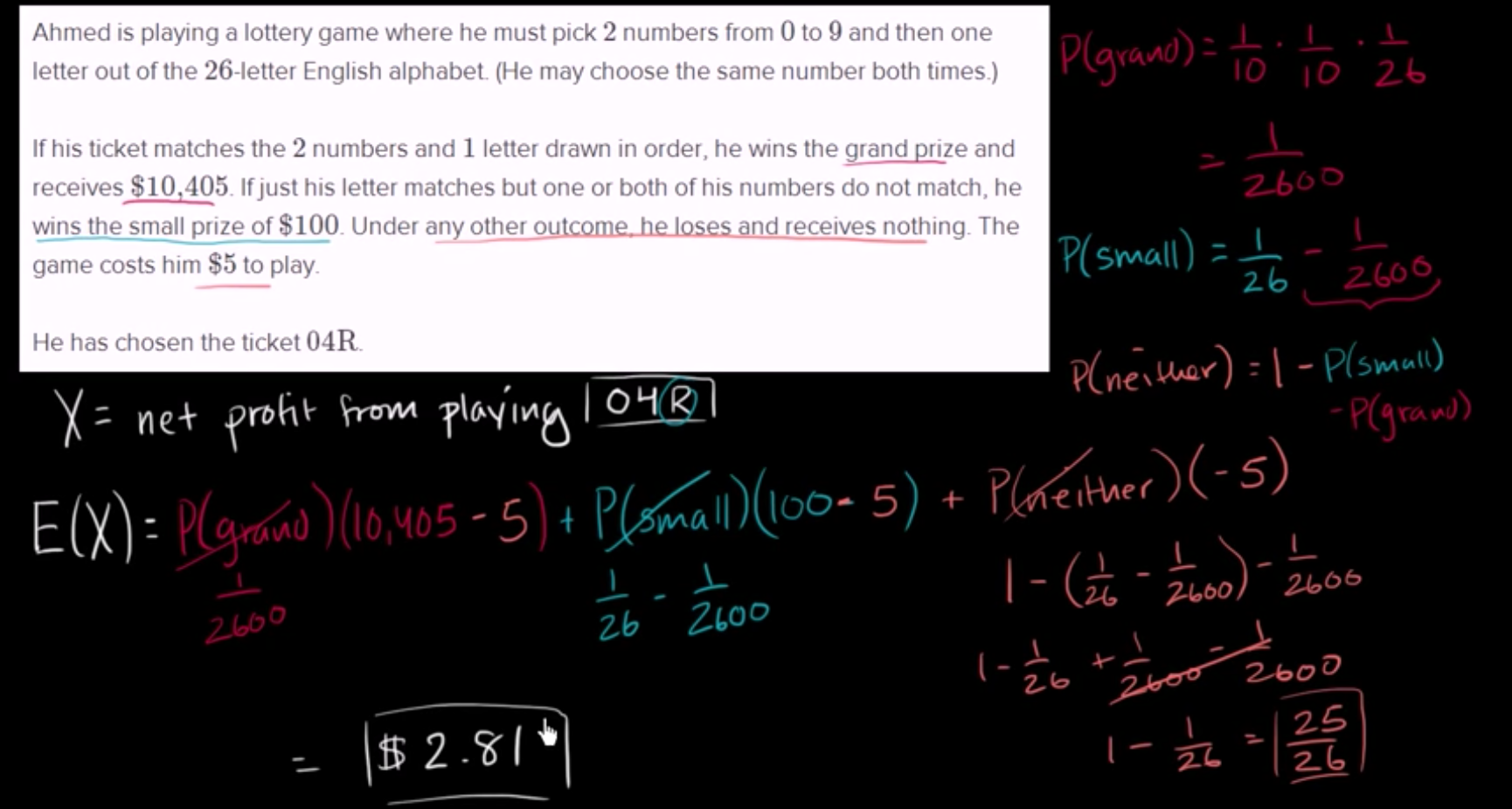
### **Mixed type :** can not be described by PMF or PDF

An example of a random variable of mixed type would be based on an experiment where a coin is flipped and the spinner is spun only if the result of the coin toss is heads. If the result is tails, ***X*** = −1; otherwise ***X*** = the value of the spinner as in the preceding example. There is a probability of 1⁄2 that this random variable will have the value −1. Other ranges of values would have half the probability of the last example.

**Expected Value :**

In [probability theory](https://en.wikipedia.org/wiki/Probability_theory), the expected value of a [random variable](https://en.wikipedia.org/wiki/Random_variable), intuitively, is the long-run average value of repetitions of the experiment it represents. For example, the expected value in rolling a six-sided [die](https://en.wikipedia.org/wiki/Dice) is 3.5 because, roughly speaking, the average of all the numbers that come up in an extremely large number of rolls is very nearly always quite close to three and a half. Less roughly, the [law of large numbers](https://en.wikipedia.org/wiki/Law_of_large_numbers) states that the [arithmetic mean](https://en.wikipedia.org/wiki/Arithmetic_mean) of the values [almost surely](https://en.wikipedia.org/wiki/Almost_surely) converges to the expected value as the number of repetitions approaches infinity. The expected value is also known as the expectation, mathematical expectation, EV, average, mean value, mean, or first moment.

<https://en.wikipedia.org/wiki/Expected_value>



**Law of large numbers :**

Let’s say a random variables X, and we know it’s expected value or population mean E(X), if we take a sample of our random variable X with n observation and sample mean is =(x1+x2+...xn)/n. So Law of large numbers tells that sample mean will approach to expected value of random variable or population mean for n approaching infinity.

In [probability theory](https://en.wikipedia.org/wiki/Probability_theory), the **law of large numbers** (**LLN**) is a [theorem](https://en.wikipedia.org/wiki/Theorem) that describes the result of performing the same experiment a large number of times. According to the law, the [average](https://en.wikipedia.org/wiki/Average) of the results obtained from a large number of trials should be close to the [expected value](https://en.wikipedia.org/wiki/Expected_value), and will tend to become closer as more trials are performed.

The LLN is important because it "guarantees" stable long-term results for the averages of some random events. For example, while a casino may lose money in a single spin of the [roulette](https://en.wikipedia.org/wiki/Roulette) wheel, its earnings will tend towards a predictable percentage over a large number of spins. Any winning streak by a player will eventually be overcome by the parameters of the game. It is important to remember that the LLN only applies (as the name indicates) when a *large number* of observations are considered. There is no principle that a small number of observations will coincide with the expected value or that a streak of one value will immediately be "balanced" by the others (see the [gambler's fallacy](https://en.wikipedia.org/wiki/Gambler%27s_fallacy)).

Weak LLN:

Strong LLN:<https://en.wikipedia.org/wiki/Law_of_large_numbers>

**Gambler’s Fallacy**: if we are getting head in all previous flips it doesn’t mean that heads will come up in the nes flip or tails will come up to balance the expected value. If it is a fair coin it can be anything either head or tail.

Useful Links:

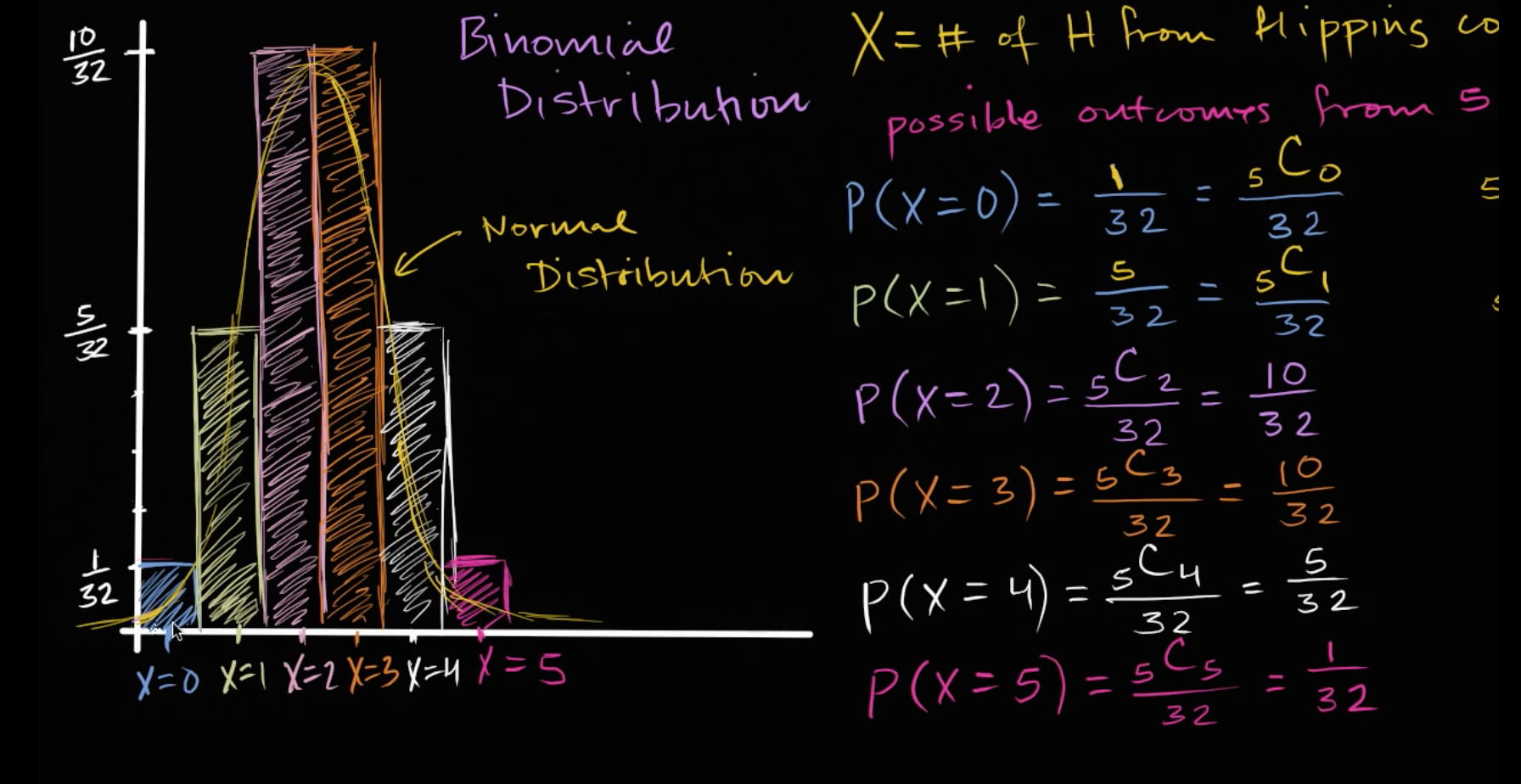
<https://en.wikipedia.org/wiki/Random_variable>

<http://www.pindling.org/Math/Statistics/Textbook/Chapter5_Discrete_Dist/discrete_dist_intro.htm>

**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) yes/no experiments, each of which yields success with [probability](https://en.wikipedia.org/wiki/Probability) *p*. A success/failure experiment is also called a Bernoulli experiment or [Bernoulli trial](https://en.wikipedia.org/wiki/Bernoulli_trial); when *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).

Example : Lets say RV, X = Number of Heads from flipping a coin 5 times



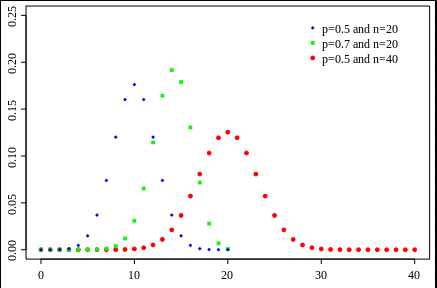
→ This is very important distribution because lot of discrete processes assumes that underlying probability distribution is binomial distribution.

→ Above, probability distribution is a PMF ( a discrete probability distribution) which looks like a bell curve (like a normal curve), so binomial distribution is a discrete type of normal distribution.

→ The binomial distribution is for two outcomes only (hence the "bi"). For cases where there are 3 or more possible outcomes, we'd use the multinomial distribution. I.e for dice

→ expected value of binomial distribution is E(X) = n.p where n = number of attempts and p is the probability of happening an event and 1-p is the probability of happening another event.

**Example:** 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/Sampling_(statistics)#Replacement_of_selected_units) 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.



**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 and/or space if these events occur with a known average 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-haight-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 doesn't 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 per day obeys a Poisson distribution.[[2]](https://en.wikipedia.org/wiki/Poisson_distribution#cite_note-umass1-2) Other examples that may follow a Poisson: the number of phone calls received by a call center per hour, the number of decay events per second from a radioactive source, or the number of pedicabs in queue in a particular street in a given hour of a day.

The number of patients arriving in an emergency room between 11 and 12 pm.

The number of meteors greater than 1 meter diameter that strike earth in a year

### **Examples that violate the Poisson assumptions:**The number of students who arrive at the student union per minute will likely not follow a Poisson distribution, because the rate is not constant (low rate during class time, high rate between class times) and the arrivals of individual students are not independent (students tend to come in groups).

### **Assumptions: When is the Poisson distribution an appropriate mode**

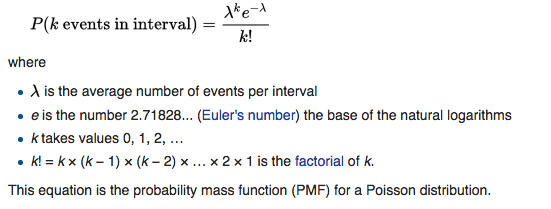
The Poisson distribution is an appropriate model if the following assumptions are true.

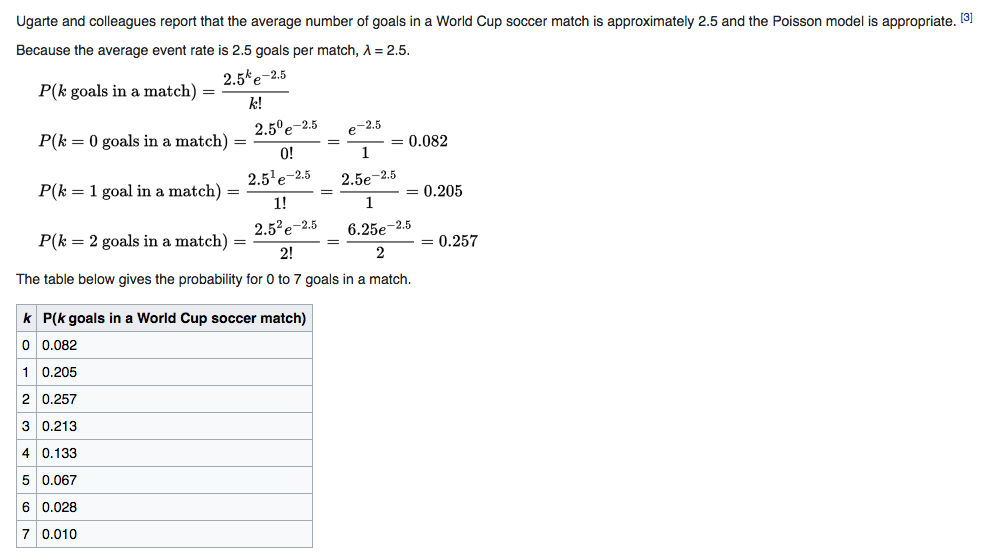
* K is the number of times an event occurs in an interval and K can take values 0, 1, 2, …
* The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
* The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.
* Two events cannot occur at exactly the same instant.
* The probability of an event in an interval is proportional to the length of the interval.

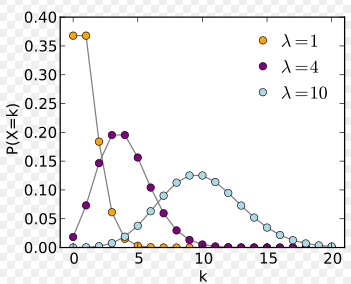
If these conditions are true, then K is a Poisson random variable, and the distribution of K is a Poisson distribution.

The average number of events in an interval is designated (lambda).

The probability of observing k events in an interval is given by the equation



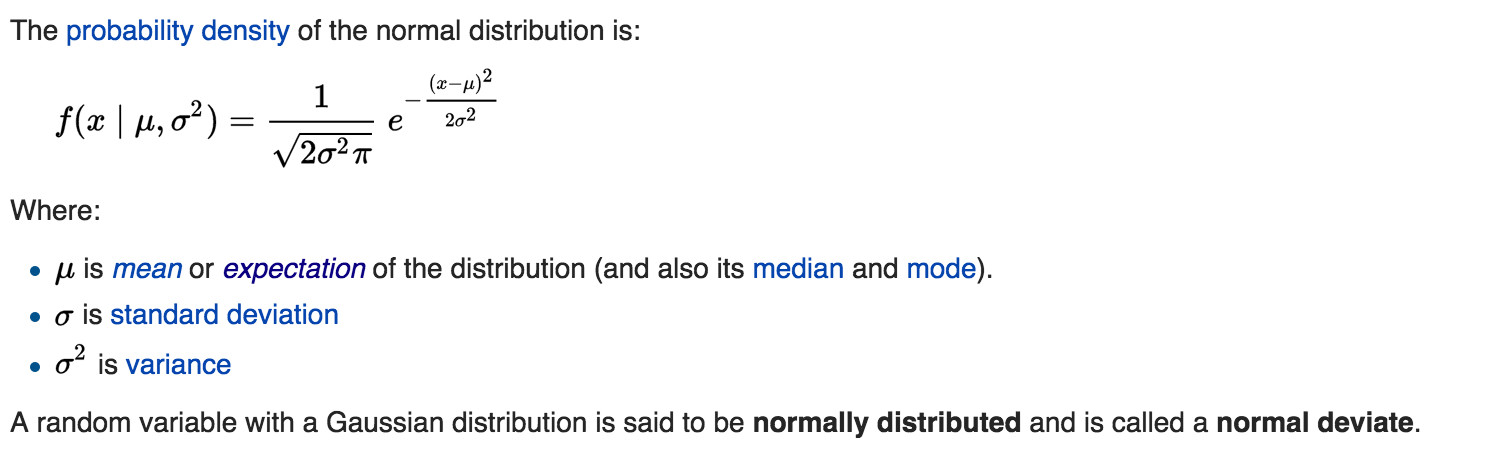




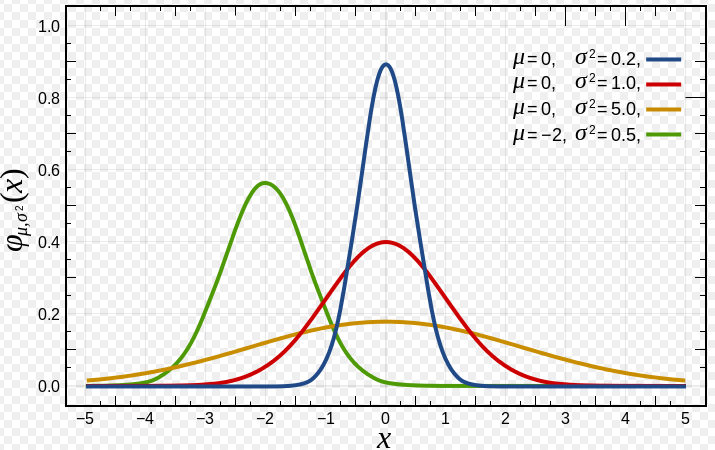
**Poisson Distribution from Binomial Distribution :**

**Normal Distribution:**

In [probability theory](https://en.wikipedia.org/wiki/Probability_theory), the **normal** (or **Gaussian**) **distribution** is a very common [continuous probability distribution](https://en.wikipedia.org/wiki/Continuous_probability_distribution). Normal distributions are important in [statistics](https://en.wikipedia.org/wiki/Statistics) and are often used in the [natural](https://en.wikipedia.org/wiki/Natural_science) and [social sciences](https://en.wikipedia.org/wiki/Social_science) to represent real-valued [random variables](https://en.wikipedia.org/wiki/Random_variable) whose distributions are not known.



The above equation gives the probability curve ( y value of graph), not the probability of particular value of ‘x’ (which is area under curve y.dx). Indeed in continuous probability distribution p(x) = 0. So you always have to find probability of a range of x i.e. p( () to () ).



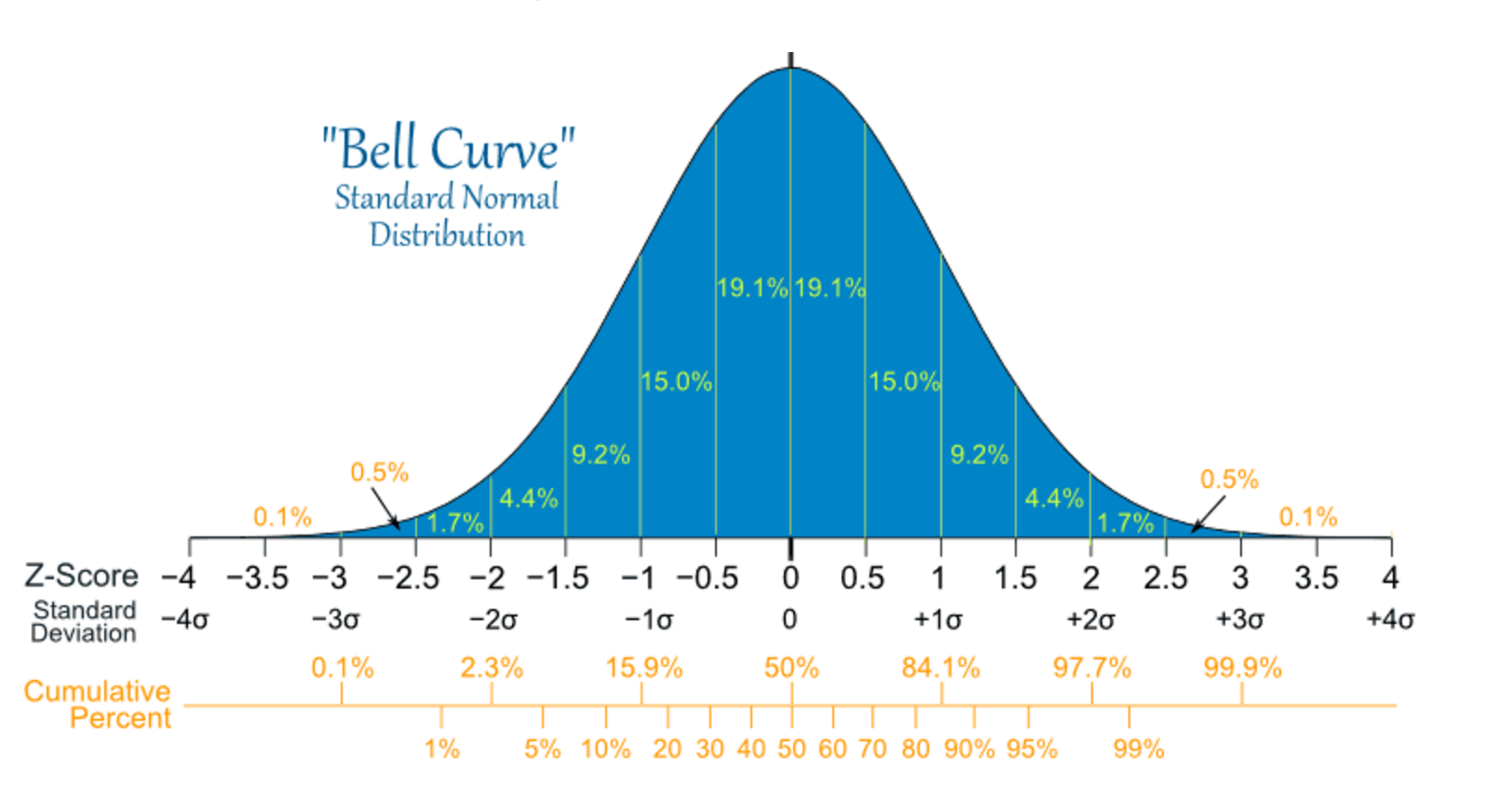
### 

### 

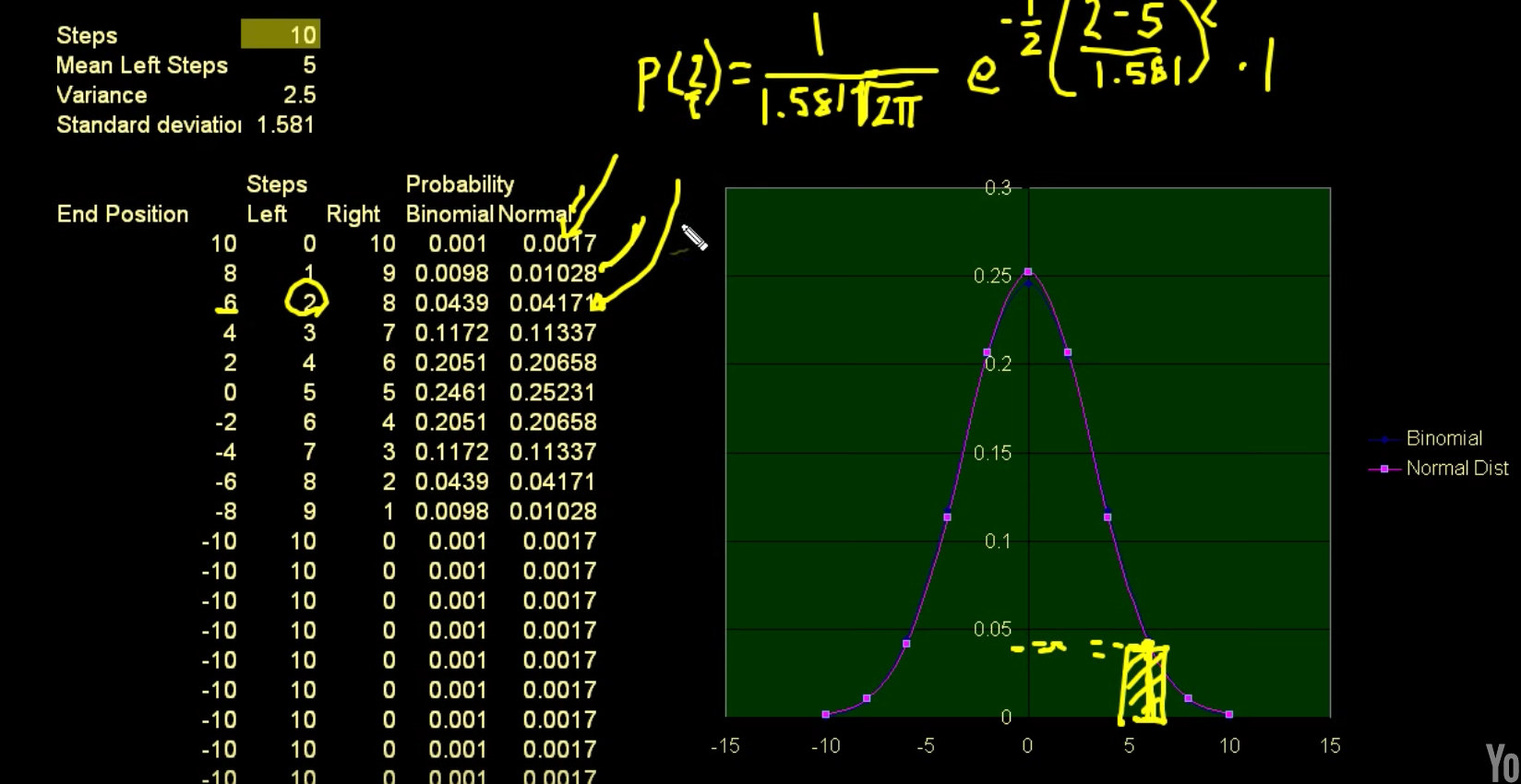
### **Standard normal distribution:**

The simplest case of a normal distribution is known as the *standard normal distribution*. This is a special case when μ=0 and σ=1, and it is described by this [probability density function](https://en.wikipedia.org/wiki/Probability_density_function):

### Screen Shot 2016-07-26 at 4.07.24 PM.png

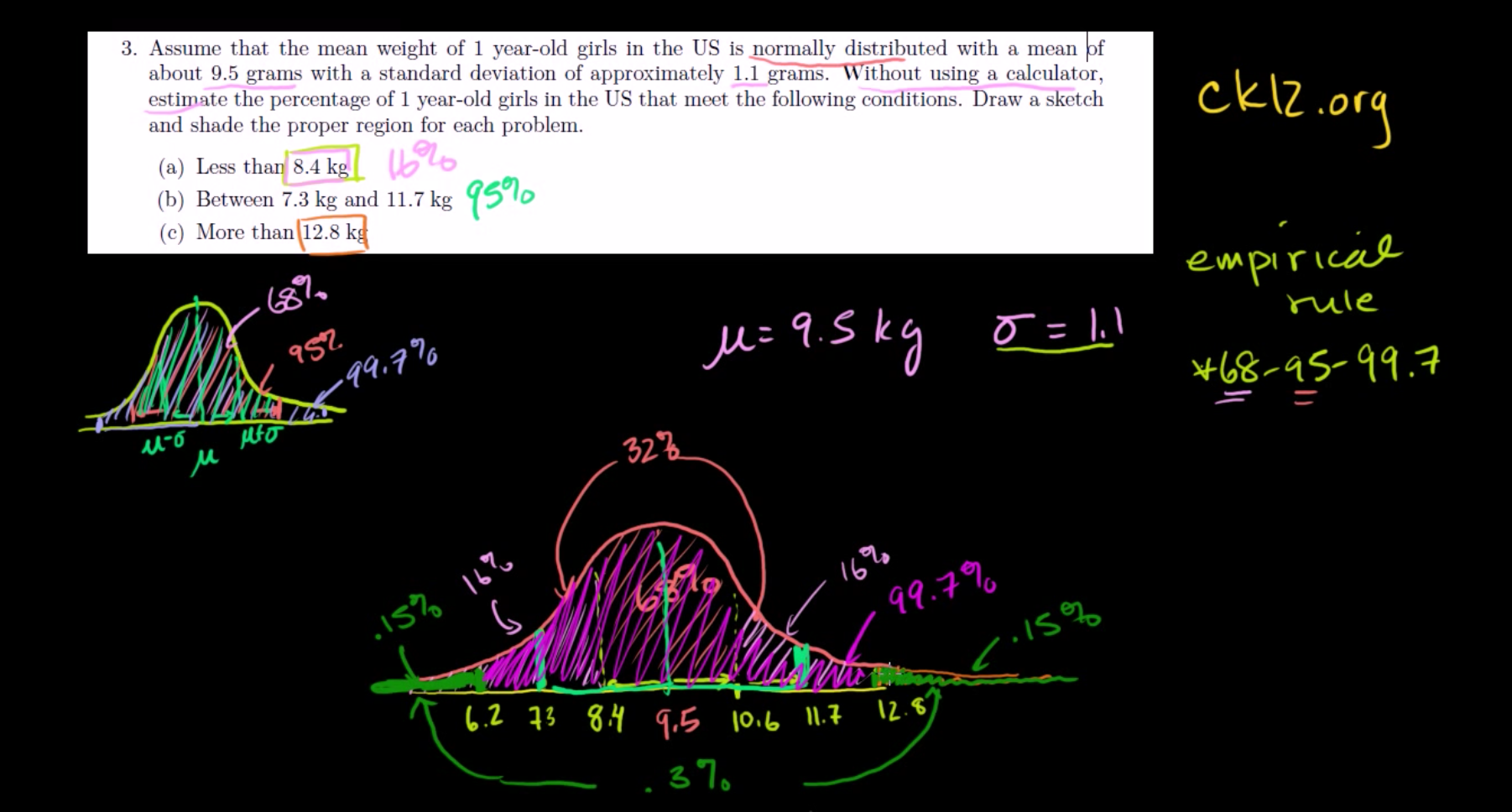


\*\*Normal distribution can arguably approximate almost all distributions. Like in following example, we can approximate binomial distribution with normal distribution



Here, if we want to find probability of 2 in normal distribution we can integrate the p(x).dx from interval 2.5 to 3.5 and can get the p(2). Or we can approximate it by multiplying p(x=2) \* 1, where 1 = 3.5-2.5

**Calculation of probability in normal distribution, empirically :**



Z-score : how many standard deviation you are away from the mean, which could be calculated for any type of distribution.

**Central Limit Theorem :** The central limit theorem states that given a distribution ( with any shape, population) with a mean μ and variance σ², the **sampling** [**distribution of the mean**](http://davidmlane.com/hyperstat/A13660.html) approaches a [normal distribution](http://davidmlane.com/hyperstat/A6929.html) with a mean (μ) and a variance σ²/N as N, the [sample size,](http://davidmlane.com/hyperstat/A104571.html)increases. The amazing and counter-intuitive thing about the central limit theorem is that no matter what the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution. Furthermore, for most distributions, a normal distribution is approached very quickly as N increases. Keep in mind that N is the sample size for each mean and not the number of samples. Remember in a [sampling distribution](http://davidmlane.com/hyperstat/A11150.html) the number of samples is assumed to be infinite. The sample size is the number of scores in each sample; it is the number of scores that goes into the computation of each mean.

So as sample size ‘N’ increases sampling distribution have smaller standard deviation. Also, if

N, sampling distribution becomes perfect normal distribution, but as sample size increase sampling distribution approached to normal quickly.

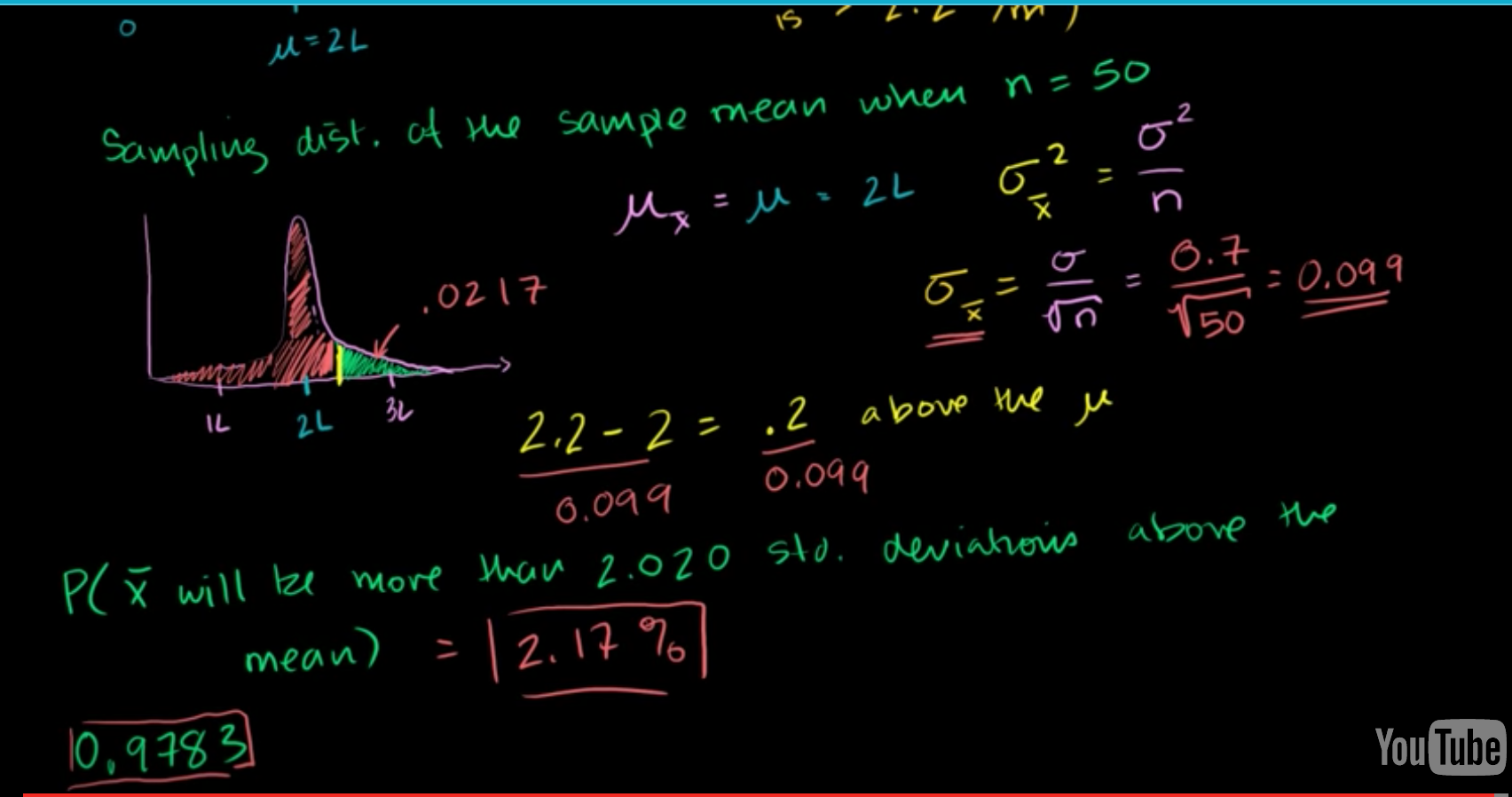
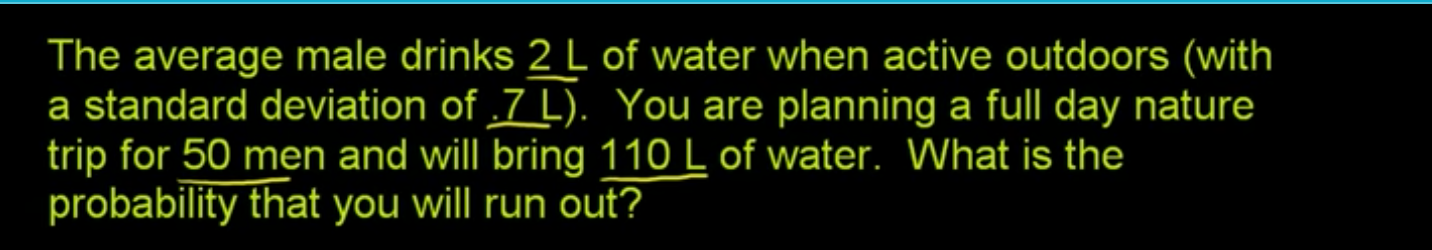
**USE : This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.**

**For the random samples we take from the population, we can compute the mean of the sample means:**

Explained with example :

<https://www.youtube.com/watch?v=Pujol1yC1_A>

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704\_Probability/BS704\_Probability12.html



**Student’s t distribution:**

**Student's *t*-distribution** (or simply the ***t*-distribution**) is any member of a family of continuous [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution) that arises when estimating the [mean](https://en.wikipedia.org/wiki/Expected_value) of a [normally distributed](https://en.wikipedia.org/wiki/Normal_distribution) [population](https://en.wikipedia.org/wiki/Statistical_population) in situations where the [sample size](https://en.wikipedia.org/wiki/Sample_size) is small and population [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) is unknown.

A t distribution has more scores in its tails and fewer in the center than a normal distribution, so it is leptokurtic. Also, a t distribution with 20 degrees of freedom has 95% of its distribution within 2.086 standard deviations of its mean.

According to the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central_limit_theorem), the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of a statistic (like a sample mean) will follow a [normal distribution](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution), as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a [z-score](http://stattrek.com/Help/Glossary.aspx?Target=Z-score), and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the **t statistic**(also known as the **t score**), whose values are given by:

t = [ x - μ ] / [ s / sqrt( n ) ]

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size. The distribution of the *t* statistic is called the **t distribution** or the **Student t distribution**.

The t distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution.

When to Use the t Distribution

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). The sampling distribution of a statistic should be bell-shaped if any of the following conditions apply.

* The population distribution is normal.
* The population distribution is [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry), [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution), without [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier), and the sample size is at least 30.
* The population distribution is moderately [skewed](http://stattrek.com/Help/Glossary.aspx?Target=Skewness), unimodal, without outliers, and the sample size is at least 40.
* The sample size is greater than 40, without outliers.

The t distribution should *not* be used with small samples from populations that are not approximately normal.

Watch this video : <https://www.youtube.com/watch?v=Uv6nGIgZMVw>

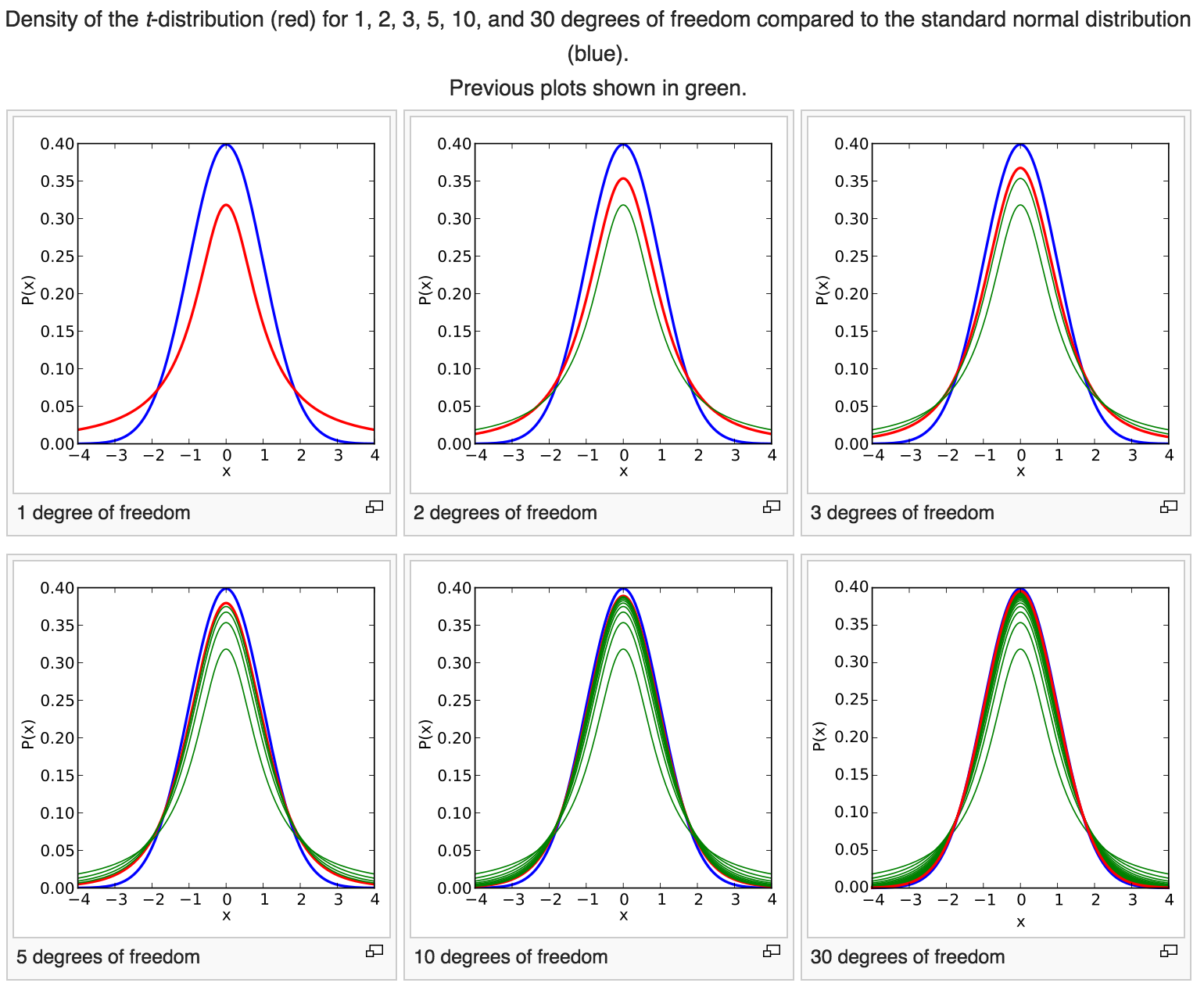
And browse : <http://stattrek.com/probability-distributions/t-distribution.aspx>

Roughly speaking, When dealing with a distribution ( of any shape) of small sample size (<30), then according to central limit theorem, the sampling distribution can not be considered normal. Instead of we can assume that sampling distribution of means have a shape of t-distribution.

Whereas a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution) describes a full population, *t*-distributions describe samples drawn from a full population; accordingly, the *t*-distribution for each sample size is different, and the larger the sample, the more the distribution resembles a normal distribution.

If we take a sample of *n* observations from a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution), then the *t*-distribution with degrees of freedom can be defined as the distribution of the location of the true mean, relative to the sample mean and divided by the sample standard deviation, after multiplying by the normalizing term . In this way, the *t*-distribution can be used to estimate how likely it is that the true mean lies in any given range.

The following images show the density of the *t*-distribution for increasing values of The normal distribution is shown as a blue line for comparison. Note that the *t*-distribution (red line) becomes closer to the normal distribution as increases.



### **Stratified Sampling**

**In stratified sampling, we split the population into non-overlapping groups or strata (e.g., men and women, people under 30 years of age and people 30 years of age and older), and then sample within each strata. The purpose is to ensure adequate representation of subjects in each stratum.**

**Sampling within each stratum can be by simple random sampling or systematic sampling. For example, if a population contains 70% men and 30% women, and we want to ensure the same representation in the sample, we can stratify and sample the numbers of men and women to ensure the same representation. For example, if the desired sample size is n=200, then n=140 men and n=60 women could be sampled either by simple random sampling or by systematic sampling.**

# **Evaluating Screening Tests:**

<http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability4.html#headingtaglink_1>

**Analysing Data:**

**Frequency Distribution:** which is a graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set. Frequency distributions come in many different shapes and sizes. In an ideal world our data would be distributed symmetrically around the centre of all scores. This is known as a **normal distribution** and is characterized by the bell-shaped curve with which you might already be familiar.

There are two main ways in which a distribution can deviate from normal:

(1) lack of symmetry (called skew )

positively skewed :the frequent scores are clustered at the lower end and the tail points towards the higher or more positives cores)

negatively skewed :the frequent scores are clustered at the higher end and the tail points

towards the lower or more negative scores

(2) pointyness (called kurtosis ) : refers to the degree to which scores cluster at the ends of the distribution

leptokurtic distribution: A distribution with positive kurtosis has many scores in the tails (a so-called heavy-tailed distribution) and is pointy.

Platykurtic distributtion : A distribution with negative kurtosis is relatively thin in the tails (has light tails) and tends to be flatter than normal

To calculate the Z score of a distribution :

Z = cuts off 5%, 1.96 cutsoff the top 2.5% of the distribution, -1.96 cuts off the bottom 2.5% distribution.

Z = +-2.58 = 1%, 99% of value, .5%-99.5%

Z = +- 3.29 = .1%

Population : The real time (real world) data generated from every process in one context.

Sample : We never have access to the entire population and so we collect smaller samples and use the behaviour

within the sample to infer things about the behaviour in the population. The bigger the sample,

the more likely it is to reflect the whole population. If we take several random samples from the

population, each of these samples will give us slightly different results. However, on average,

large samples should be fairly similar.

**The mean: a very simple statistical model**

**Assessing the fit of the mean: sums of squares, variance and standard deviations**

**Measures to Fit**

Deviance/Error : Difference between actual value and the mean. is the mean

Sum of Errors: Sum of Deviances ( Should be equal to zero in case of The Mean Model)

Sum of Squared Errors(SS) :

\*\*SS is the good measure but it depends on number of observation and increases if more the number of observation. So we can use average SS instead in which we divide SS by n-1.

Variance : . The average error between the mean and observation.

\*\* problem with Variance is it gives us measure in units squared.

Standard Deviation : It gives the error in the same unit what the observation is in. It is measure of how well the mean represents data. It also tell us about the shape of distribution of scores.

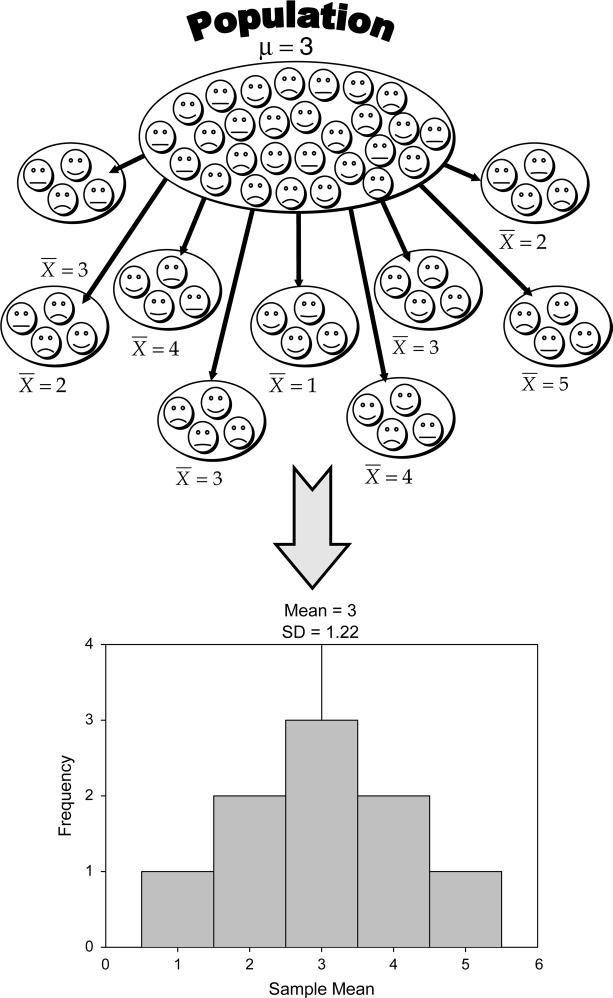
Large standard deviation: means data is not well centered around mean, mean is here not good fit for model.

Bessel’s Correction :

That is, when [estimating](https://en.wikipedia.org/wiki/Estimation_theory) the population [variance](https://en.wikipedia.org/wiki/Variance) and [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) from a sample when the population mean is unknown, the sample variance estimated as the *mean* of the squared deviations of sample values from their mean (that is, using a multiplicative factor 1/*n*) is a [biased estimator](https://en.wikipedia.org/wiki/Biased_estimator) of the population variance, and for the average sample underestimates it. Multiplying the standard sample variance as computed in that fashion by *n*/*n* − 1 (equivalently, using 1/*n* − 1 instead of 1/*n* in the estimator's formula) corrects for this, and gives an unbiased estimator of the population variance. In some terminology,[[1]](https://en.wikipedia.org/wiki/Bessel%27s_correction#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Bessel%27s_correction#cite_note-2) the factor *n*/*n* − 1 is itself called **Bessel's correction**.

**The Standard Error:**

We’ve seen that the standard deviation tells us something about how well the mean represents the sample data, but usually we collect data from samples because we don’t have access to the entire population. If you take several samples from a population, then these samples will differ slightly; therefore, it’s also important to know how well a particular sample represents the population. This is where we use the standard error .



sampling distribution: is simply the frequency distribution of sample means from the same population.

Standard Error : standard deviation of sample means.

Where is population mean . is mean of sample. n is number of samples.

In reality we cannot collect hundreds of samples and so we rely on approximations

of the standard error. Statisticians have demonstrated that as samples get large (usually defined as greater than 30), the sampling distribution has a normal distribution with a mean equal to the population mean, and a standard deviation of:

This is known as the **central limit theorem** and it is useful in this context because it means

that if our sample is large we can use the above equation to approximate the standard error

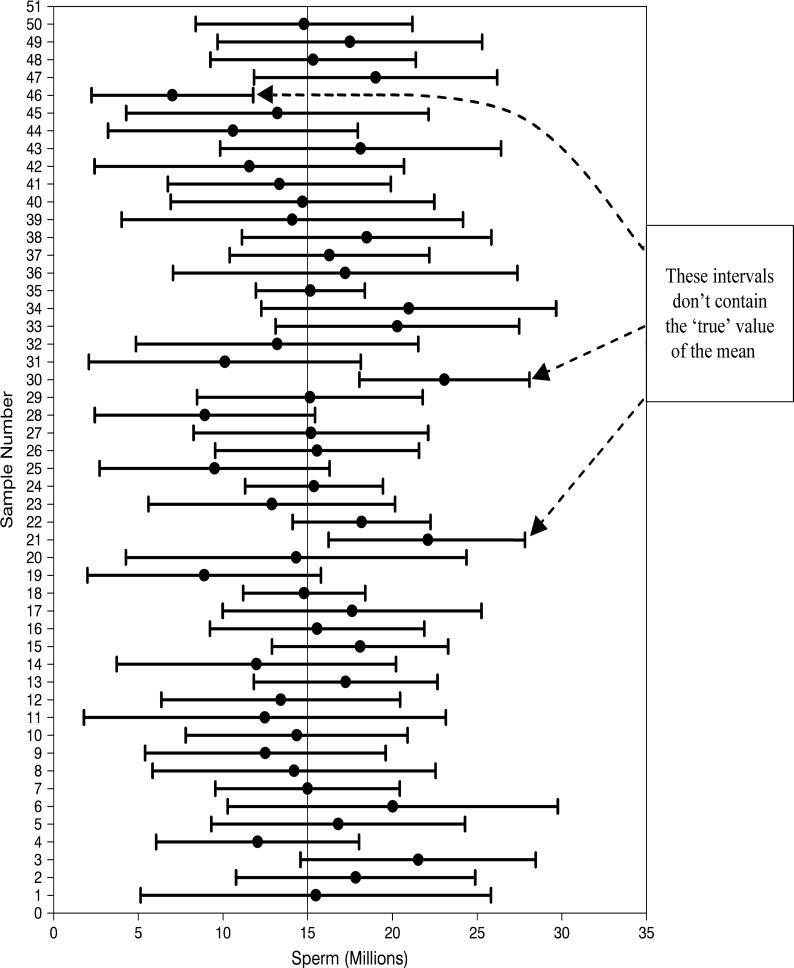
(because, remember, it is the standard deviation of the sampling distribution). When the

sample is relatively small (fewer than 30) the sampling distribution has a different shape,

known as a t -distribution.

**Confidence intervals:**

Standard error to get some idea of the extent to which sample means differ. A different approach to assessing the accuracy of the sample mean as an estimate of the mean in the population is to calculate boundaries within which we believe the true value of the mean will fall. Such boundaries are called confidence intervals .



Statisticians use a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) to describe the amount of uncertainty associated with a sample estimate of a population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter).

## How to Interpret Confidence Intervals

Suppose that a 90% confidence interval states that the population mean is greater than 100 and less than 200. How would you interpret this statement?

Some people think this means there is a 90% chance that the population mean falls between 100 and 200. This is incorrect. Like any population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter), the population mean is a constant, not a [random variable](http://stattrek.com/Help/Glossary.aspx?Target=Random%20variable). It does not change. The probability that a constant falls within any given range is always 0.00 or 1.00.

The [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=confidence_level) describes the uncertainty associated with a *sampling method*. Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter; A 95% confidence level means that 95% of the intervals would include the parameter; and so on.

A 95% confidence interval for a mean, think of it like this: if we’d collected 100 samples, calculated the mean and then calculated a confidence interval for that mean then for 95 of these samples, the confidence intervals we constructed would contain the true value of the mean in the population.

To calculate 95% confidence interval:

95% confidence interval is equal to 2.5%-97.5%. The SD at 2.5% and 97.5% is 1.96 in a normal curve.

lower boundary of confidence interval = − (1.96 × SE)

upper boundary of confidence interval = + (1.96 × SE)

=1.96

Examples :

<http://stattrek.com/estimation/confidence-interval.aspx>

<https://www.khanacademy.org/math/probability/statistics-inferential/confidence-intervals/v/small-sample-size-confidence-intervals>

**Use of confidence Interval:**

\*\*\*if the confidence intervals of any two means(samples) do not overlap then we can infer that these means are from different populations – they are significantly different.

**Calculating confidence intervals in small samples**

**t-distribution :**

In [probability](https://en.wikipedia.org/wiki/Probability) and [statistics](https://en.wikipedia.org/wiki/Statistics), Student's *t*-distribution (or simply the *t*-distribution) is any member of a family of continuous [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution) that arises when estimating the [mean](https://en.wikipedia.org/wiki/Expected_value) of a [normally distributed](https://en.wikipedia.org/wiki/Normal_distribution) [population](https://en.wikipedia.org/wiki/Statistical_population) in situations where the [sample size](https://en.wikipedia.org/wiki/Sample_size) is small and population [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) is unknown.

The t -distribution is a family of probability distributions that change shape as the sample size gets bigger (when the sample is very big, it has the shape of a normal distribution).

<https://en.wikipedia.org/wiki/Student%27s_t-distribution>

lower boundary of confidence interval = − ( × SE)

upper boundary of confidence interval = + ( × SE)

n-1 is the degree of freedom. and tells us which of the t -distributions to use.

**Using statistical models to test research questions:**

five-stage process:

1 Generate a research question through an initial observation (hopefully backed up by

some data).

2 Generate a theory to explain your initial observation.

3 Generate hypotheses: break your theory down into a set of testable predictions

4 Collect data to test the theory: decide on what variables you need to measure to test

your predictions and how best to measure or manipulate those variables.

5 Analyse the data: fit a statistical model to the data – this model will test your original

predictions. Assess this model to see whether or not it supports your initial

predictions.

**Inferential statistics:** which tell us whether the alternative hypothesis is likely to be true – they help us to confirm or reject our predictions. Crudely put, we fit a statistical model to our data that represents the alternative hypothesis and see how well it fits (in terms of the variance it explains). If it fits the data well (i.e. explains a lot of the variation in scores) then we assume our initial prediction is true: we gain confidence in the alternative hypothesis. Of course, we can never be completely sure that either hypothesis is correct, and so we calculate the probability that our model would fit if there were no effect in the population (i.e. the null hypothesis is true). As this probability decreases, we gain greater confidence that the alternative hypothesis is actually correct and that the null hypothesis can be rejected.

**Criminal Trial Analogy**

One place where you can consistently see the general idea of hypothesis testing in action is in criminal trials held in the United States. Our criminal justice system assumes "the defendant is innocent until proven guilty." That is, our initial assumption is that the defendant is innocent.

In the practice of statistics, we make our initial assumption when we state our two competing hypotheses -- the null hypothesis (*H*0) and the alternative hypothesis (*H*A). Here, our hypotheses are:

* *H*0: Defendant is not guilty (innocent)
* *H*A: Defendant is guilty

In statistics, we always **assume the null hypothesis is true**. That is, the null hypothesis is always our initial assumption.

The prosecution team then collects evidence — such as finger prints, blood spots, hair samples, carpet fibers, shoe prints, ransom notes, and handwriting samples — with the hopes of finding "sufficient evidence" to make the assumption of innocence refutable.

In statistics, the **data** are the evidence.

The jury then makes a decision based on the available evidence:

* If the jury finds sufficient evidence — beyond a reasonable doubt — to make the assumption of innocence refutable, the jury **rejects the null hypothesis** and deems the defendant guilty. We behave as if the defendant is guilty.
* If there is insufficient evidence, then the jury **does not reject the null hypothesis**. We behave as if the defendant is innocent.

In statistics, we always make one of two decisions. We either "reject the null hypothesis" or we "fail to reject the null hypothesis."

This is a very important distinction! We make our decision based on evidence not on 100% guaranteed proof. Again:

* If we reject the null hypothesis, we do not prove that the alternative hypothesis is true.
* If we do not reject the null hypothesis, we do not prove that the null hypothesis is true.

We merely state that there is enough evidence to behave one way or the other. This is always true in statistics! Because of this, whatever the decision, **there is always a chance that we made an error**.

Let's review the two types of errors that can be made in criminal trials:

|  |  |  |
| --- | --- | --- |
|  | **Truth** | |
| **Jury Decision** | Not Guilty | Guilty |
| Not Guilty | OK | **ERROR** |
| Guilty | **ERROR** | OK |

and let's see how they correspond to the two types of errors in hypothesis testing:

|  |  |  |
| --- | --- | --- |
|  | **Truth** | |
| **Decision** | Null Hypothesis | Alternative Hypothesis |
| Do not reject null | OK | **Type II ERROR** |
| Reject null | **Type I ERROR** | OK |

Note that, in statistics, we call the two types of errors by two different names -- one is called a "Type I error," and the other is called a "Type II error." Here are the formal definitions of the two types of errors:

|  |
| --- |
| **Type I error**: The null hypothesis is rejected when it is true.  **Type II error**: The null hypothesis is not rejected when it is false. |

Fisher suggested that 95% is a useful threshold for confidence: only when we are 95% certain that a result is genuine (i.e. not a chance finding) should we accept it as being true. The opposite way to look at this is to say that if there is only a 5% chance (a probability of .05) of something occurring by chance then we can accept that it is a genuine effect: we say it is a statistically significant finding. E.g. The lady tasting tea

**Test statistics:** We can test whether our statistical models (and therefore our hypotheses) are significant fits of the data we collected.

test statistic = variance explained by the model variance not explained by the model

= effect error

A test statistic is a statistic that has known properties; specifically we know how frequently

different values of this statistic occur. By knowing this, we can calculate the probability

of obtaining a particular value (just as we could estimate the probability of getting

a score of a certain size from a frequency distribution). This allows us to

establish how likely it would be that we would get a test statistic of a certain size if there

were no effect (i.e. the null hypothesis were true).

The more variation our model explains (compared to the variance it can’t explain), the bigger the test statistic will be, and the more unlikely it is to occur by chance

Consider, a person has very small probability of living for 110 years. Similar way as test statistics get bigger the probability of them occurring becomes smaller. When this probability falls below .05 (Fisher's Criterion), we accept that our model explains sufficient amount of variation.

\*\*\*\* Examples for hypothesis testing : Well explained

<https://onlinecourses.science.psu.edu/statprogram/print/book/export/html/139>

don’t be fooled by that word

‘significant’, because even if the probability of our

effect being a chance result is small (less than .05) it

doesn’t necessarily follow that the effect is important

Once you’ve calculated

your test statistic, you calculate the probability of

that test statistic occurring by chance; if this probability

is greater than .05 you reject your alternative

hypothesis. However, this does not mean that the null

hypothesis is true.

OK, we may not be able to accept

the null hypothesis as being true, but we can at least conclude

that it is false when our results are significant, right?

Wrong! A significant test statistic is based on probabilistic

reasoning, which severely limits what we can conclude.

\*\*Because the one-tailed test provides more power to detect an effect, you may be tempted to use a one-tailed test whenever you have a hypothesis about the direction of an effect. Before doing so, consider the consequences of missing an effect in the other direction. Imagine you have developed a new drug that you believe is an improvement over an existing drug. You wish to maximize your ability to detect the improvement, so you opt for a one-tailed test. In doing so, you fail to test for the possibility that the new drug is less effective than the existing drug. The consequences in this example are extreme, but they illustrate a danger of inappropriate use of a one-tailed test.

So when is a one-tailed test appropriate? If you consider the consequences of missing an effect in the untested direction and conclude that they are negligible and in no way irresponsible or unethical, then you can proceed with a one-tailed test. For example, imagine again that you have developed a new drug. It is cheaper than the existing drug and, you believe, no less effective. In testing this drug, you are only interested in testing if it less effective than the existing drug. You do not care if it is significantly more effective. You only wish to show that it is not less effective. In this scenario, a one-tailed test would be appropriate.

Effect Size:

Statistical Power:

**Assumptions of parametric data:**

* Normally Distributed Data:
* Homogeneity of Variance:
* Interval data:
* Independence:

we know from the central limit theorem that if the sample data are approximately normal or size of sample is big enough regardless the shape, then the sampling distribution will be normal. Therefore, people tend to look at their sample data to see if they are normally distributed. If so, then they have a little party to celebrate and assume that the sampling distribution (which is what actually matters) is also.

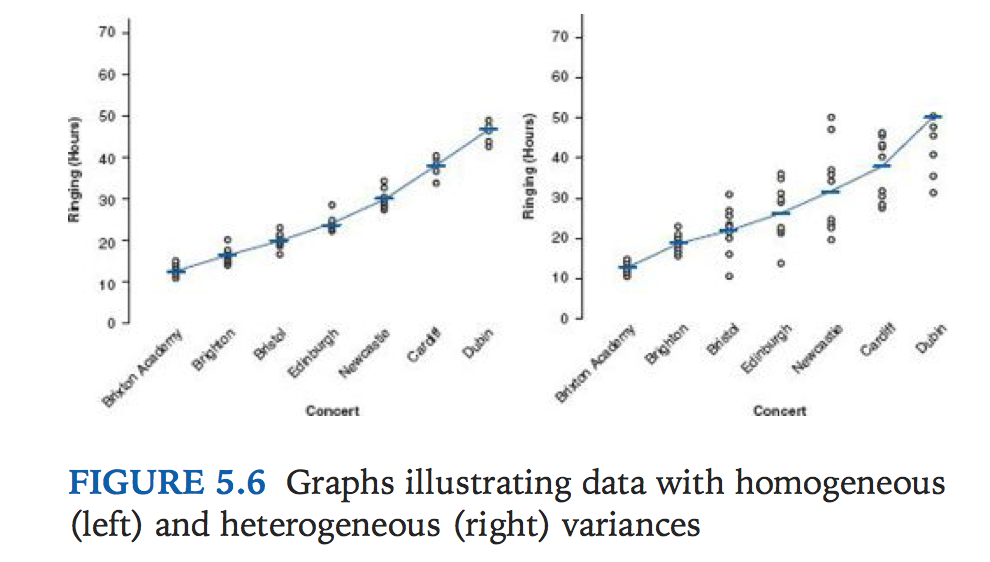
**To check normality of data:**

Q-Q plot (quantile–quantile plot): data are ranked and sorted. Each value is compared to the expected value that the score should have in a normal distribution and they are plotted against one another. If the data are normally distributed then the actual scores will have the same distribution as the score we expect from a normal distribution, and you’ll get a lovely straight diagonal line.

describe() in psych and stat.desc() in pastecs: output contains some values. Numbers for mean,median, skewness (+ve value indicate pile of scores in left side), kurtosis( +ve value indicates pointy and heavy tailed distribution),

Shapiro-Wilk Test: it compares the scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the test is non-significant (p > .05) it tells us that the distribution of the sample is not significantly different from a normal distribution. If, however, the test is significant (p < .05) then it is non-normal.

**To Check Homogeneity of Variance:**



Levene’s Test: Levene’s test tests the null hypothesis that the variances in different groups are equal. It’s a very simple and elegant test that works by doing a one-way ANOVA conducted on the deviation scores; that is, the absolute difference between each score and the mean of the group from which it came. For now, all we need to know is that if Levene’s test is significant at p ≤ .05 then we can conclude that the null hypothesis is incorrect and that the variances are significantly different. If, however, Levene’s test is non-significant (i.e., p > .05) then the variances are roughly equal and the assumption is tenable.

“Warning: In large samples Levene’s test can be significant even when group variances are not very different. Therefore, it should be interpreted in conjunction with the variance ratio.”

**Correcting Problems in Data:**

1. **Dealing with Outliers:**
   1. Remove the case: only if you have good reason to believe that the case is not from the population or the case could be a mis-entry.
   2. Transform the data: Some students feel like transforming data sounds dodgy.
      1. Log transformation
      2. Square root transformation
      3. Reciprocal transformation
      4. Reverse score transformation
   3. Change the score: This on the face of it may seem like cheating. Anyway changing the score is the lesser of two evils.
      1. The next highest score plus one.
      2. Change all scores above any percentile ( e.g 99%, 99.7% etc) to score at that percentile. Or you can say, change all scores to z-score and limit the scores above i.e. 3.29 to score 3.29.
   4. Robust Methods: There are reasons to think that transformations are not necessarily a good idea. A much more promising approach is to use robust methods.
      1. Trimmed Mean
      2. M-estimator: differs from trimmed mean in that the optimal amount of trimming is determined empirically. Disadvantage is that sometime it is not possible to reach a solution.
      3. Bootstrap: The problem that we have is that we don’t know the shape of the sampling distribution, but normality in our data allows us to infer that the sampling distribution is normal (and hence we can know the probability of a particular test statistic occurring). Lack of normality prevents us from knowing the shape of the sampling distribution unless we have big samples. It consider sample as population from which smaller samples (bootstraps samples) are taken( with replace). From this many bootstrap samples we can create a sampling distribution and can calculate standard error and hence confidence interval and significance tests.

**Correlation:**

Covariance: it is not standardized method.

Correlation : To make unit of measure standardize.

Pearson product moment correlation r =

Causality: Correlation coefficients no indication about causality. E.g. although we can conclude that as the number of adverts watched increases, the number of packets of toffees bought increases also, we cannot say that watching adverts causes you to buy packets of toffees. This caution is for two reasons:

* The third variable problem: causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.
* Direction of causality: Correlation coefficients say nothing about which variable causes the other to change. So, although it is intuitively appealing to conclude that watching adverts causes us to buy packets of toffees, there is no statistical reason why buying packets of toffees cannot cause us to watch more adverts. Although the latter conclusion makes less intuitive sense, the correlation coefficient does not tell us that it isn’t true.

Two types of correlation: bivariate and partial.

* A bivariate correlation is a correlation between two variables. E.g. Pearson’s product-moment correlation coefficient, Spearman’s rho and Kendall’s tau.
* A partial correlation looks at the relationship between two variables while ‘controlling’ the effect of one or more additional variables.

**Pearson’s Correlation:**

Pearson product moment correlation r =

Assumptions : both the variables should be normally distributed.

Function cor.test or, in package ‘Himsc’, function ‘rcorr’ provides correlation coeff as well p-value. Here corr = -.44 and p-value = .001 says ‘the p-value or significant value tell us that the probability of getting a correlation coefficient this big in a sample of n=103 data points if the null hypothesis were true is very low.

**Coefficient of Determination :**

‘The variance in y accounted for by x’ or ‘the variance in y explained by x’

\*100 %

Let say correlation is -0.4410 and so the value of R2 will be (-0.4410)^2 = 0.194. we can say that x shares 19.4% of the variability in y but it does not necessarily ‘cause’ this variation.

**Spearman’s rank correlation:**

It is non parametric statistics and so can be used when the data have violated parametric assumptions such as non-normally distributed data.

Spearman cor coef will be 1 when both variables are monotonically related.

When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.

The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's rho limits the outlier to the value of its rank.

method = “spearman” for cor() and cor.test(), and type = “spearman” for rcorr() in R

**Kendall’s tau Correlation coef:**

τ, is another non-parametric correlation and it should be used rather than Spearman’s coefficient when you have a small data set with a large number of tied ranks.

if both *xi* > *xj* and *yi* > *yj* or if both *xi* < *xj* and *yi* < *yj*. They are said to be *discordant*, if *xi* > *xj* and *yi* < *yj* or if *xi* < *xj* and y*i* > *yj*. If *xi* = *xj* or *yi* = *yj*, the pair is neither concordant nor discordant.

**Bootstrapping Correlations:**

**Biserial and Point-biserial correlations:**

The point-biserial correlation coefficient (rpb) is used when one variable is a discrete dichotomy (e.g., pregnancy: either a lady is pregnant or not), whereas the biserial correlation coefficient (rb) is used when one variable is a continuous dichotomy (e.g., passing or failing an exam: passing or failing by large margin or small, a continuum exists).

Use ‘cor.test’ for rpb and function ‘polyserial’ for biseria rbl.

**Partial Correlation :**

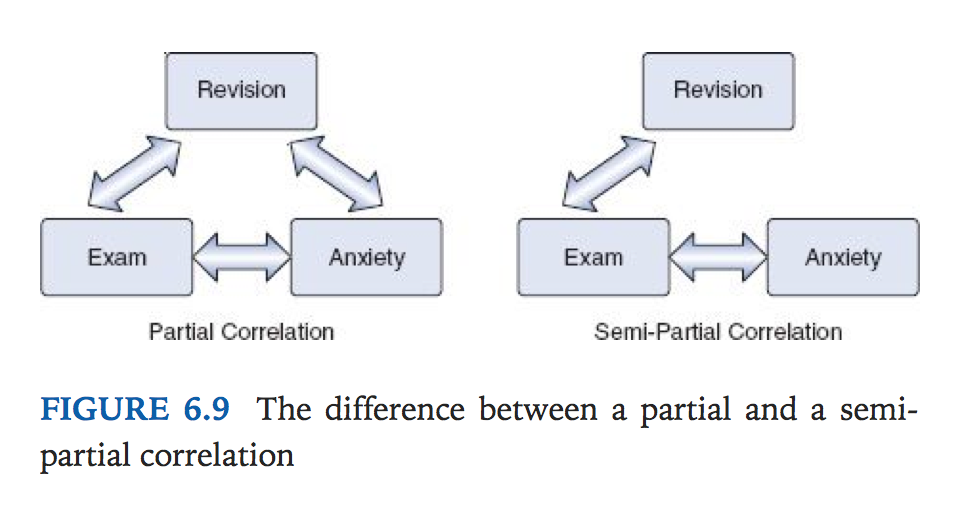
Exam anxiety accounts for 19.4% of the variance in exam performance, that revision time accounts for 15.7% of the variance in exam performance, and that revision time accounts for 50.2% of the variance in exam anxiety. If revision time accounts for half of the variance in exam anxiety, then it seems feasible that at least some of the 19.4% of variance in exam performance that is accounted for by anxiety is the same variance that is accounted for by revision time. As such, some of the variance in exam performance explained by exam anxiety is not unique and can be accounted for by revision time. A correlation between two variables in which the effects of other variables are held constant is known as a **partial correlation.**

we could conduct a partial correlation between exam anxiety and exam performance while ‘controlling’ for the effect of revision time, Let say for partial cor is .06 which means that exam anxiety can now account for only 6% of the variance in exam performance. Likewise, we could carry out a partial correlation between revision time and exam performance while ‘controlling’ for the effects of exam anxiety.

These partial correlations can be done when variables are dichotomous.

Package ‘ggm’, function ‘pcor’: pcor(c("var1", "var2", "control1", "control2" etc.), var(dataframe))

**Semi-partial (or part) correlations:**

When we do a partial correlation between two variables, we control for the effects of a third variable. Specifically, the effect that the third variable has on both variables in the correlation is controlled. In a semi-partial correlation we control for the effect that the third variable has on only one of the variables in the correlation.

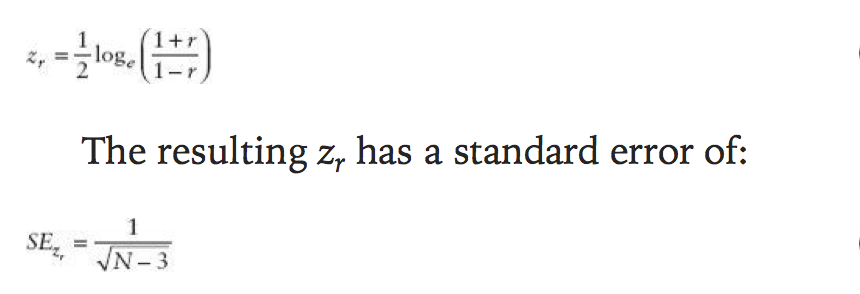
**Comparing independent correlations (rs) :**

When we looked at the effect of exam anxiety on exam performance, we might have been interested to know whether this correlation was different in men and women. We could compute the correlation in these two samples, but then how would we assess whether the difference was meaningful?

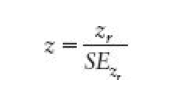
we would find that the correlations were rMale = –.506 and rFemale = –.381. These two samples are independent; that is, they contain different entities. Covert coef ‘r’ into zr by formula.

----------------------------

As we have seen, z-scores are useful because we know the probability of a given value of z occurring, if the distribution from which it comes is normal. There is one problem with Pearson’s r, which is that it is known to have a sampling distribution that is not normally distributed. This is a bit of a nuisance, but luckily, thanks to our friend Fisher, we can adjust r so that its sampling distribution is normal as follows. Here zr is adjusted r, not z-score

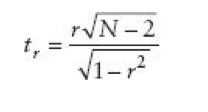


Normally we want to see whether the correlation is different from 0, in which case we can subtract 0 from the observed value of r and divide by the standard error.



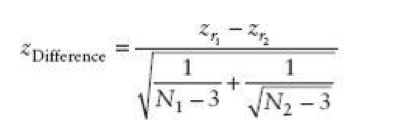
For our advert data this gives us 1.33/.71 = 1.87. We can look up this value of z (1.87) in the table for the normal distribution in the Appendix and get the one-tailed probability from the column labelled ‘Smaller Portion’. In this case the value is .0307. To get the two-tailed probability we simply multiply the one-tailed probability value by 2, which gives us .0614. As such the correlation is significant, p < .05, one-tailed, but not two-tailed.

In fact, the hypothesis that the correlation coefficient is different from 0 is usually (R, for example, does this) tested not using a z-score, but using a t-statistic with N - 2 degrees of freedom, which can be directly obtained from r:



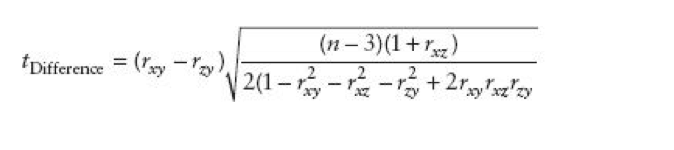
---------------------

Calculate z-score for difference of correlation for male and female as follows:



**Comparing Dependent correlations (rs) :**

For example, in our exam anxiety data we might want to see whether the relationship between exam anxiety (x) and exam performance (y) is stronger than the relationship between revision (z) and exam performance. To calculate this, all we need are the three rs that quantify the relationships between these variables: rxy, the relationship between exam anxiety and exam performance (–.441); rzy, the relationship between revision and exam performance (.397); and rxz, the relationship between exam anxiety and revision (–.709). The t-statistic is computed as:



**Regression:**

The most basic model, we can guess without any predictor variables, is mean. So we can compare the regression model with basic model (mean) to assess how good that regression model is doing.

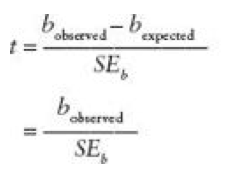
**The goodness of fit :**

* **Total sum of squares :** It’s sum of squared differences (deviations) when most basic model is applied to data. This value represents how good mean is as model of the observed data.
* **Residual sum of squares :** It’s sum of squared of residuals for the best fit model (regression).
* **Model sum of squares :** This difference shows us the reduction in the inaccuracy of the model resulting from fitting the regression model to the data.
* : It represents the amount of variance in the outcome explained by the model () relative to how much variation there was to explain in the first place ().

**Assessing individual predictors :**

If a variable significantly predicts an outcome then is should have coefficient *b* significantly different from 0. This hypothesis is tested using a t-test. The t-statistic tests the null hypothesis that the value of b is 0: therefore, if it is significant we gain confidence in the hypothesis that the b-value is significantly different from 0 and that the predictor variable contributes significantly to our ability to estimate values of the outcome.

We could take lots and lots of samples of data regarding and calculate the b-values for each sample. We could plot a frequency distribution of these samples to discover whether the b-values from all samples would be relatively similar, or whether they would be very different . We can use the standard deviation of this distribution (known as the standard error) as a measure of the similarity of b-values across samples. If the standard error is very small, then it means that most samples are likely to have a b-value similar to the one in our sample (because there is little variation across samples). The t-test tells us whether the b-value is different from 0 relative to the variation in b-values across samples. When the standard error is small even a small deviation from zero can reflect a meaningful difference because b is representative of the majority of possible samples. The b expected term is simply the value of b that we would expect to obtain if the null hypothesis were true and that is 0.

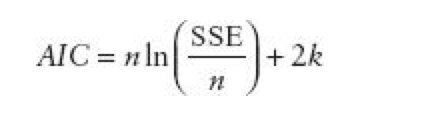


The values of t have a special distribution that differs according to the degrees of freedom for the test. In regression, the degrees of freedom are N – p – 1, where N is the total sample size and p is the number of predictors. In simple regression when we have only one predictor, so this gives N – 2. Having established which t-distribution needs to be used, the observed value of t can then be compared to the values that we would expect to find if there was no effect (i.e., b = 0): if t is very large then it is unlikely to have occurred when there is no effect. R provides the exact probability that the observed value (or a larger one) of t would occur if the value of b was, in fact, 0. As a general rule, if this observed significance is less than .05, then scientists assume that b is significantly different from 0; put another way, the predictor makes a significant contribution to predicting the outcome.

**Multiple Regression:**

**Multiple :** is the square of the correlation between the observed values of Y and the values of Y predicted by the multiple regression model. It is the amount of variation in the outcome variable that is accounted for by the model.

The big problem with is that when you add more variables to the model, it will always go up. If you are deciding which of two models fits the data better, the model with more predictor variables in will always fit better. The **Akaike information criterion (AIC)** is a measure of fit which penalizes the model for having more variables – a little like adjusted R2. The AIC is defined as:



Imagine we add a variable to the model; usually this would increase R2, and hence SSE would be reduced. But imagine that this variable does not change the fit of the model at all. What will happen to the AIC? Well, the first part will be the same: n and SSE are unchanged. What will change is k: it will be higher, by one (because we have added a variable). Hence, when we add this variable to the model, the AIC will be higher by 2. A larger value of the AIC indicates worse fit, corrected for the number of variables.

**Bayesian information criterion (BIC)**

**Methods of Regressions:**

Based on the order of entry of predictor in model:

Hierarchical:

Forced Entry:

Stepwise:

Forward direction: start only constatn b0 then select the predictor that has highest simple correlation with outcome. Then second predictor is chosen, if it has largest semi-partial correlation with the outcome.

Backward elimination: It starts with all predictor used for model and then by looking to see if AIC goes down when each variable is removed. If a variable is removed, the contribution of the remaining predictors is then reassessed and the process continues until removing any variable causes AIC to increase.

Both Direction: It starts the in same way as the forward method, except that each time a predictor is added to the equation, a removal test is made of the least useful predictor. As such the regression equation is constantly being reassessed to see whether any redundant predictors can be removed.

If you do decide to use a stepwise method then the backward direction is preferable to the forward method. This is because of suppressor effects, which occur when a predictor has an effect but only when another variable is held constant. Forward selection is more likely than backward elimination to exclude predictors involved in suppressor effects. As such, the forward method runs a higher risk of making a Type II error (i.e., missing a predictor that does in fact predict the outcome).

All subset methods: Make model for all subsets of the predictors.

**Which method to choose?**

Stepwise methods are best avoided except for exploratory model building. If you must do a stepwise regression then it is advisable to cross-validate your model by splitting the data. Stepwise methods may overfit or underfit.

\*\*So we should not only depend on the computer to choose predictor for us, but we should research and include theoretical meaningful and important predictors.

**How accurate is my regression model?**

1. Does the model fit the observed data well, or it influenced by small number of cases?
2. Can my model generalize to other samples?
3. **Assessing the regression model I: diagnostics**

A. Outliers and residuals: An outlier is a case that substantially differs from main trend of the data. How to detect an outlier?? We can do it by residuals ( error between predicted value and observed value) for each case in data we will have a residual. We can convert these residuals to standardized residuals (z-scores), and can compare residuals from different models.

Some general rules for standardized residuals:

1. standardized residuals with an absolute value greater than 3.29 (3) are cause for concern because in an average sample a value this high is unlikely to happen by chance
2. if more than 1% of our sample cases have standardized residuals with an absolute value greater than 2.58 (2.5) there is evidence that the level of error within our model is unacceptable (the model is a fairly poor fit of the sample data
3. if more than 5% of cases have standardized residuals with an absolute value greater than 1.96 (2) then there is also evidence that the model is a poor representation of the actual data.

B. Influential Case: Again this process will unveil outliers. Some residual statistics is as follows

1. Adjusted predicted value: If we have doubt on any case, we can exclude that case while calculating a new model and then predicting the excluded case by the new model. If a case does not exert a large influence over the model then we would expect the adjusted predicted value to be very similar to the predicted value when the case is included. Studentized residual, cook’s distance and hat values.

**2. Assessing the regression model II: generalization**

For a regression model to generalize we must be sure that underlying assumptions have been met, and to test whether the model does generalize we can look at cross-validating it.

Checking Assumption :

* Variable type:
* Non-zero variance: at least some variation in predictor variables.
* No Perfect multicollinearity: Predictor variables should not correlate to highly.
* Predictors are uncorrelated with ‘external variables’:There should be no external variables(which is not included in regression) that correlate with any of the variables included in the regression model. Obviously, if it do correlate with the predictors, then the conclusions we draw from the model become unreliable (because other variables exist that can predict the outcome just as well).
* Homoscedasticity: At each level of the predictor variable(s), the variance of the residual terms should be constant.when the variances are very unequal there is said to be heteroscedasticity.
* Independent Errors:For any two observations the residual terms should be uncorrelated or can say lack of autocorrelation. Tested with Durbin watson test.
* Normally distributed error: It is assumed that the residuals in the model are random, normally distributed variables with a mean of 0. Predictors do not need to be normally distributed.
* Independence: It is assumed that all of the values of the outcome variable are independent (in other words, each value of the outcome variable comes from a separate entity).
* Linearity: The mean values of the outcome variable for each increment of the predictor(s) lie along a straight line. In plain English this means that it is assumed that the relationship we are modelling is a linear one. If we model a non-linear relationship using a linear model then this obviously limits the generalizability of the findings.

**Cross Validation of Model:**