**Mean:**

The "average" number; found by adding all data points and dividing by the number of data points.

**Median:** The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

**Mode:** The most frequent number—that is, the number that occurs the highest number of times.

Range :

The "range" of a list a numbers is just the difference between the largest and smallest values.

Examples :-

13, 18, 13, 14, 13, 16, 14, 21, 13

Calculating Mean :-

The mean is the usual average, so I'll add and then divide:

(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15

Calculating Median :

The median is the middle value, so first I'll have to rewrite the list in numerical order:

13, 13, 13, 13, 14, 14, 16, 18, 21

There are nine numbers in the list, so the middle one will be the (9 + 1) ÷ 2 = 10 ÷ 2 = 5th number:

13, 13, 13, 13, 14, 14, 16, 18, 21

So the median is 14.

Calculating Mode :-

The mode is the number that is repeated more often than any other, so 13 is the mode.

Calculating Range :

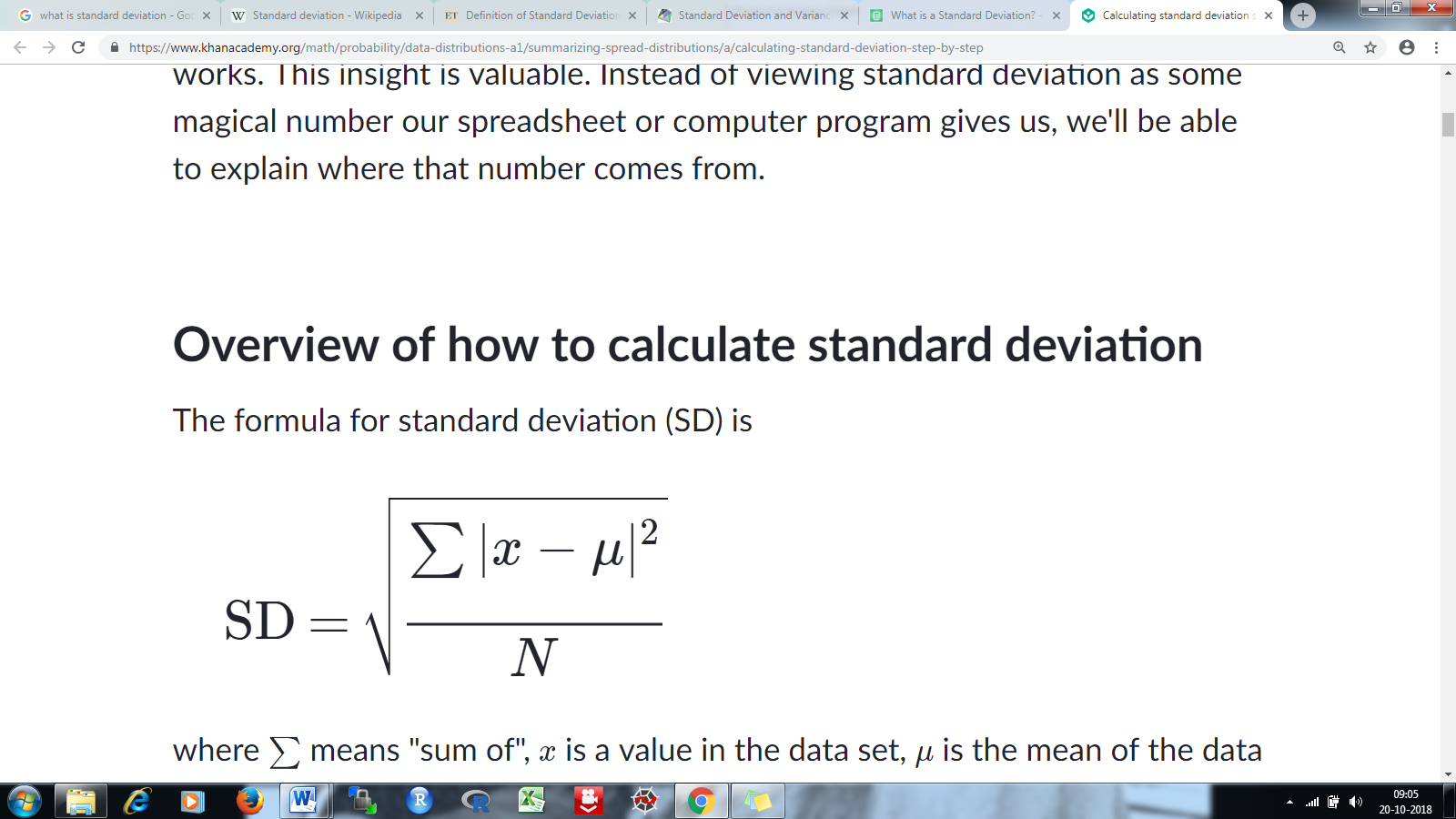
The largest value in the list is 21, and the smallest is 13, so the range is 21 – 13 = 8.

Standard Deviation

Standard Deviation (SD) is a statistical measure that captures the difference between the average and the outliers in a set of data.

 A low standard deviation indicates that the data points tend to be close to the [mean](https://en.wikipedia.org/wiki/Mean) (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

The formula for standard deviation (SD) is



where ∑ is "sum of"---

x is a value in the data set

*μ* is the mean of the data set,

N is the number of data points in the population.

Step 1: Find the mean.

Step 2: For each data point, find the square of its distance to the mean.

Step 3: Sum the values from Step 2.

Step 4: Divide by the number of data points.

Step 5: Take the square root.

Examples :-

6,2,3,1

Step 1:

1. Average = 6+2+3+1 = 12/4 = 3

Step 2:

1. 6 – 3 = 3 = square of 3 = 9
2. 2 – 3 = -1 = square of -1 = 1
3. 3-3=0 = square of 0 = 0
4. 1-3 = -2 = square of -2 = 4

Step 3:

1. 9+1+0+4 = 14

Step 4 :

1. 14/4 = 3.5

Step 5:

Sqrt of 3.5 = 1.87

* 1. is standard deviation of the above dataset.

Variance :-

It measures how far a set of (random) numbers are spread out from their average value.

The average of the squared differences from the Mean

Here 3.5 is the variance.

Variability

Variability is the extent to which data points in a statistical distribution or data set diverge from the average, or mean, value as well as the extent to which these data points differ from each other.

interquartile range

The interquartile range (IQR) is a measure of variability, based on dividing a data set into [quartiles](https://stattrek.com/Help/Glossary.aspx?Target=Quartile).

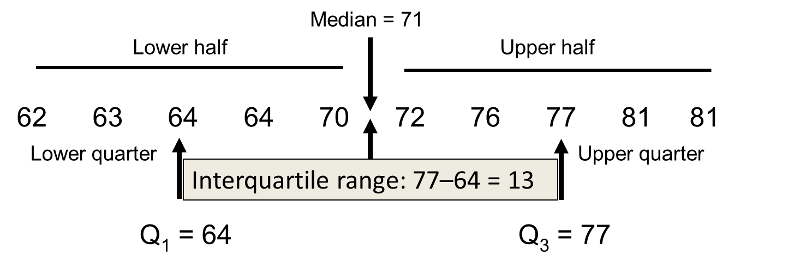
Q1 is the "middle" value in the *first* half of the rank-ordered data set.

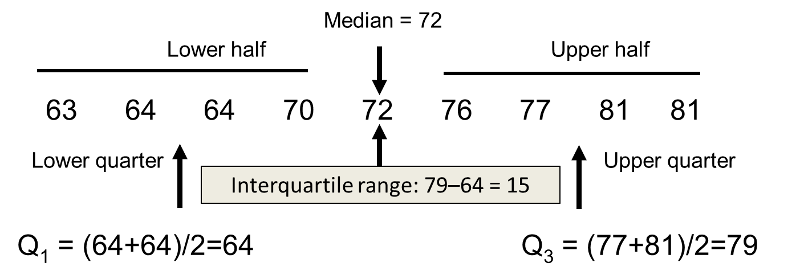
Q2 is the [median](https://stattrek.com/Help/Glossary.aspx?Target=Median) value in the set.

Q3 is the "middle" value in the second half of the rank-ordered data set.

The interquartile range is equal to Q3 minus Q1.

For example:





The IQR is used to build [box plots](https://en.wikipedia.org/wiki/Box_plot), simple graphical representations of a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution).

The IQR can be used to identify [outliers](https://en.wikipedia.org/wiki/Outlier) (see [below](https://en.wikipedia.org/wiki/Interquartile_range#Outliers)).

Types of Distributions

1. Bernoulli Distribution
2. Uniform Distribution
3. Binomial Distribution
4. Normal Distribution
5. Poisson Distribution
6. Exponential Distribution

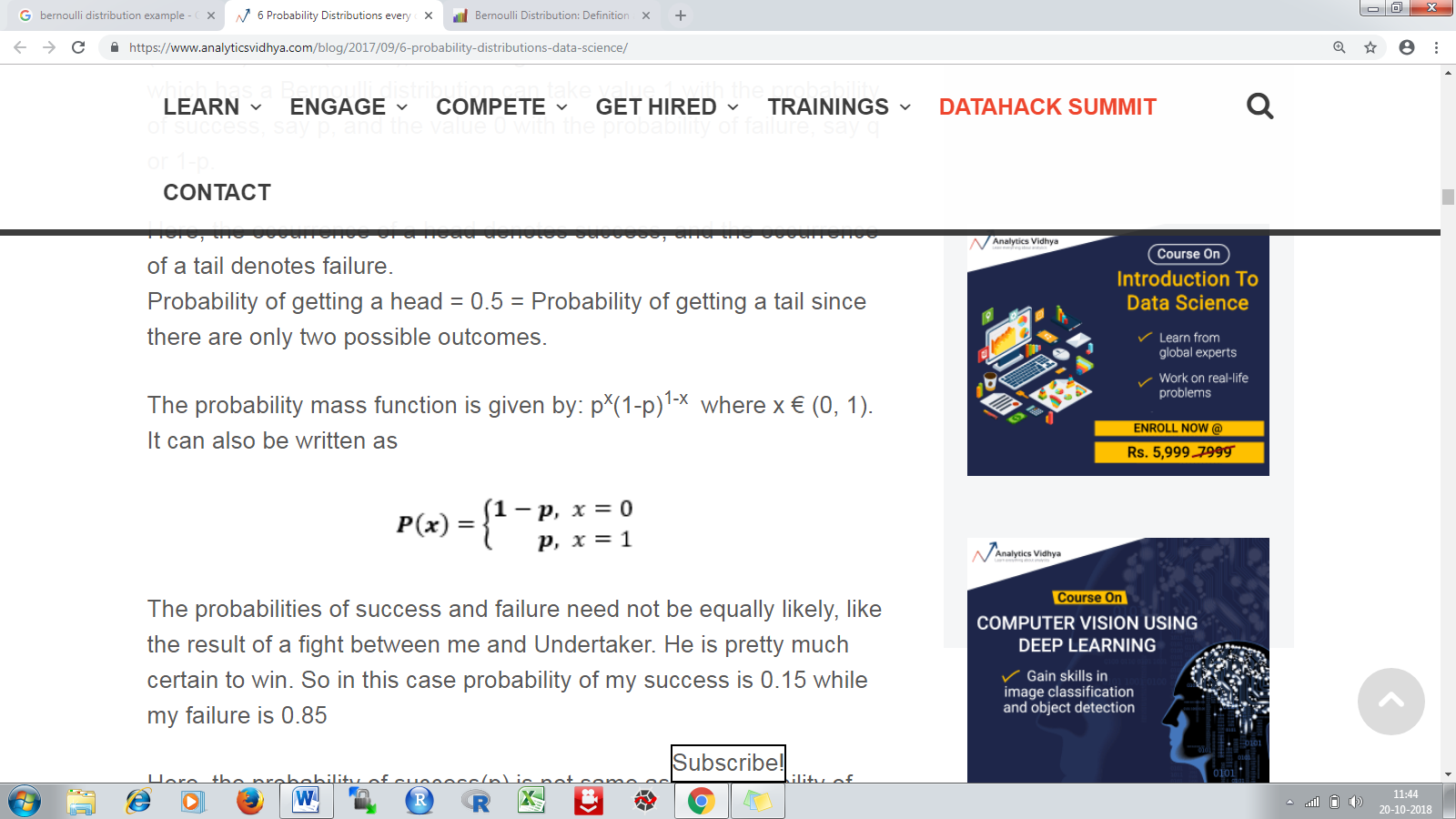
## Common Data Types

Discrete Data, as the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5 or 6 and not 1.5 or 2.45.

Continuous Data can take any value within a given range. The range may be finite or infinite. For example, A girl’s weight or height, the length of the road. The weight of a girl can be any value from 54 kgs, or 54.5 kgs, or 54.5436kgs.

### Bernoulli Distribution

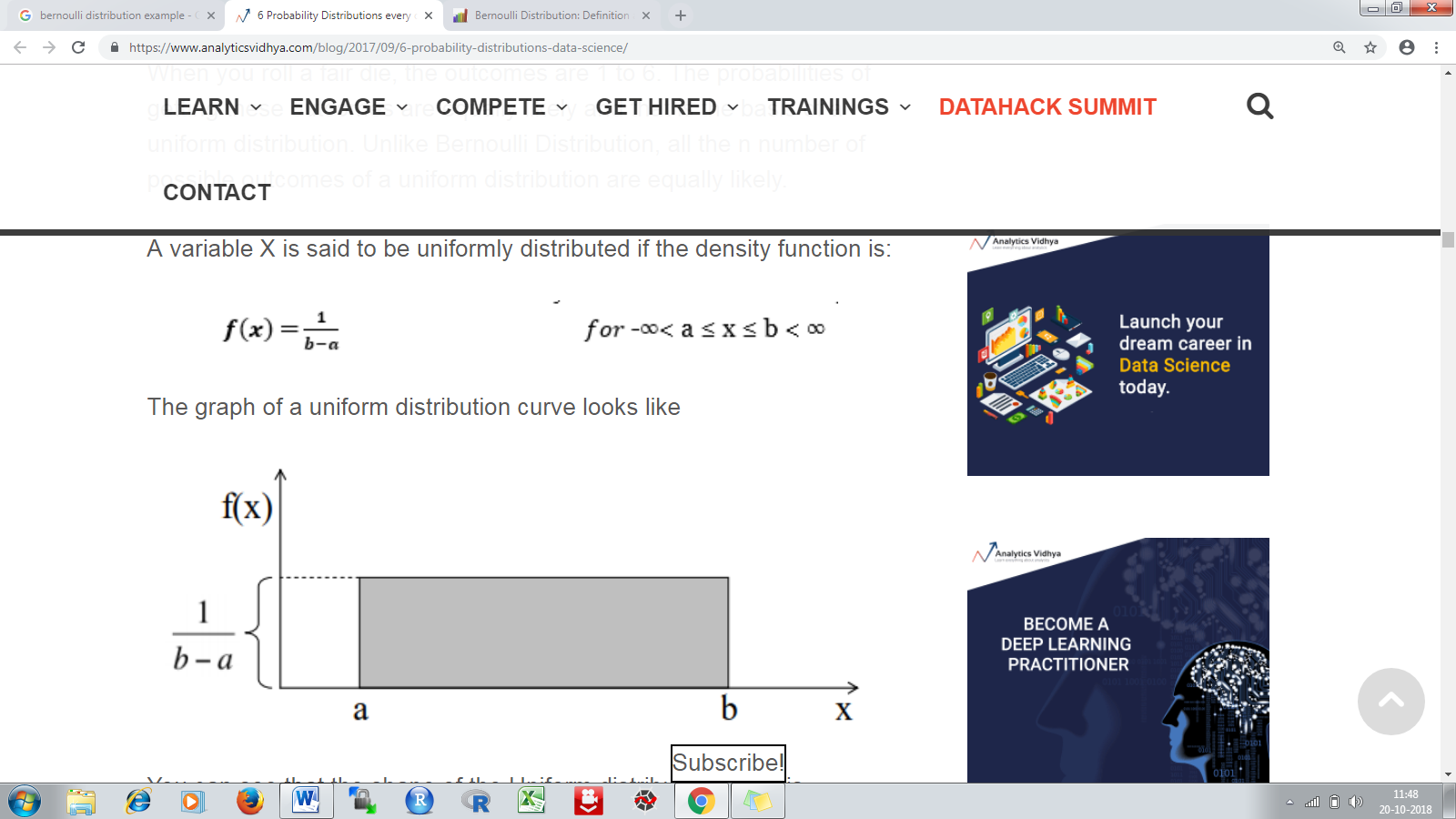
A **Bernoulli distribution** has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

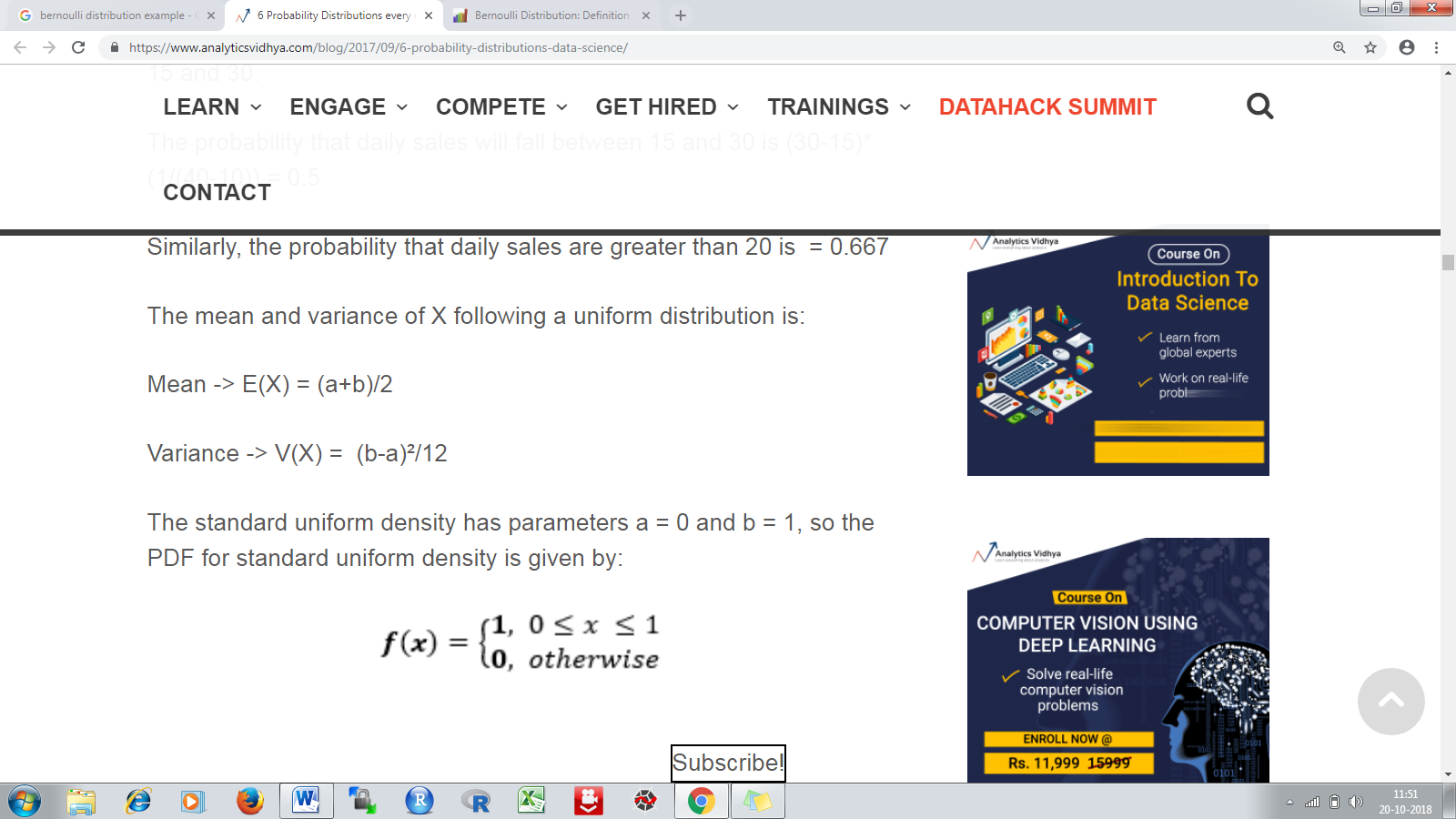


### Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution.

A variable X is said to be uniformly distributed if the density function is:





### Binomial Distribution

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

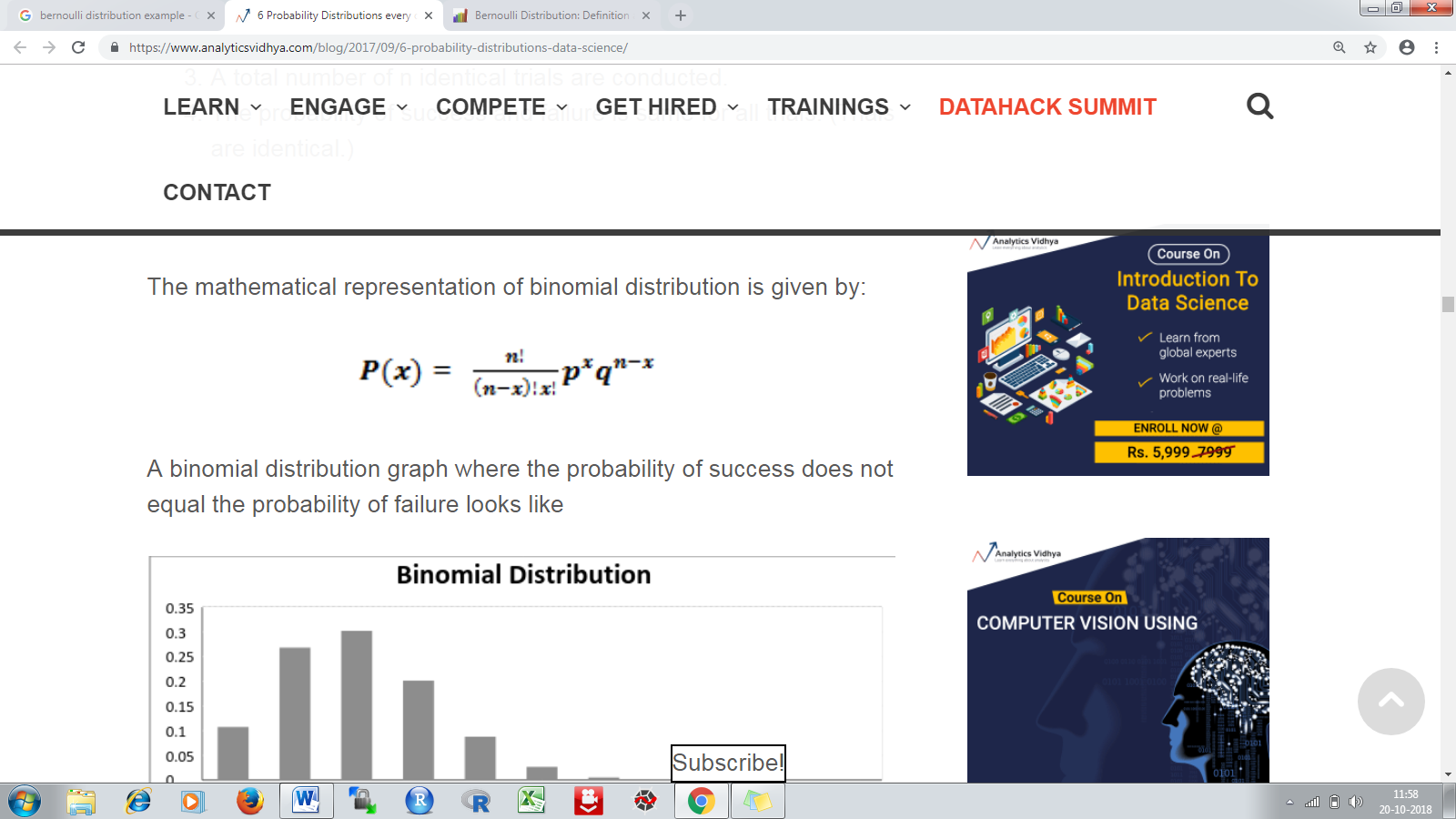
Each trial is independent since the outcome of the previous toss doesn’t determine or affect the outcome of the current toss.

An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

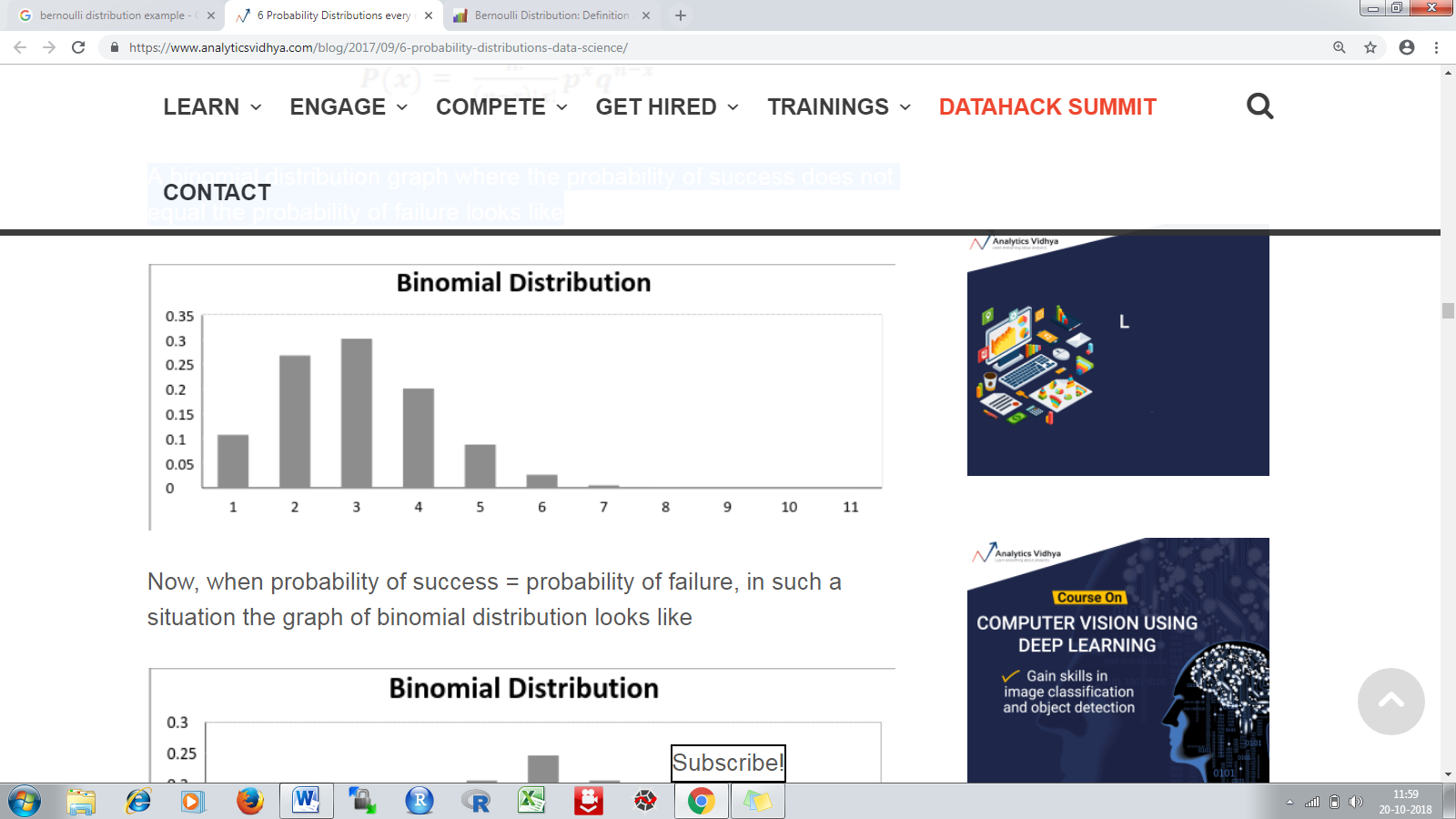
The properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

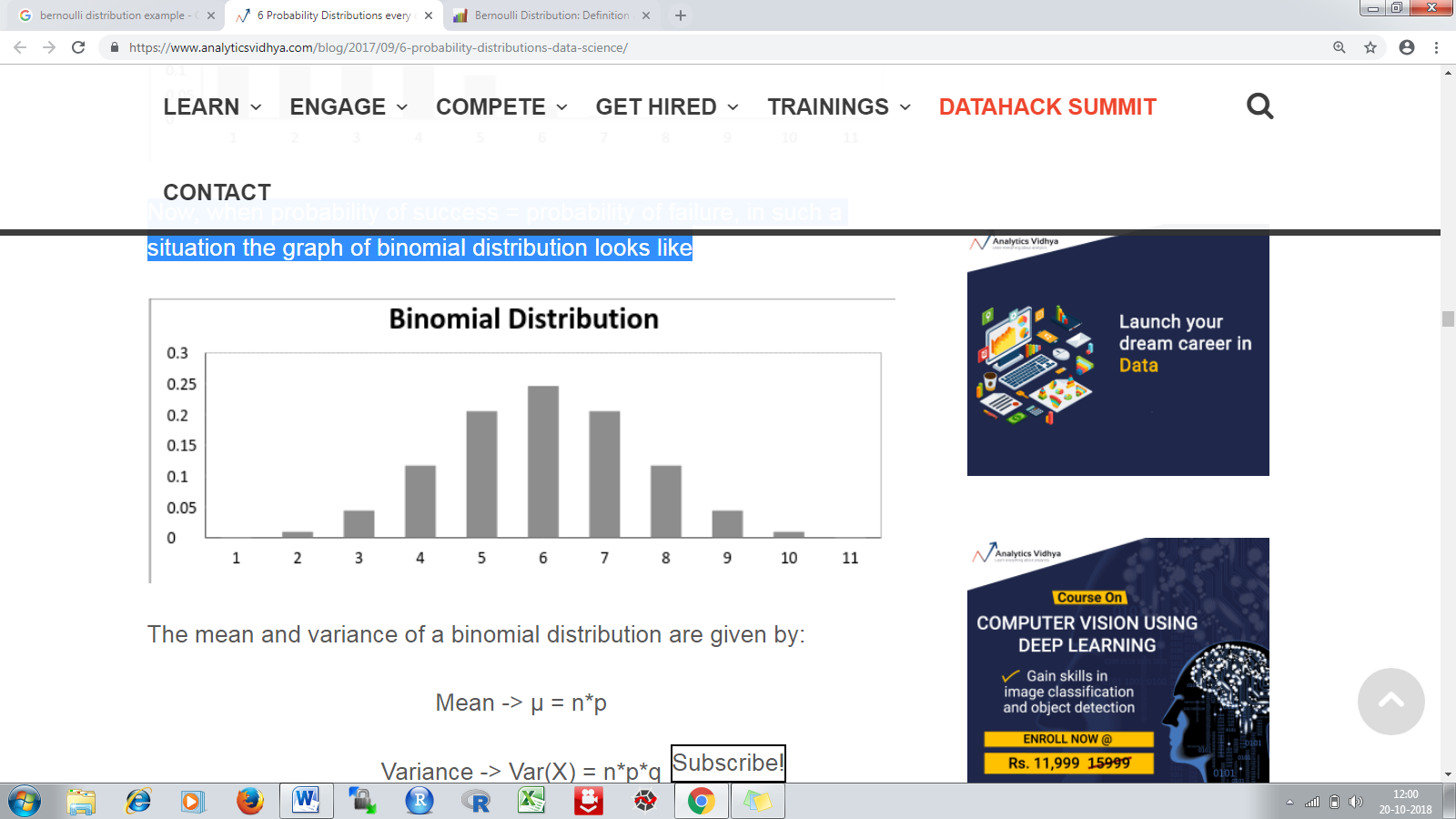
The mathematical representation of binomial distribution is given by:



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



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



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

Mean -> µ = n\*p

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

### Normal Distribution

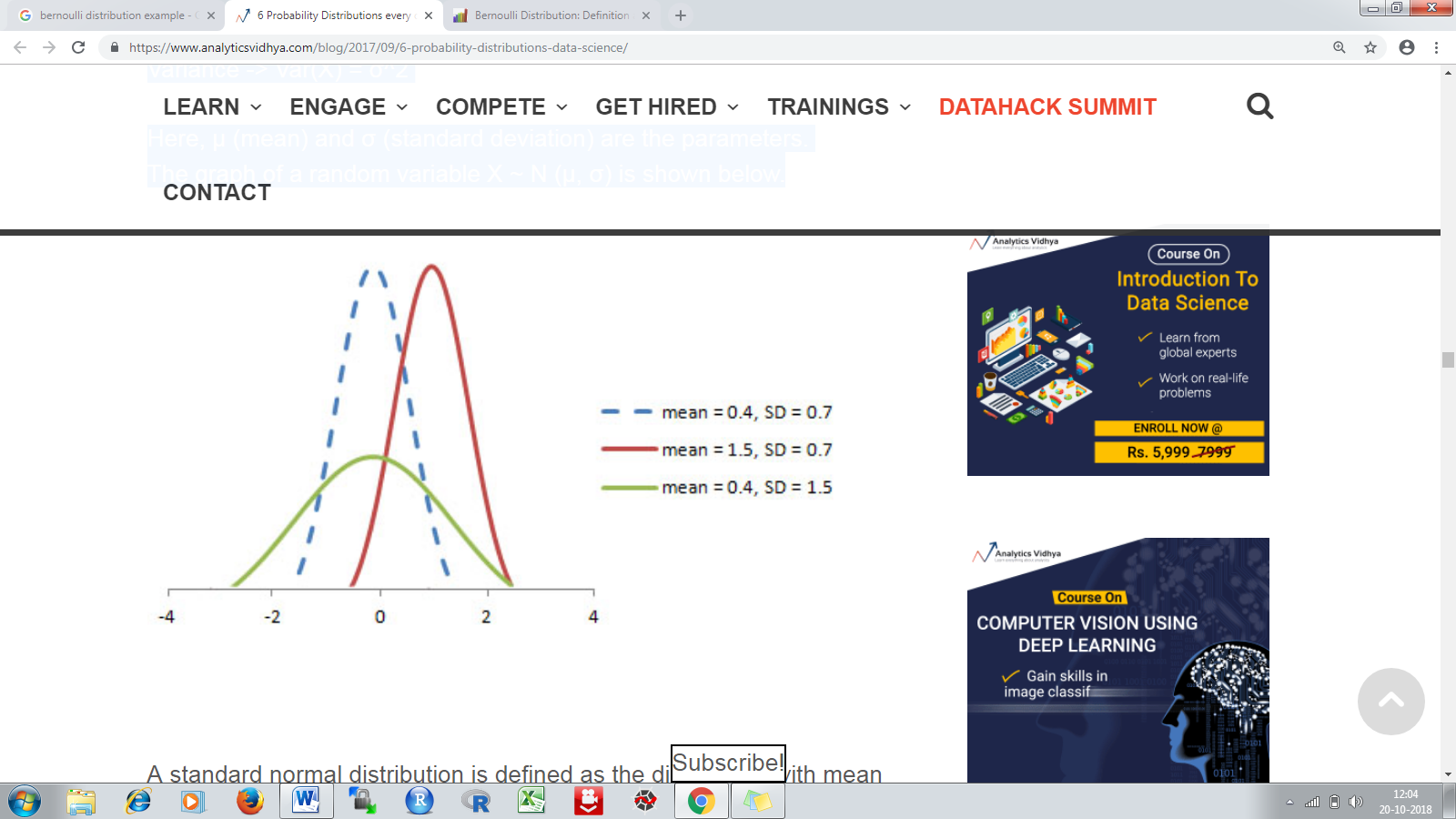
Any distribution is known as Normal distribution if it has the following characteristics:

1. The mean, median and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center and the other half to the right.
5. The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean -> E(X) = µ

Variance -> Var(X) = σ^2

Here, µ (mean) and σ (standard deviation) are the parameters. The graph of a random variable X ~ N (µ, σ) is shown below.



### Poisson Distribution

Suppose you work at a call center, approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are

1. The number of emergency calls recorded at a hospital in a day.
2. The number of thefts reported in an area on a day.
3. The number of customers arriving at a salon in an hour.
4. The number of suicides reported in a particular city.
5. The number of printing errors at each page of the book.

Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called Poisson distribution when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.  
2. The probability of success over a short interval must equal the probability of success over a longer interval.  
3. The probability of success in an interval approaches zero as the interval becomes smaller.

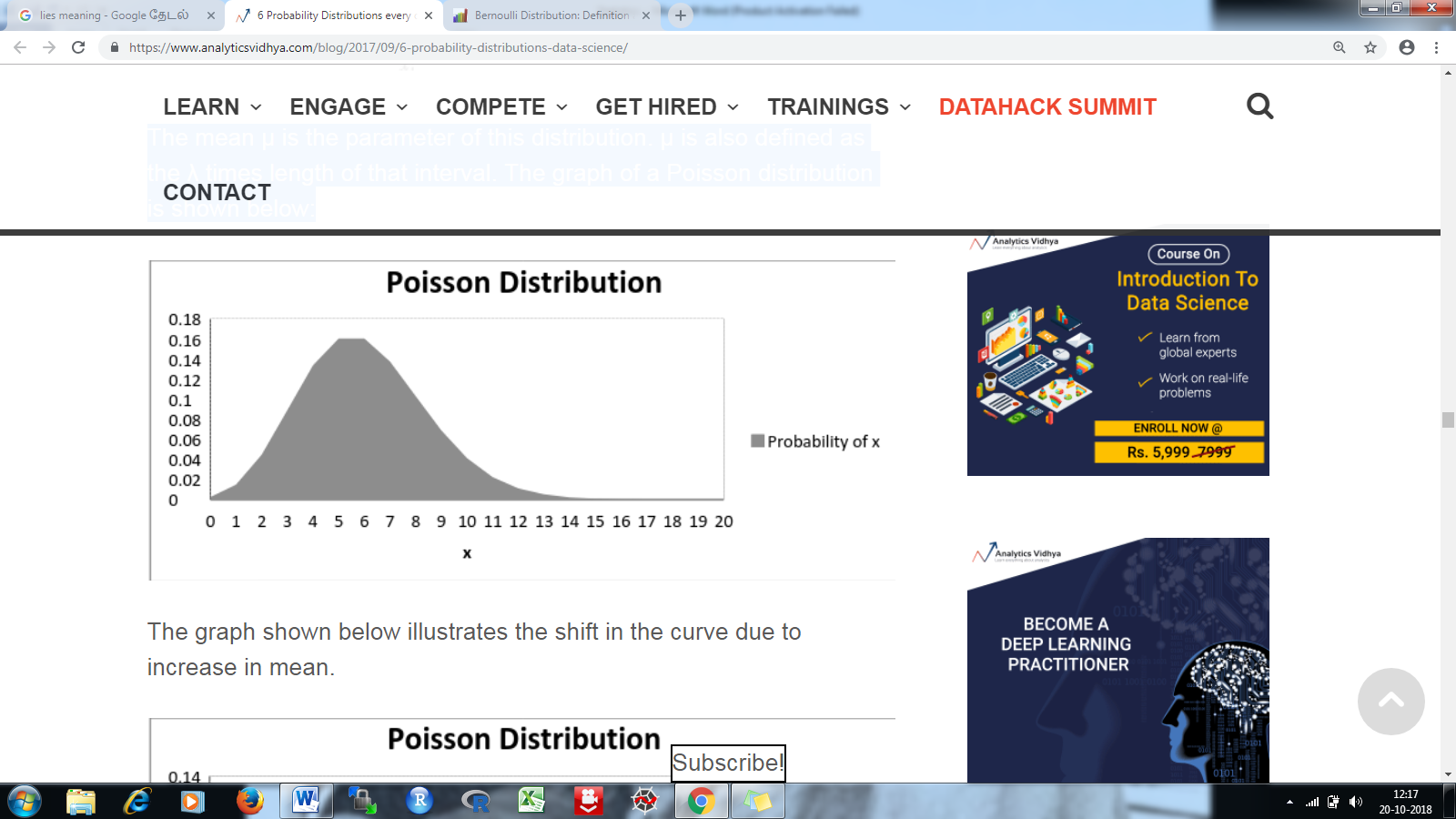
Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Let µ denote the mean number of events in an interval of length t. Then, µ = λ\*t.

The PMF of X following a Poisson distribution is given by:



The mean µ is the parameter of this distribution. µ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:



The mean and variance of X following a Poisson distribution:

Mean -> E(X) = µ  
Variance -> Var(X) = µ

### Exponential Distribution

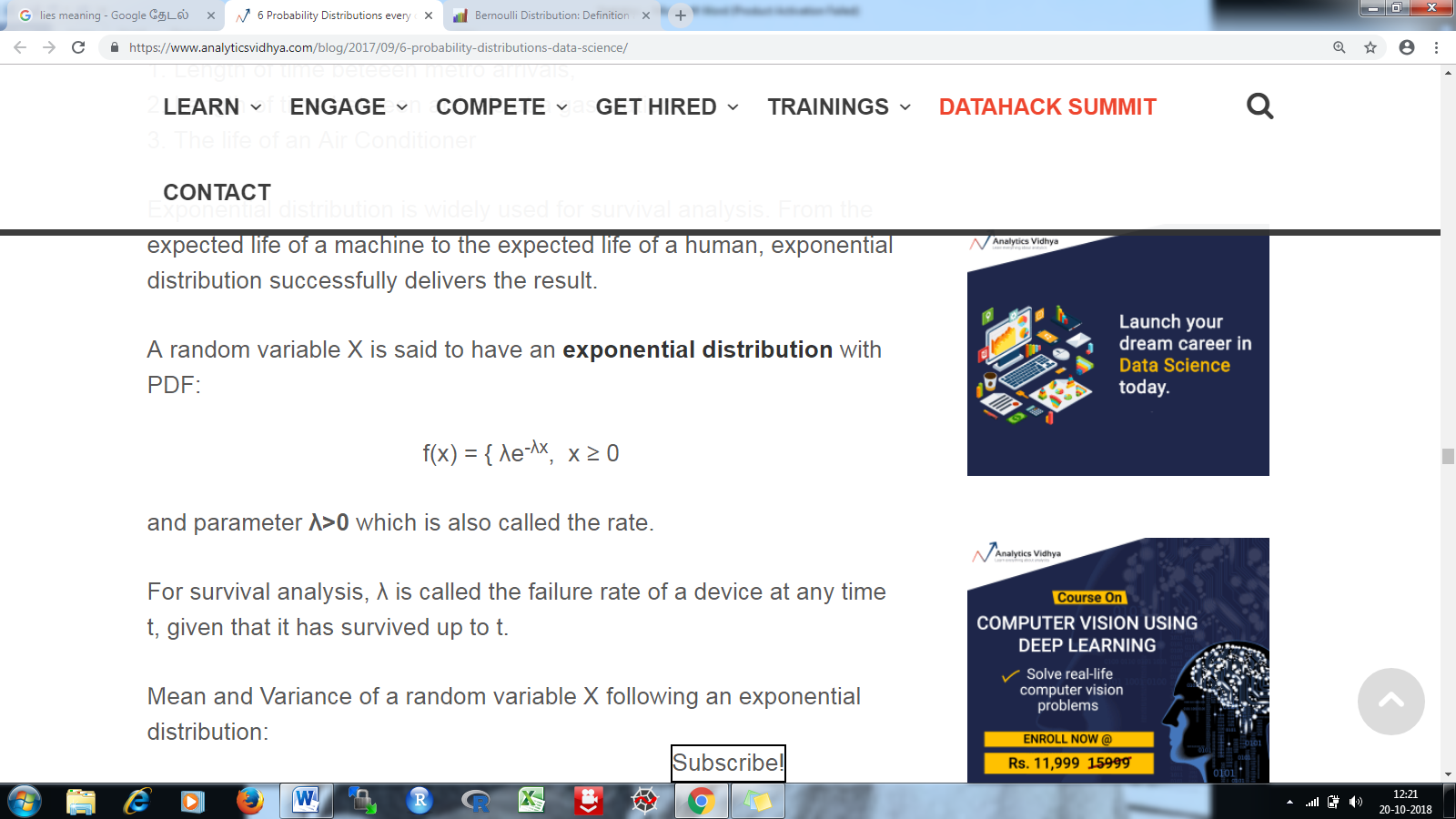
Let’s consider the call center example one more time. What about the interval of time between the calls ? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

Other examples are:

1. Length of time beteeen metro arrivals,  
2. Length of time between arrivals at a gas station  
3. The life of an Air Conditioner

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

A random variable X is said to have an exponential distribution with PDF:



and parameter λ>0 which is also called the rate.

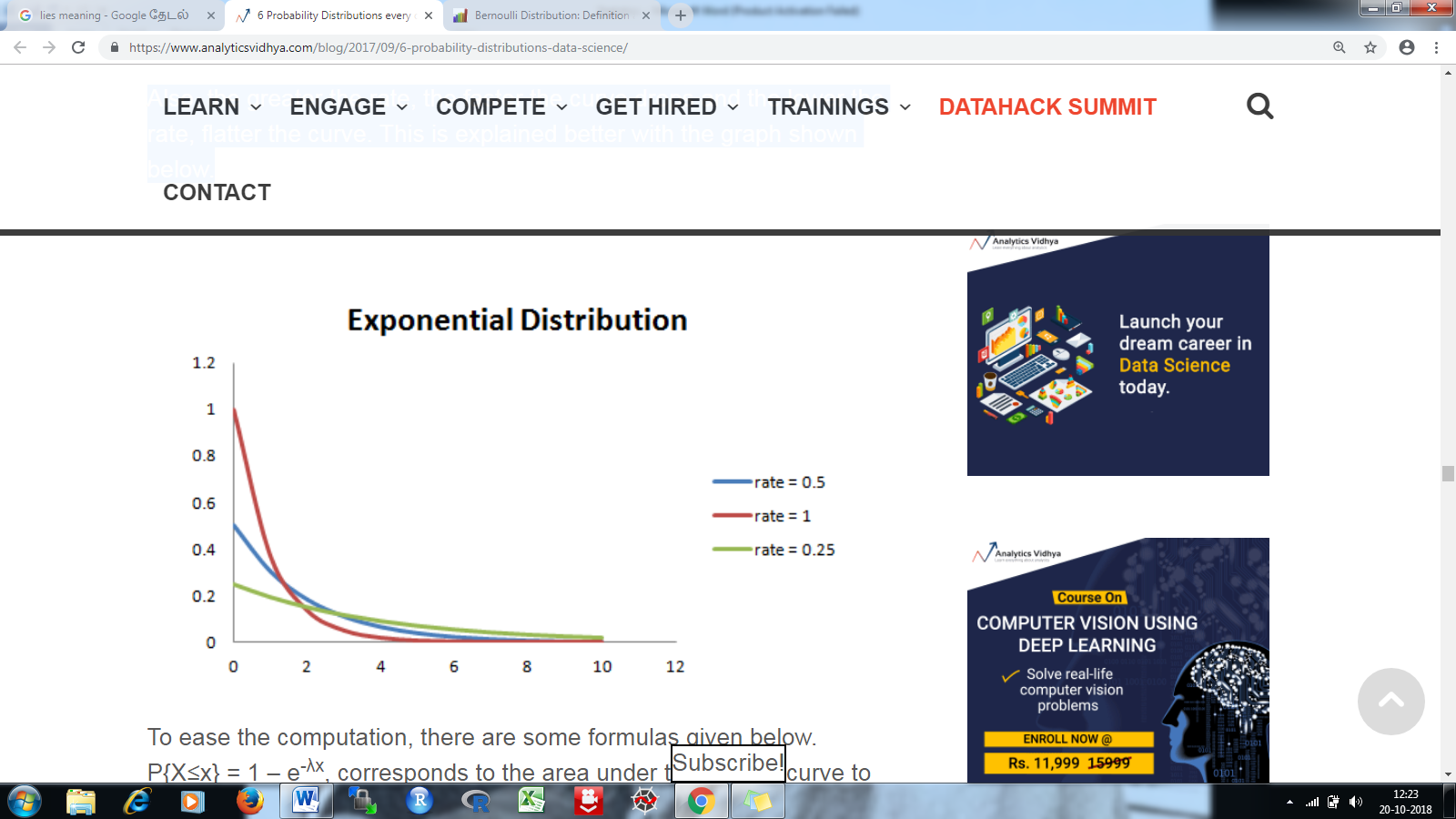
For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

Mean and Variance of a random variable X following an exponential distribution:

Mean -> E(X) = 1/λ

Variance -> Var(X) = (1/λ)²

Also, the greater the rate, the faster the curve drops and the lower the rate, flatter the curve. This is explained better with the graph shown below.



### Relation between Bernoulli and Binomial Distribution

1. Bernoulli Distribution is a special case of Binomial Distribution with a single trial.

2. There are only two possible outcomes of a Bernoulli and Binomial distribution, namely success and failure.

3. Both Bernoulli and Binomial Distributions have independent trails.

### Relation between Poisson and Binomial Distribution

Poisson Distribution is a limiting case of binomial distribution under the following conditions:

1. The number of trials is indefinitely large or n → ∞.
2. The probability of success for each trial is same and indefinitely small or p →0.
3. np = λ, is finite.

### Relation between Normal and Binomial Distribution & Normal and Poisson Distribution:

Normal distribution is another limiting form of binomial distribution under the following conditions:

1. The number of trials is indefinitely large, n → ∞.
2. Both p and q are not indefinitely small.

The normal distribution is also a limiting case of Poisson distribution with the parameter λ →∞.

### Relation between Exponential and Poisson Distribution:

If the times between random events follow exponential distribution with rate λ, then the total number of events in a time period of length t follows the Poisson distribution with parameter λt.

## What are Outliers?

An outlier is an observation that is unlike the other observations.

It is rare, or distinct, or does not fit in some way.

Outliers can have many causes, such as:

* Measurement or input error.
* Data corruption.
* True outlier observation (e.g. Michael Jordan in basketball)

# correlation

Correlation is a statistical measure that indicates the extent to which two or more [variables](https://whatis.techtarget.com/definition/variable) fluctuate together.

A [positive correlation](https://whatis.techtarget.com/definition/positive-correlation) indicates the extent to which those variables increase or decrease in parallel;

a [negative correlation](https://whatis.techtarget.com/definition/negative-correlation) indicates the extent to which one variable increases as the other decreases.

A **correlation coefficient** is a [numerical measure](https://en.wikipedia.org/wiki/Numerical_measure) of some type of [correlation](https://en.wikipedia.org/wiki/Correlation_and_dependence), meaning a statistical relationship between two [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)).

**Sample question**: Find the value of the correlation coefficient from the following table:

|  |  |  |
| --- | --- | --- |
| SUBJECT | AGE X | GLUCOSE LEVEL Y |
| 1 | 43 | 99 |
| 2 | 21 | 65 |
| 3 | 25 | 79 |
| 4 | 42 | 75 |
| 5 | 57 | 87 |
| 6 | 59 | 81 |

**Step 1:***Make a chart.* Use the given data, and add three more columns: xy, x2, and y2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUBJECT | AGE X | GLUCOSE LEVEL Y | XY | X2 | Y2 |
| 1 | 43 | 99 |  |  |  |
| 2 | 21 | 65 |  |  |  |
| 3 | 25 | 79 |  |  |  |
| 4 | 42 | 75 |  |  |  |
| 5 | 57 | 87 |  |  |  |
| 6 | 59 | 81 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUBJECT | AGE X | GLUCOSE LEVEL Y | XY | X2 | Y2 |
| 1 | 43 | 99 | 4257 | 1849 | 9801 |
| 2 | 21 | 65 | 1365 | 441 | 4225 |
| 3 | 25 | 79 | 1975 | 625 | 6241 |
| 4 | 42 | 75 | 3150 | 1764 | 5625 |
| 5 | 57 | 87 | 4959 | 3249 | 7569 |
| 6 | 59 | 81 | 4779 | 3481 | 6561 |

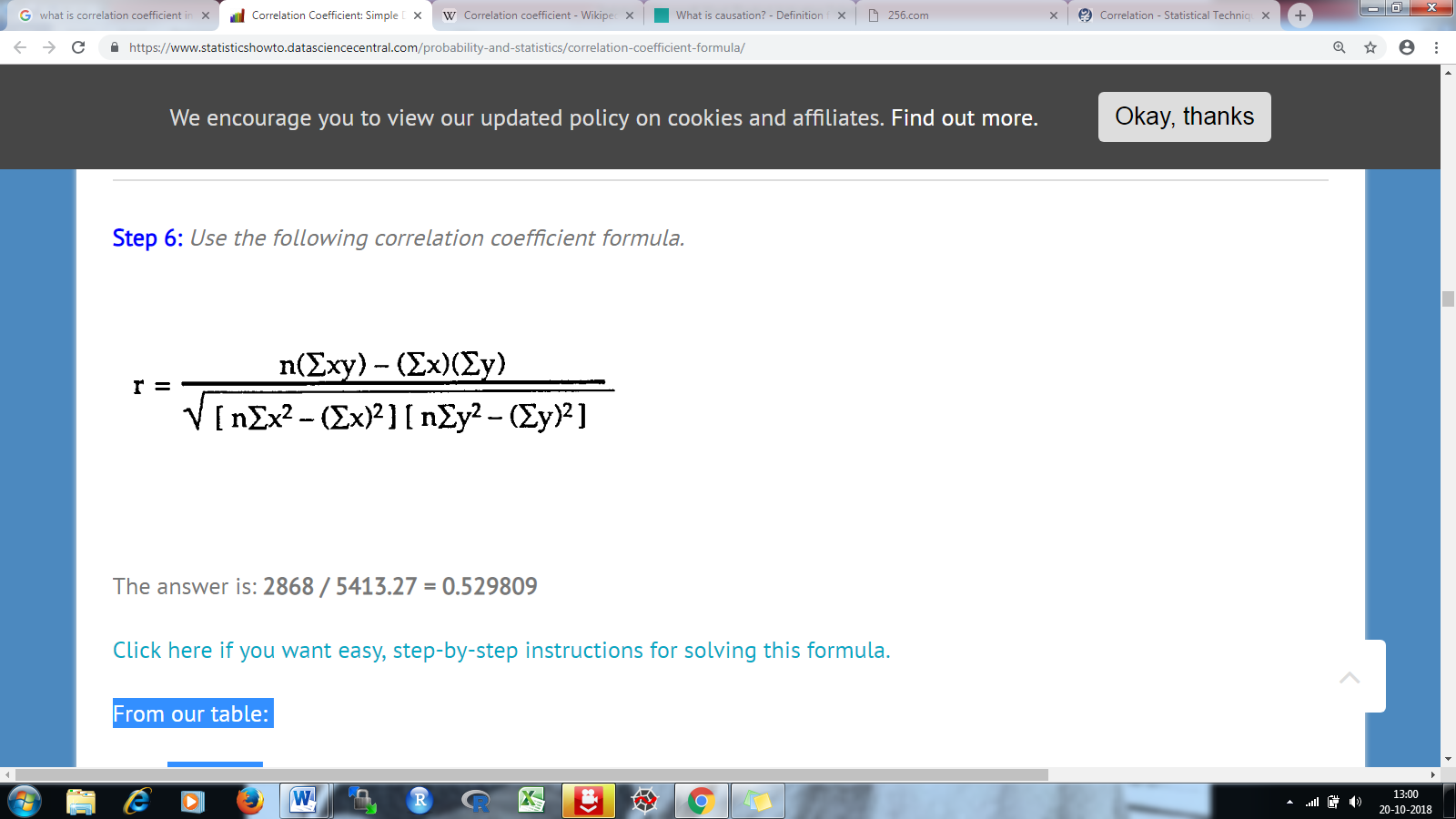
**Step 5:** *Add up all of the numbers in the columns and put the result at the bottom of the column.* The Greek letter sigma (Σ) is a short way of saying “sum of.”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUBJECT | AGE X | GLUCOSE LEVEL Y | XY | X2 | Y2 |
| 1 | 43 | 99 | 4257 | 1849 | 9801 |
| 2 | 21 | 65 | 1365 | 441 | 4225 |
| 3 | 25 | 79 | 1975 | 625 | 6241 |
| 4 | 42 | 75 | 3150 | 1764 | 5625 |
| 5 | 57 | 87 | 4959 | 3249 | 7569 |
| 6 | 59 | 81 | 4779 | 3481 | 6561 |
| Σ | 247 | 486 | 20485 | 11409 | 40022 |

**Step 6:**Use the following correlation coefficient formula.

From our table:

* Σx = 247
* Σy = 486
* Σxy = 20,485
* Σx2 = 11,409
* Σy2 = 40,022
* n is the [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/), in our case = 6



The answer is: **2868 / 5413.27 = 0.529809**

The correlation coefficient =

6(20,485) – (247 × 486) / [√[[6(11,409) – (2472)] × [6(40,022) – 4862]]]

= 0.5298

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.

Positive correlation range = 0 to 1

Negative correlation range = 0 to -1

### Coefficient of Determination

The **coefficient of determination** (denoted by R2) is a key output of [regression](https://stattrek.com/Help/Glossary.aspx?Target=Regression) analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

* The coefficient of determination is the square of the [correlation](https://stattrek.com/Help/Glossary.aspx?Target=Correlation) (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
* An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.
* An R2 of 1 means the dependent variable can be predicted without error from the independent variable.
* An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R2 of 0.20 means that 20 percent is predictable; and so on.

### Residual

In [regression analysis](https://stattrek.com/Help/Glossary.aspx?Target=Regression), the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the **residual** (e). Each data point has one residual.

Residual = Observed value - Predicted value   
e = y - ŷ

Both the sum and the mean of the residuals are equal to zero. That is, Σ e = 0 and e = 0.

### Residual Plot

A **residual plot** is a graph that shows the [residuals](https://stattrek.com/Help/Glossary.aspx?Target=Residual) on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model.

