Statistics

1.**Random variable**

**Types of random variable:**

i)Numerical random variable

a)Discrete random variable

b)continuous random variable

ii)Categorical random variable

eg:dataset with categorical column contains categorical random variable like in gender.

a)Discrete Random variable

Discrete are those random variable which takes a whole number,it will not be a floating point no. (eg:no. of bank accts of a person so it will be a whole no.)

b)continuous random variable

Within a range of values we can have any value like 10-15 it can be whole no or floating point no.

For eg:height =5.8 inches ,5.9 inches

2.**Gaussian distribution/Normal distribution**

X ∝G.D(μ, σ)

μ –mean

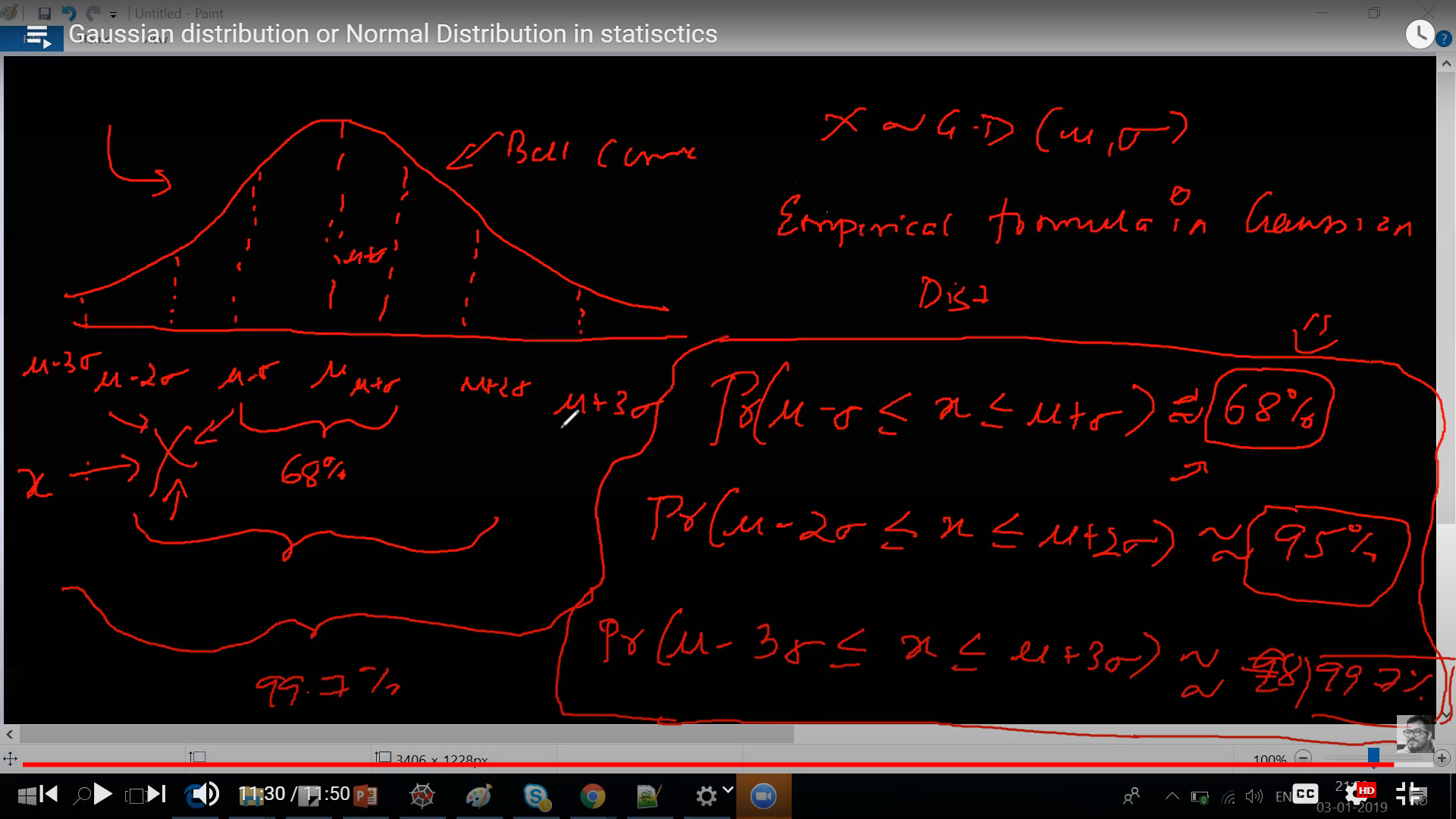
σ – standard deviation

μ=1/n( i=1 to n) Σ Xi

variance=1/n(i=1 to n) Σ (Xi – μ)2

σ=√var

standard deviation and mean actually helps us to specify from the mean how far is the other element distributed whether it is one standard deviation to right or one standard deviation to the left.



x🡪X if this random variable follows gaussian distribution it will form a bell curve

probabilty of x between the first standard deviation to the left and right is 68%

probabilty of x between the second standard deviation to the left and right is 95%

probabilty of x between the third standard deviation to the left and right is 99.7%

3.**Log normal distribution**

X∝Log normal distribution

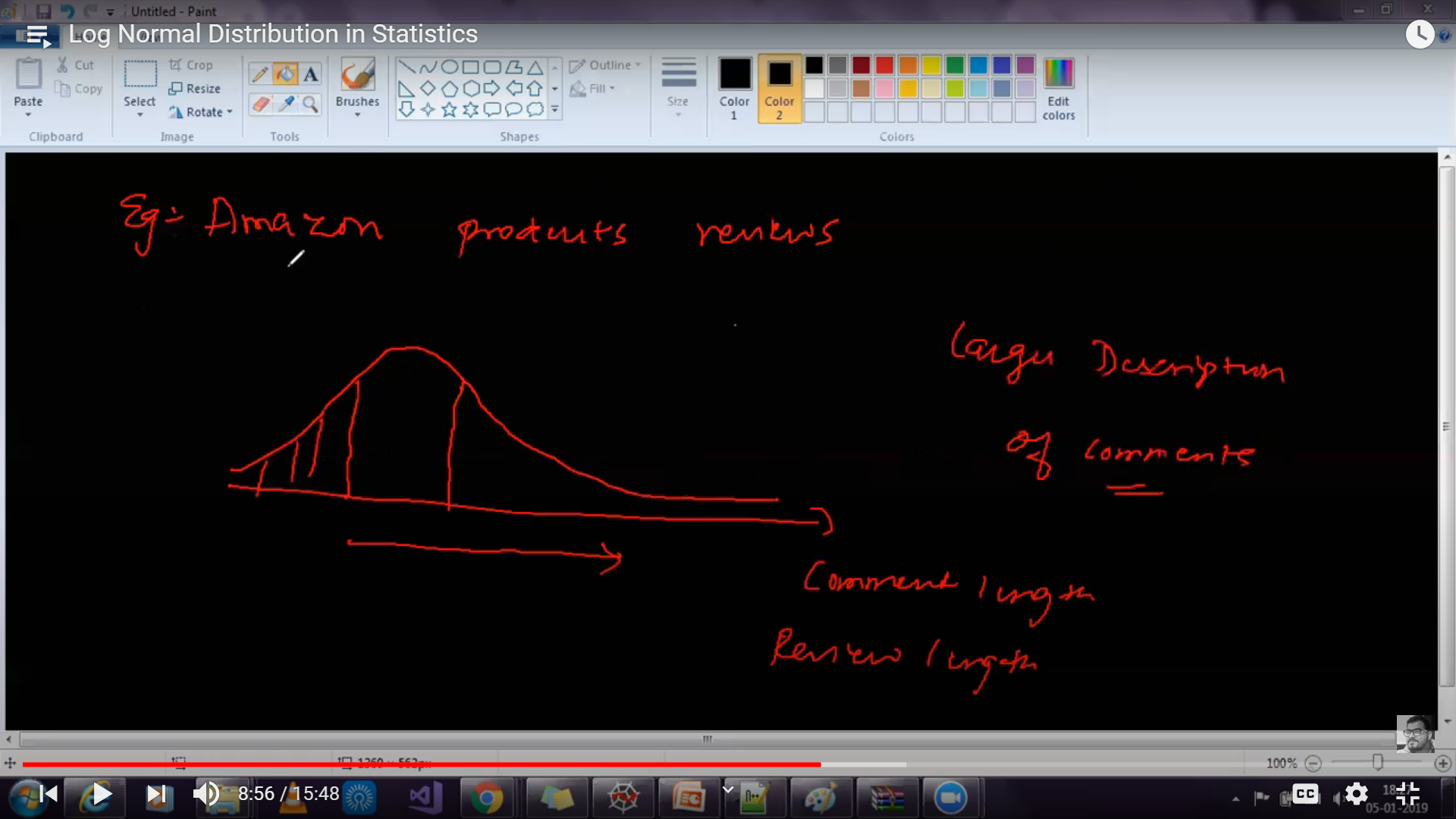
If ln(x) ∝G.D(μ, σ)

X={x1,x2,x3………….xn}

Ln(x1),ln(x2),ln(x3)

Gaussian distribution-(use case)-height datacollection will follow G.D

Log normal distribution-income of the people,amazon product reviews,sentiment analysis



Now Why are we learning distribution?

If we have a data with R&D spent,Marketing ,campaingn features for target value profit so if our R&D spent follows gaussian distribution what can we do is convert it into Standard normal distribution(mean=0,standard deviation=1).

Gaussian => Standard normal distribution

Snd=(Xi- μ)/σ

Now in marketing column if we know by our domain knowledge that it is following log normal distribution we will find the log of each value and we know it follows log normal distribution so log values follows gaussian distribution then we can convert it into standard normal distribution and now all values are standard scaled so if I feed this data into my model it will give me better accuracy and with fine tuning we will get better accuracy.

4. **Covariance**

In probability theory and statistics, covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, the covariance is positive.

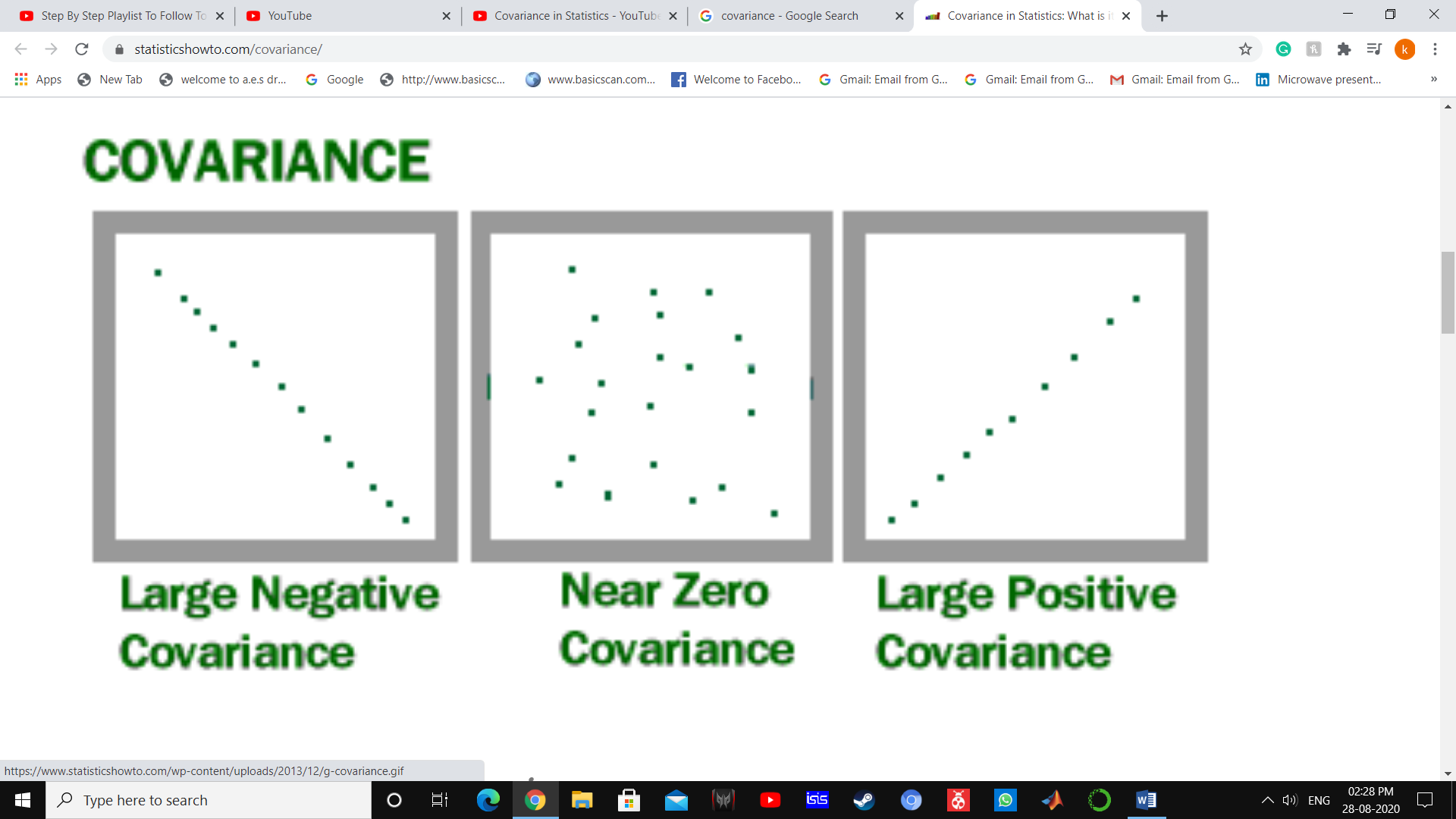
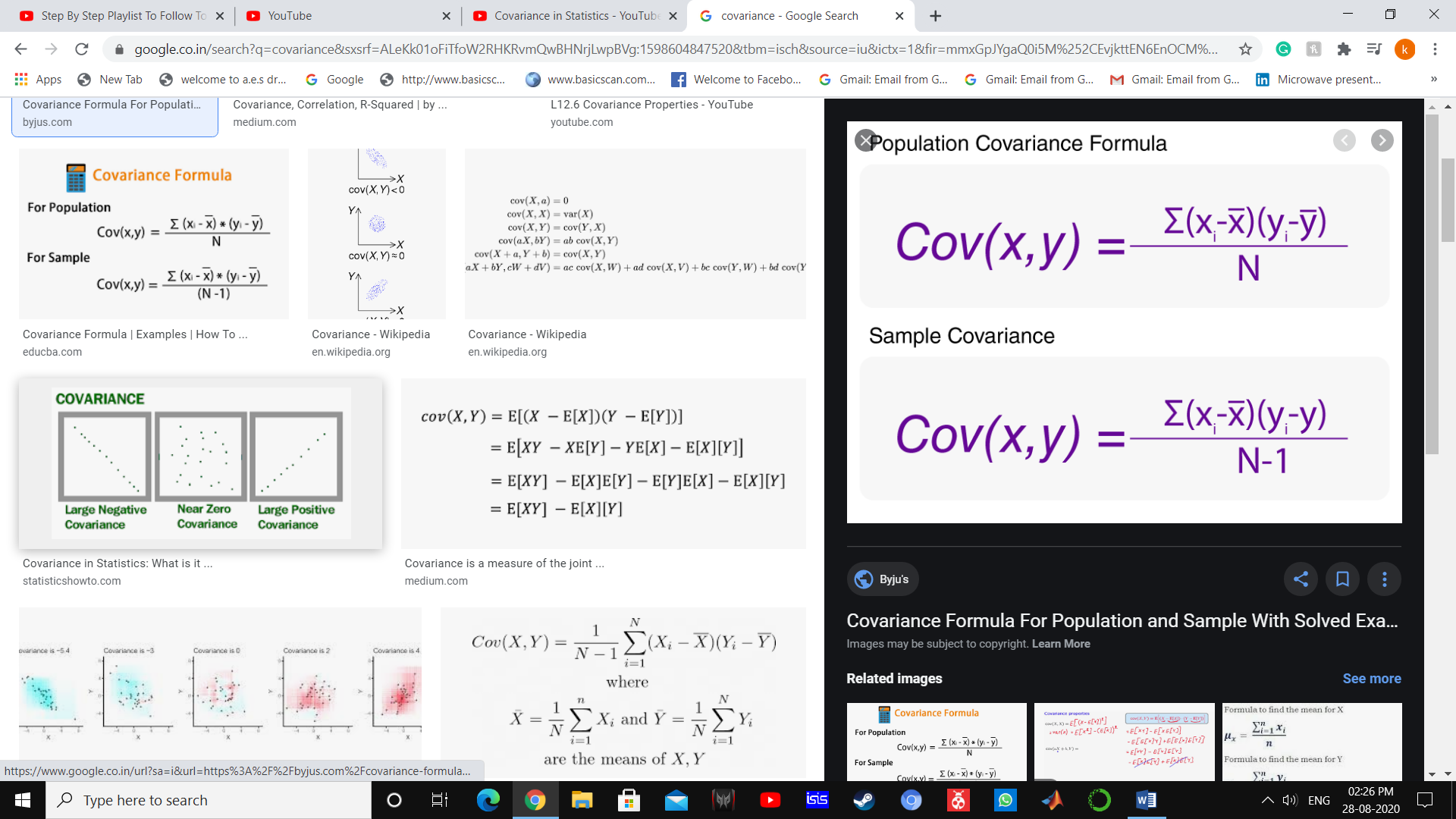
**Cov(x,y) =1/n i=1Σn (xi – μx) X (yi – μy)**

Var(x)=1/n **i=1Σn**(xi – μx)2

=1/n **i=1Σn**(xi – μx)X(xi – μx)

Cov(x,x)=var(x)

**Covariance** is a statistical tool that is used to determine the relationship between the movement of two asset prices. When two stocks tend to move together, they are seen as having a positive **covariance**; when they move inversely, the **covariance** is negative.



**5.Population & Sample**

The main difference between a **population** and **sample** has to do with how observations are assigned to the **data** set. A **population** includes all of the elements from a set of **data**. A **sample** consists one or more observations drawn from the **population**.

Population=N

Population mean = μ

Sample = n

Sample mean= X̄

Population mean:

μ =1/N\***i=1ΣN(xi)**

Sample mean:

X̄=1/n\***i=1Σn(xi)**

**Example: Exit poll not taking whole population data taking a small chunk of data (sample) to get mean of results.**

**Central Limit Theorem**

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population [**with replacement**](javascript:void(0);)https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/ada-reference.gif, then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n > 30). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that min(np, n(1-p))> 5, where n is the sample size and p is the probability of success in the population. 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:

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/ada-reference.gifhttps://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/lessonimages/equation_image123.gif

and the standard deviation of the sample means:

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/ada-reference.gifhttps://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/lessonimages/equation_image124.gif

Before illustrating the use of the Central Limit Theorem (CLT) we will first illustrate the result. In order for the result of the CLT to hold, the sample must be sufficiently large (n > 30). Again, there are two exceptions to this. If the population is normal, then the result holds for samples of any size (i..e, the sampling distribution of the sample means will be approximately normal even for samples of size less than 30).

**Chebysev’s Inequality**

chebeshev's inequality is basically defined for terms where variance is quite low so intuitively to say when your variance is low the most of data points are likely to fall closer to the mean of observed data when your data is not distributed normally the data point's likelihood of occurrence or the random variable occurrence of a specific data point in that dataset or random variable set can be estimated by chebshev's inequality.

Y ≠ Gaussian distribution

Pr(μ-kσ<x< μ+ kσ)>1-1/k2

K=2

Pr(μ-2σ<x< μ+2σ)>1-1/22

>75%

K=3

Pr(μ-3σ<x< μ+3σ)>1-1/32

>85%

**Pearson Correlation Coefficient vs Covariance**

**Covariance=Cov(x,y) =1/n i=1Σn (xi – μx) X (yi – μy)**

Pearson CC= **ρ**(x,y)=Cov(x,y)/ σx σy

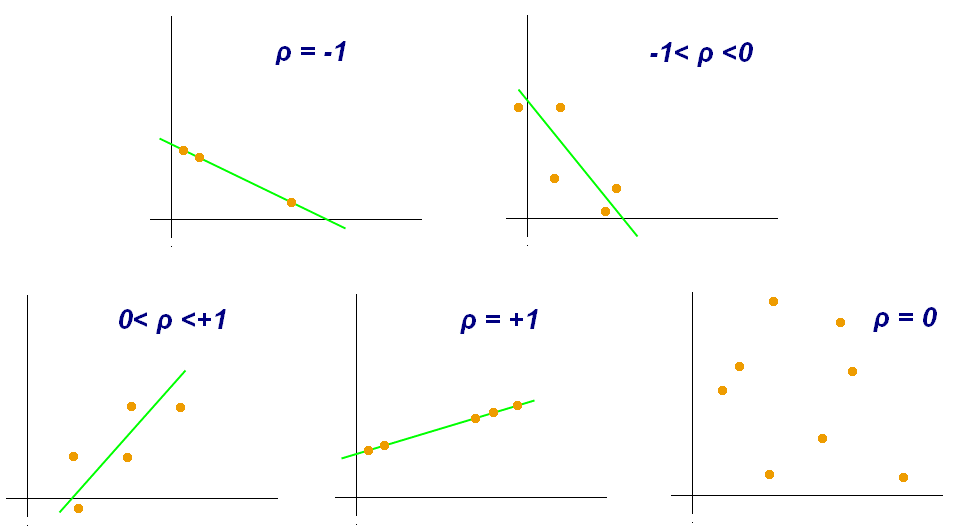
-> Covariance is helping us to find the direction of relationship.

X **↑** Y **↑ =** +ve Covariance

X **↑** Y ↓ = -ve Covariance

->Pearson Correlation coefficient helps us to fint the magnitude and the direction of relationship or correlation

-1<**ρ**(x,y)<1



So if X and Y are correlated with magnitude 1 so we know both variable are same so we can drop 1 of them to find the target variable.

**Spearman rank correlation coefficient**

The Spearman correlation between two variables is equal to the [Pearson correlation](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient) between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

The Spearman correlation coefficient is defined as the [Pearson correlation coefficient](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient) between the [rank variables](https://en.wikipedia.org/wiki/Ranking).[[3]](https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient#cite_note-myers2003-3)

For a sample of size *n*, the *n* [raw scores](https://en.wikipedia.org/wiki/Raw_score) {\displaystyle X\_{i},Y\_{i}}are converted to ranks {\displaystyle \operatorname {rg} \_{X\_{i}},\operatorname {rg} \_{Y\_{i}}}, and {\displaystyle r\_{s}}is computed as

rs =ρ rgx, rgy =cov(rgX, rgY)/ σ rgx σ rgy

ρ- denotes the usual [Pearson correlation coefficient](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient), but applied to the rank variables

cov(rgX, rgY)- is the [covariance](https://en.wikipedia.org/wiki/Covariance) of the rank variables,

σ rgx σ rgy are the [standard deviations](https://en.wikipedia.org/wiki/Standard_deviation) of the rank variables.

rs =1- 6Σdi2/n(n2-1)

di= rg(Xi) – rg(Yi)

| [**IQ**](https://en.wikipedia.org/wiki/IQ)**, {\displaystyle X\_{i}}** | **Hours of**[**TV**](https://en.wikipedia.org/wiki/TV)**per/week, {\displaystyle Y\_{i}}** |
| --- | --- |
| 106 | 7 |
| 100 | 27 |
| 86 | 2 |
| 101 | 50 |
| 99 | 28 |
| 103 | 29 |
| 97 | 20 |
| 113 | 12 |
| 112 | 6 |
| 110 | 17 |

Firstly, evaluate di2 {\displaystyle d\_{i}^{2}}.To do so use the following steps, reflected in the table below.

1. Sort the data by the first column ({\displaystyle X\_{i}} Xi). Create a new column {\displaystyle x\_{i}}and assign it the ranked values 1, 2, 3, ..., *n*.
2. Next, sort the data by the second column ({\displaystyle Y\_{i}}Yi). Create a fourth column {\displaystyle y\_{i}}and similarly assign it the ranked values 1, 2, 3, ..., *n*.
3. Create a fifth column di {\displaystyle d\_{i}}to hold the differences between the two rank columns ({\displaystyle x\_{i}}xi and {\displaystyle y\_{i}}yi).
4. Create one final column {\displaystyle d\_{i}^{2}}di2 to hold the value of column {\displaystyle d\_{i}}di squared.

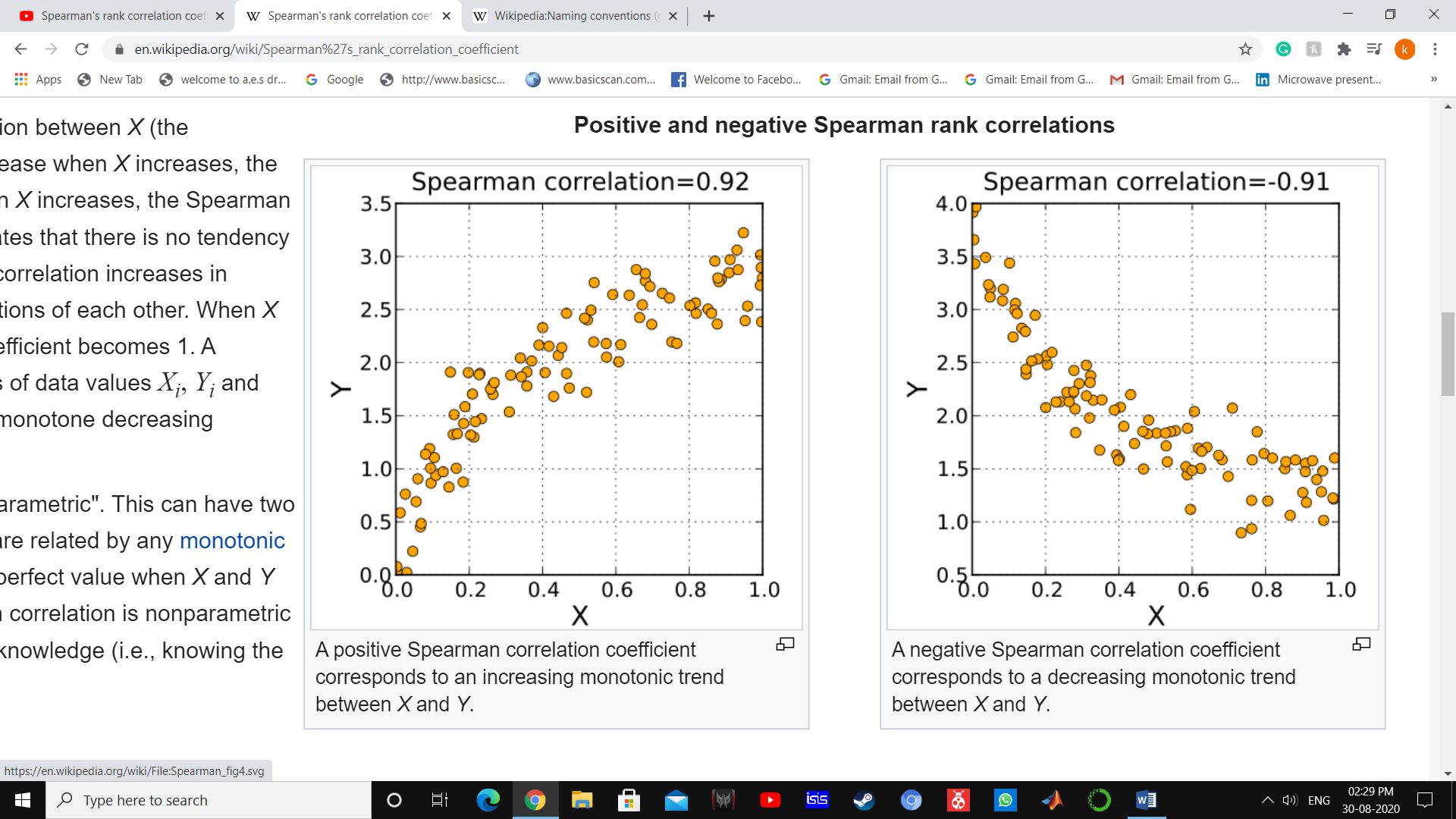
| [**IQ**](https://en.wikipedia.org/wiki/IQ)**,  {\displaystyle X\_{i}}** | **Hours of**[**TV**](https://en.wikipedia.org/wiki/TV)**per week, {\displaystyle Y\_{i}}** | **Rank(Xi) {\displaystyle x\_{i}}** | **Rank(Yi) {\displaystyle y\_{i}}** | **{\displaystyle d\_{i}}di** | **{\displaystyle d\_{i}^{2}}dd d**  **di2** |
| --- | --- | --- | --- | --- | --- |
| 86 | 2 | 1 | 1 | 0 | 0 |
| 97 | 20 | 2 | 6 | −4 | 16 |
| 99 | 28 | 3 | 8 | −5 | 25 |
| 100 | 27 | 4 | 7 | −3 | 9 |
| 101 | 50 | 5 | 10 | −5 | 25 |
| 103 | 29 | 6 | 9 | −3 | 9 |
| 106 | 7 | 7 | 3 | 4 | 16 |
| 110 | 17 | 8 | 5 | 3 | 9 |
| 112 | 6 | 9 | 2 | 7 | 49 |
| 113 | 12 | 10 | 4 | 6 | 36 |

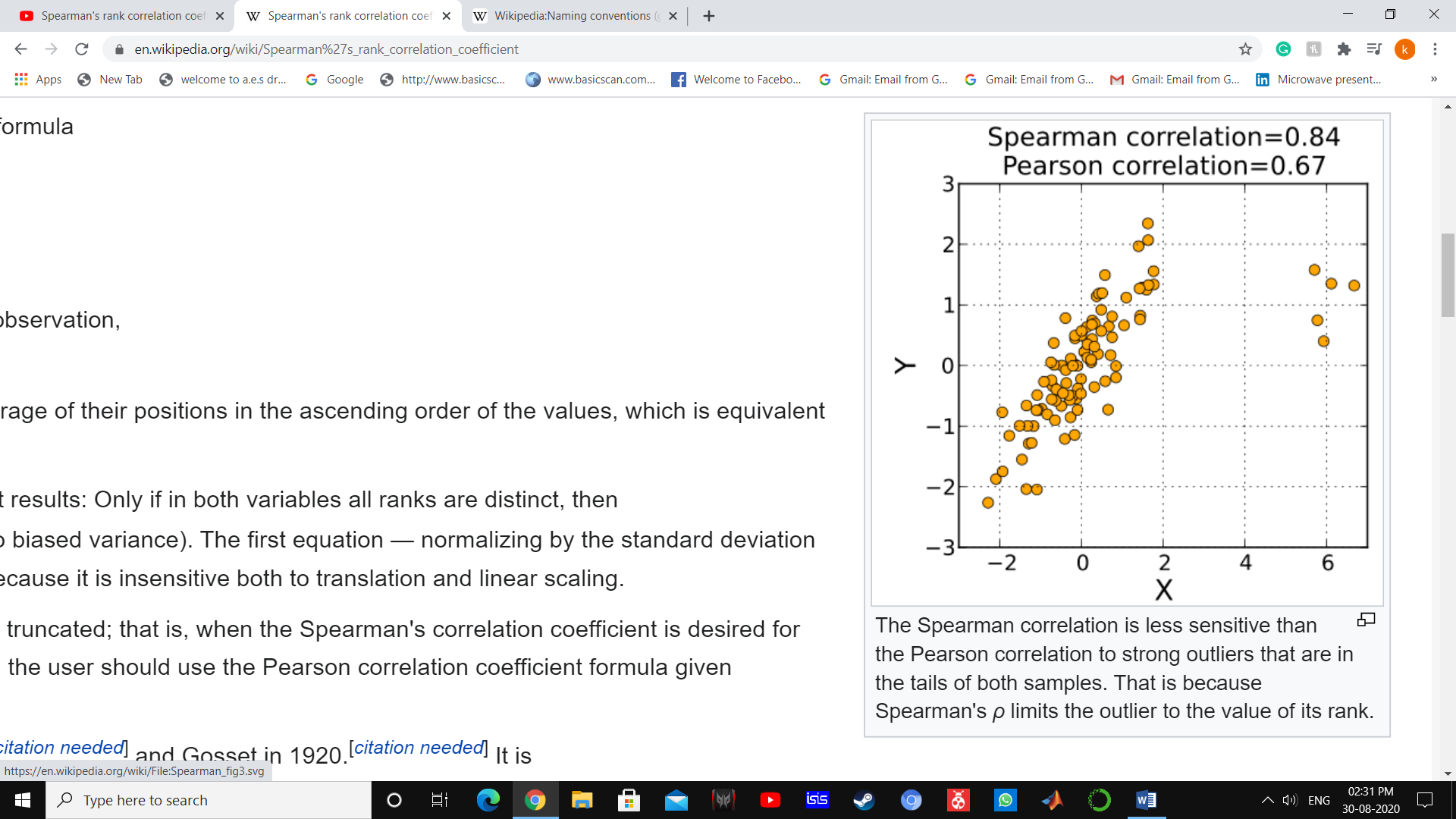
With {\displaystyle d\_{i}^{2}}di2 found, add them to find i=1Σn di2=194{\displaystyle \sum d\_{i}^{2}=194}  
. The value of *n* is 10. These values can now be substituted back into the equation

ρ =1- 6Σdi2/n(n2-1)

ρ=1- 6\*194/10(102-1)

which evaluates to *ρ* = −29/165 = −0.175757575





{\displaystyle \rho =1-{\frac {6\sum d\_{i}^{2}}{n(n^{2}-1)}}}

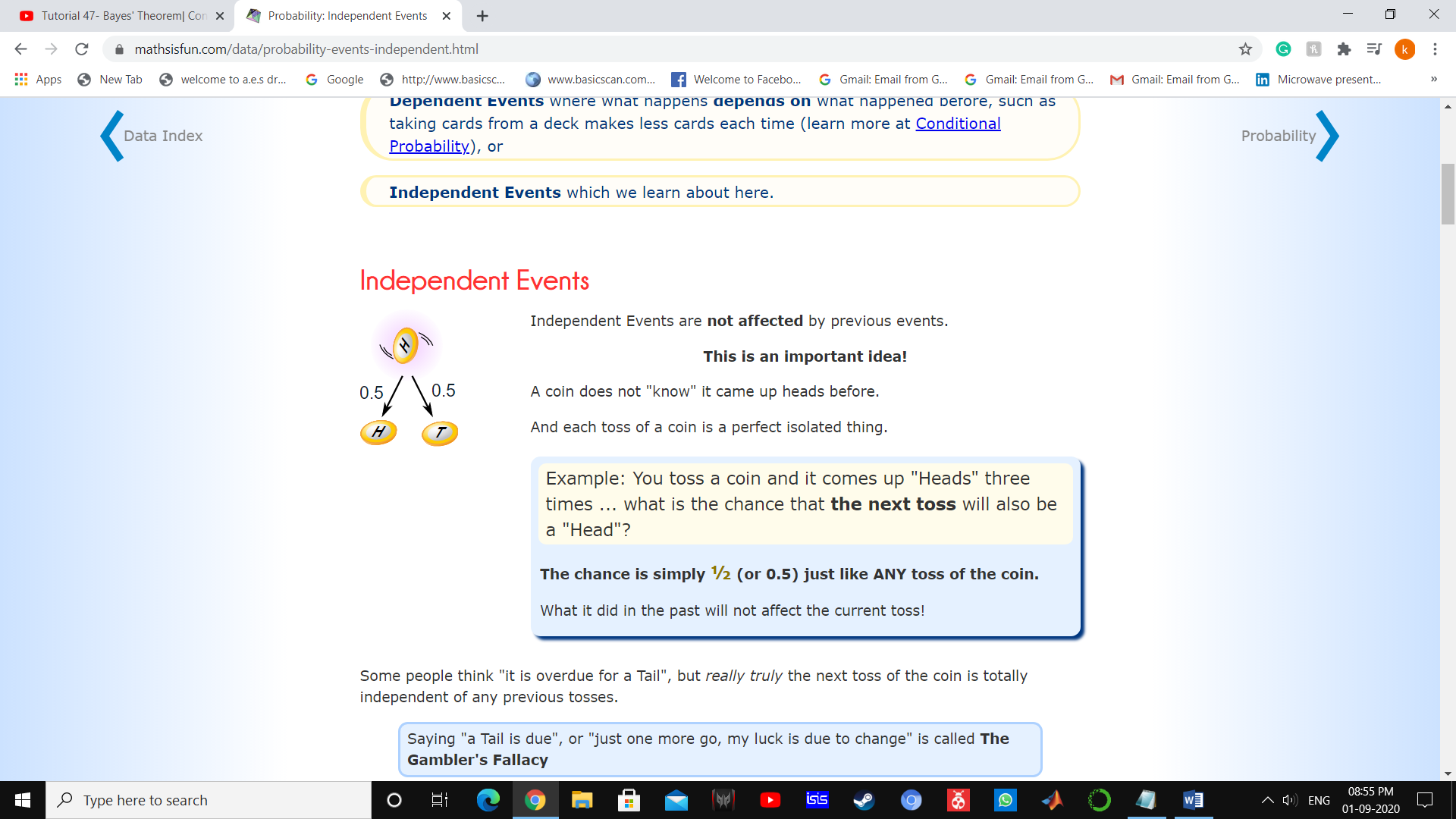
**Bayes’s Theorem**

1)conditional Probability

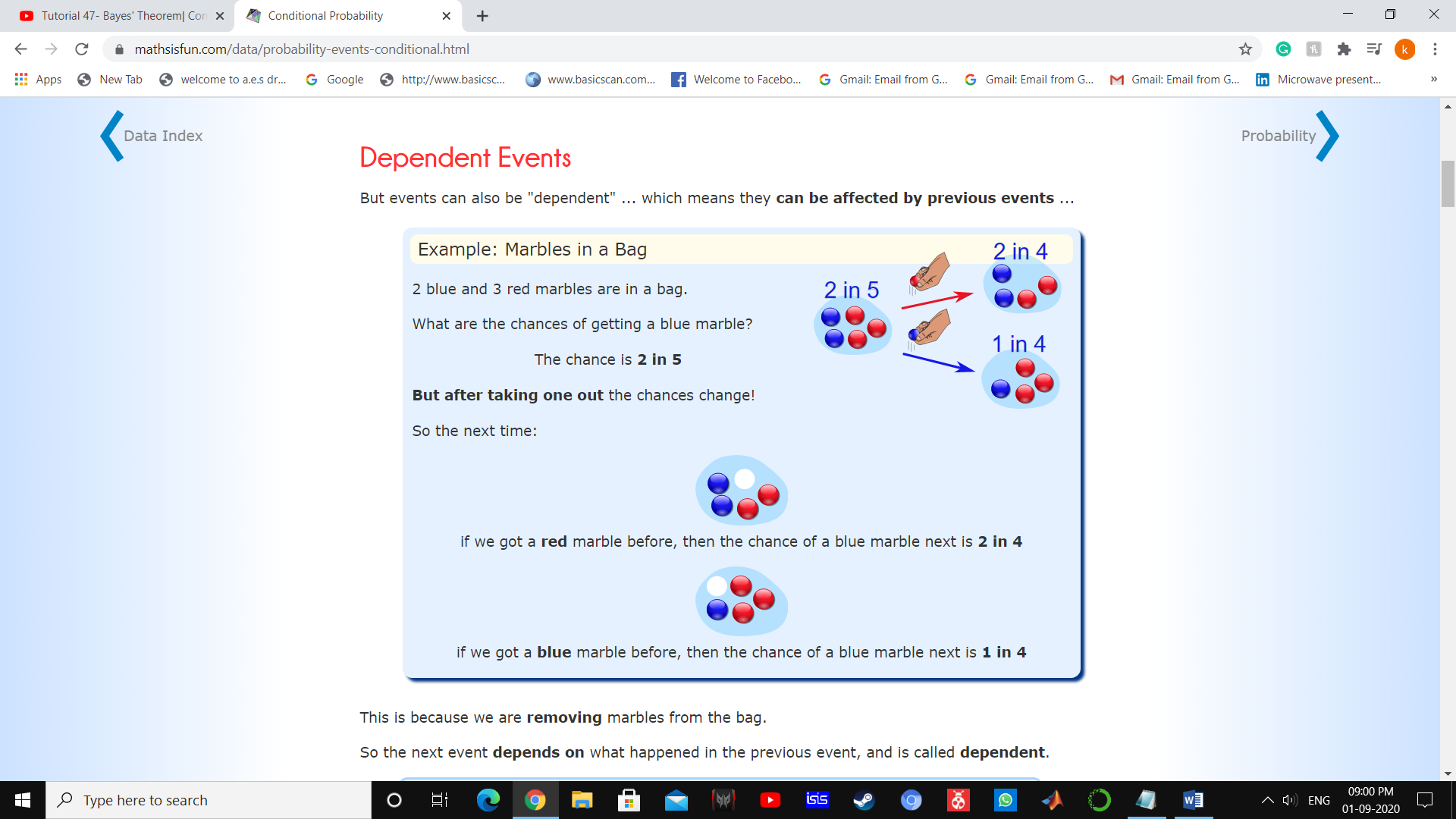
P(A|B)= P(A⋂B)/P(B)

A:Event 1 B:Event 2

2)Independent Events



3)dependent Events



P(A)=2/5

P(B|A)=1/4

P(A⋂B)=2/5\*1/4=1/10

Bayes Theorem

P(A|B)=P(A⋂B)/P(B)

P(B|A)=P(B⋂A)/P(A)

P(A⋂B)=P(B⋂A)

P(A|B)\*P(B)=P(B|A)\*P(A)

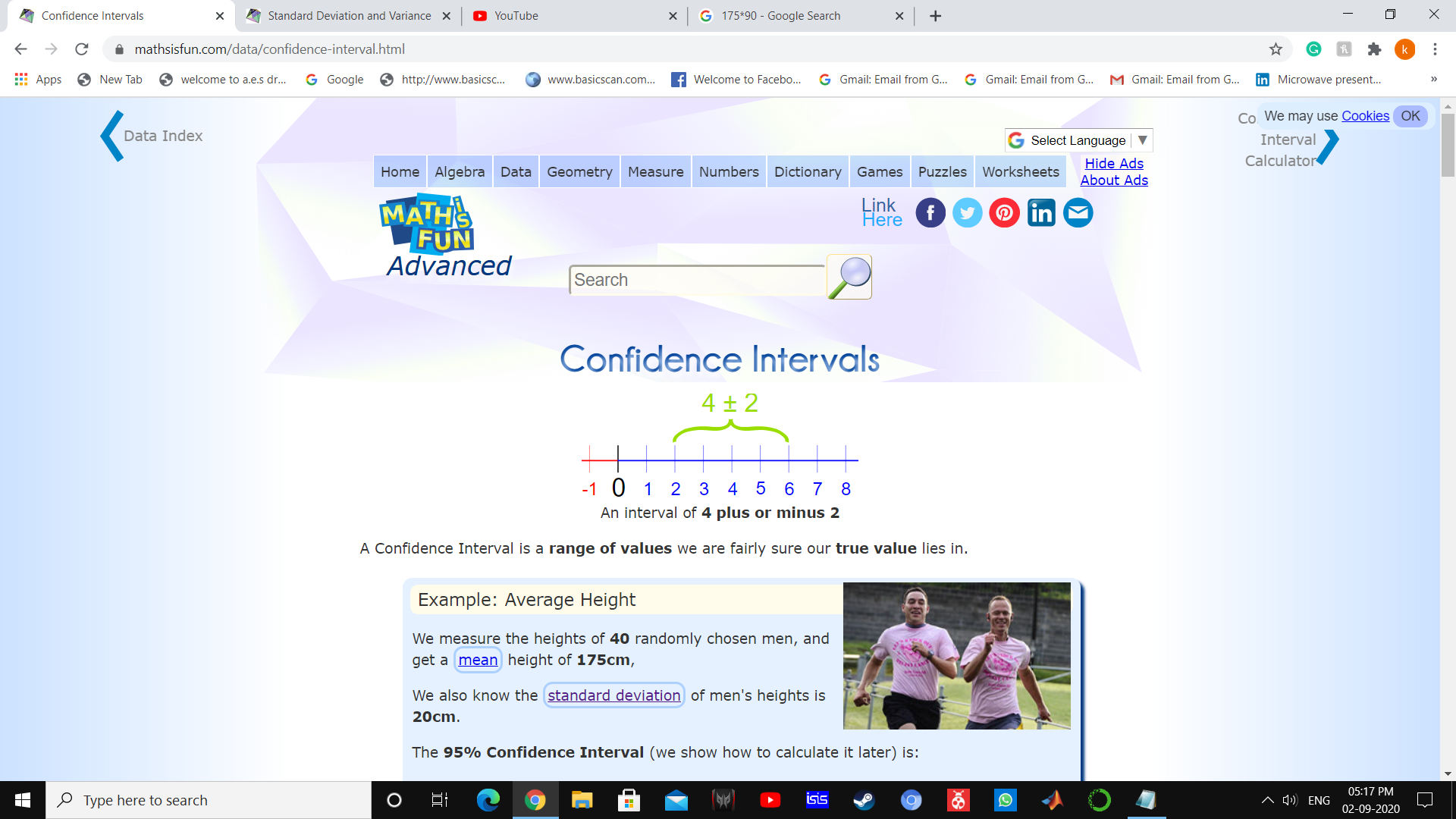
P(A|B)=P(B|A)\*P(A)/P(B)

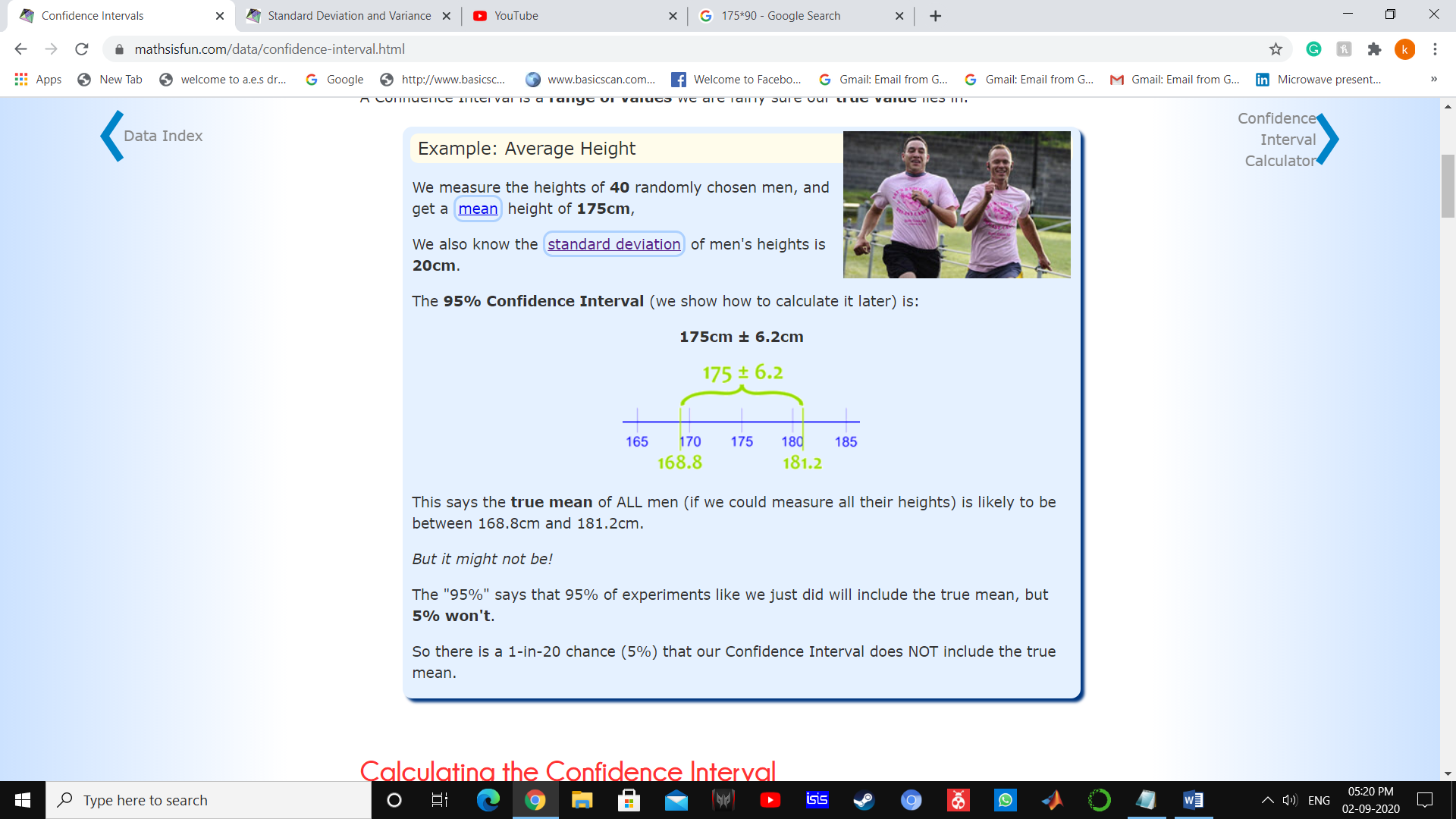
P(A|B): Posterior Probability

P(B|A):Likelihood Probability

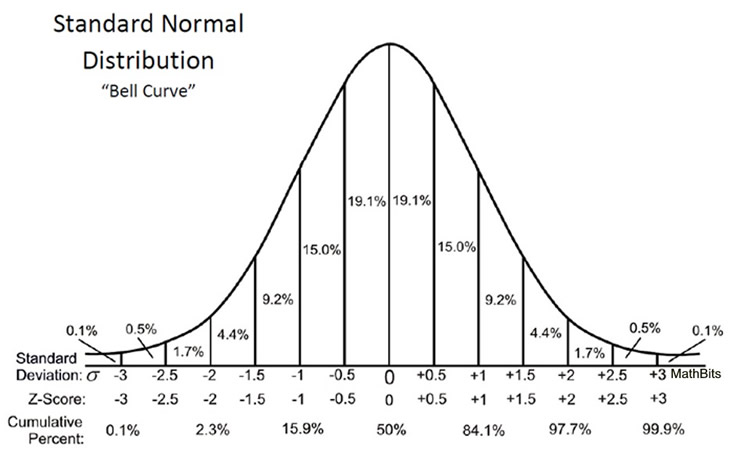
P(A):Prior prob

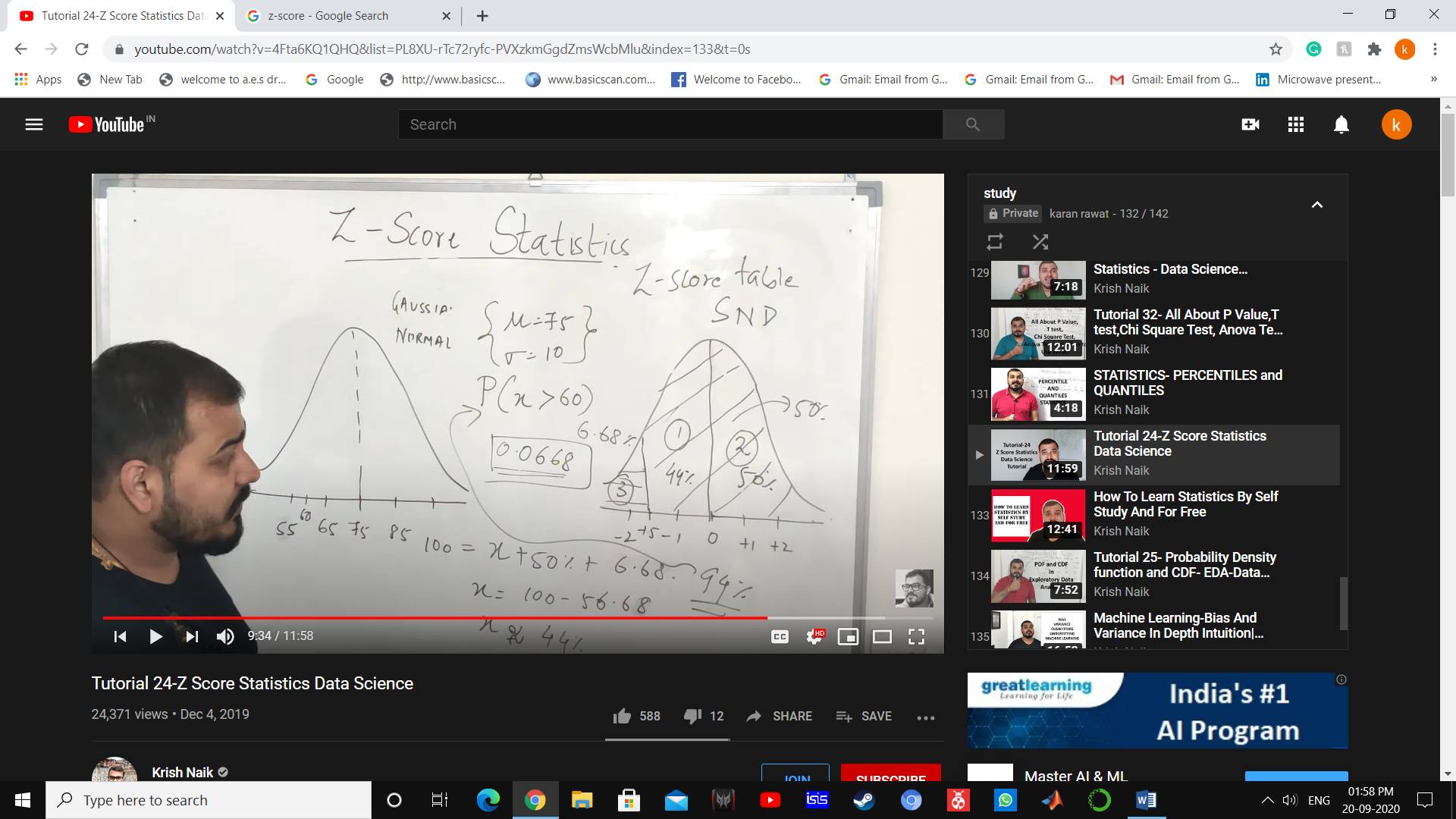
P(B):Margin prob





**Z-Score Statistics**





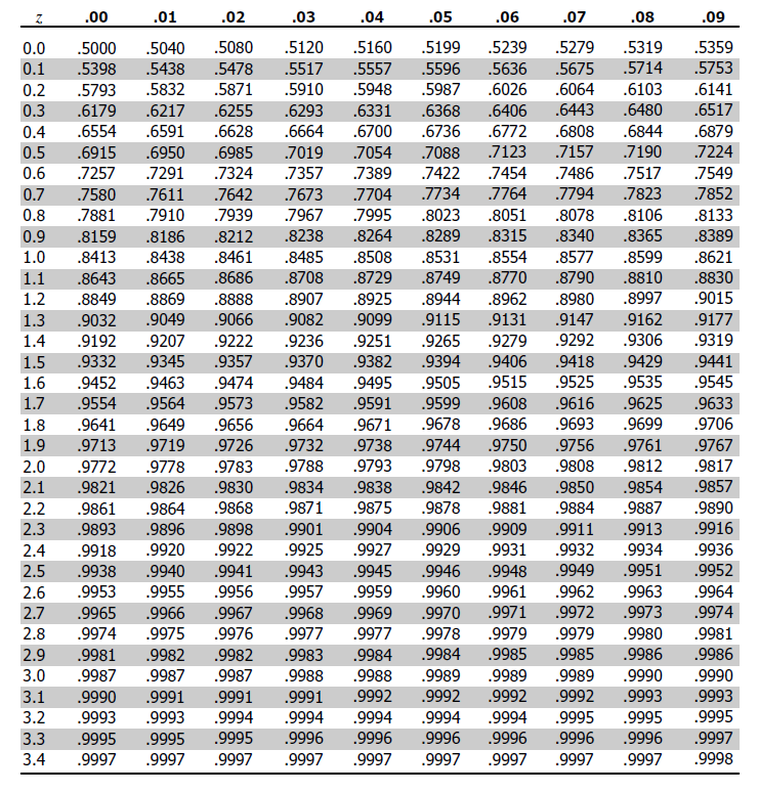
μ=75 , σ=10

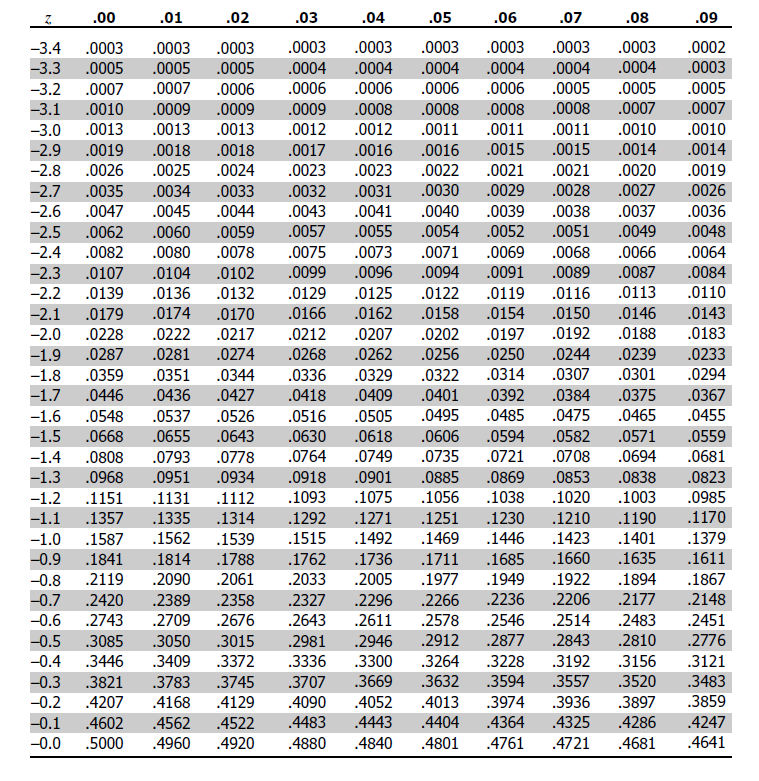
P(x>60)

6.7% less than -1.5

100=x+50%+6.7%

X=44%



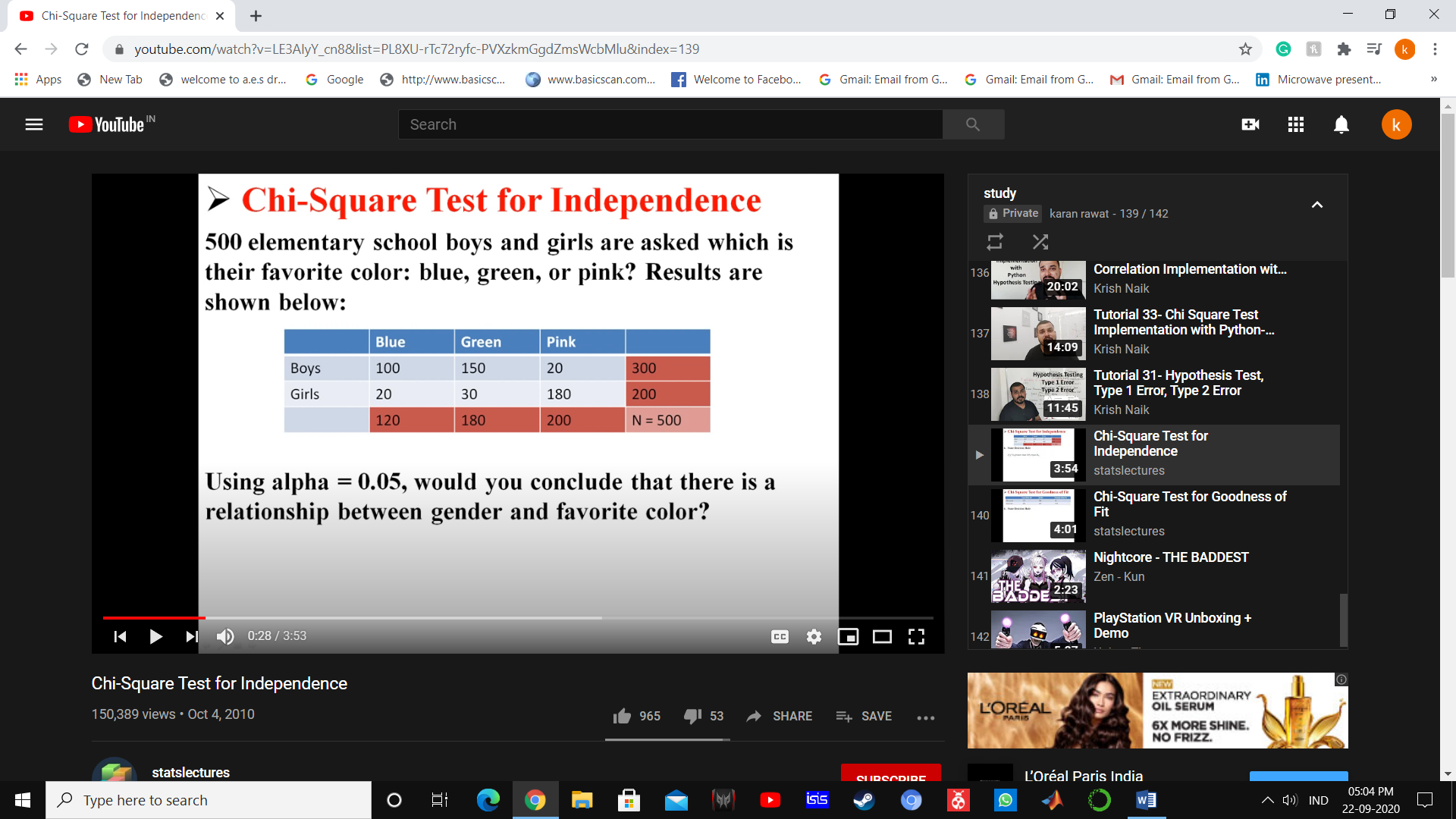


-1.5 left in SND =.0668

**Chi square test for independence**

The Chi square test for independence evaluates the relation between two variables.

It is a non parametric test that is performed on categorical (nominal or ordinal) data.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Blue | Green | Pink |  |
| Boys | 100 | 150 | 20 | 270 |
| Girls | 20 | 30 | 180 | 230 |
|  | 120 | 180 | 200 | N=500 |

1.Define null and alternative hypothesis

H0=null hypothesis

H1=alternative hypothesis



2.State Alpha

Alpha=0.05

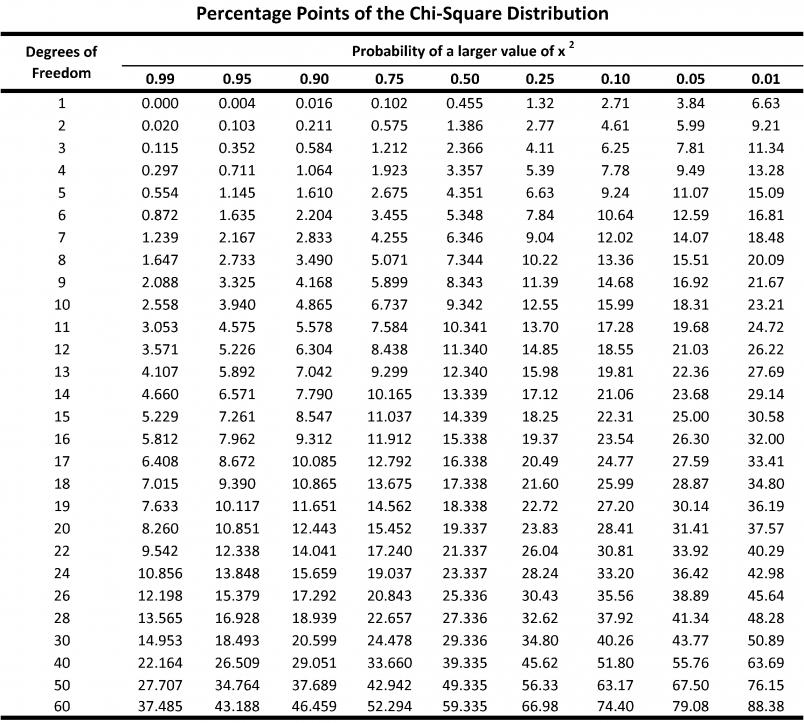
3.Calculate degrees of freedom

Df=(rows-1)(columns-1)

Df=(2-1)(3-1)

Df=(1)(2)=2

4.State decision rule



So for df=2 and alpha=0.05

If X2 is greater than 5.99,reject H0

5.Calculate Test Statistic

f0=frequency of original value

fe=frequency of expected value

X2= Σ(f0 –fe)2/fe

fc=frequency of column

fr=frequency of rows

fe =fcfr/n

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Blue | Green | Pink |  |
| Boys | 100(64.8) | 150(97.2) | 20(108) | 270 |
| Girls | 20(55.2) | 30(82.8) | 180(92) | 230 |
|  | 120 | 180 | 200 | N=500 |

X2= (100-64.8)2/64.8 +(150-97.2)2/97.2 +(20-108)2/108 +(20-55.2)2/55.2 +(30-82.8)2/82.8 +(180-92)2/92

X2=19.12+28.68+71.7+22.44+33.66+84.17=259.77

X2=259.77

6.state results

So X2 is greater than 5.99

So we reject null hypothesis

7.State Conclusion

In the population ,there is a relationship between gender and favourite color .