Saturday, August 04, 2018 6:48 PM

1. **Residuals**

In **regression** analysis, 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.

From <https://www.google.co.in/search?ei=0p5lW8zDCcyKvQTGz5SoAQ&q=residual+vs+fitted+what+is+fitted+&oq=residual+vs+fitted+what+is+fitted+&gs_l=psy-ab.3..33i160k1.2683886.2703276.0.2704388.37.33.0.4.4.0.449.3701.0j16j2j1j1.20.0....0...1c.1.64.psy-ab..13.23.3558...0j0i22i30k1j33i22i29i30k1j33i21k1.0.P3AGo8m5BgQ>

2. **Fitted**

A **fitted value is** simply another name for a predicted **value** as it describes where a particular x-**value** fits the line of best fit. It **is** found by substituting a given **value** of x into the regression equation . A residual denoted (e) **is** the difference or error between an observed observation and a predicted or fit **value**.

From <https://www.google.co.in/search?ei=0p5lW8zDCcyKvQTGz5SoAQ&q=residual+vs+fitted+what+is+fitted+&oq=residual+vs+fitted+what+is+fitted+&gs_l=psy-ab.3..33i160k1.2683886.2703276.0.2704388.37.33.0.4.4.0.449.3701.0j16j2j1j1.20.0....0...1c.1.64.psy-ab..13.23.3558...0j0i22i30k1j33i22i29i30k1j33i21k1.0.P3AGo8m5BgQ>

3. **q-q plot**

The **quantile-quantile** (**q-q**) **plot** is a graphical technique for determining if two data sets come from populations with a common distribution. A **q-q plot** is a **plot** of the quantiles of the first data set against the quantiles of the second data set.

From <https://www.google.co.in/search?ei=0p5lW8zDCcyKvQTGz5SoAQ&q=residual+vs+fitted+what+is+fitted+&oq=residual+vs+fitted+what+is+fitted+&gs_l=psy-ab.3..33i160k1.2683886.2703276.0.2704388.37.33.0.4.4.0.449.3701.0j16j2j1j1.20.0....0...1c.1.64.psy-ab..13.23.3558...0j0i22i30k1j33i22i29i30k1j33i21k1.0.P3AGo8m5BgQ>

4. **Latent variables** are not observed directly (construct/ dependent variable/ response/ Endogenous)
5. **Measured**(manifest / Observed/ indicator/ predictor/ Exogenous)
6. **Degree of freedom(df)**

In a general sense, DF are the number of observations in a sample that are free to vary while estimating statistical parameters. You can also think of it as the amount of independent data that you can use to estimate a parameter.

From <http://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/>

Degrees of freedom also define the probability distributions for the test statistics of various hypothesis tests. For example, hypothesis tests use the t-distribution, F-distribution, and the chi-square distribution to determine statistical significance. Each of these probability distributions is a family of distributions where the degrees of freedom define the shape. Hypothesis tests use these distributions to calculate p-values. So, the DF are directly linked to p-values through these distributions!

From <http://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/>

7. **P Value**
   A small **p-value** (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. A large **p-value** (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.

From <https://www.google.co.in/search?q=what+is+z-score+and+p-value&oq=what+is+z+value+and+p+&aqs=chrome.1.69i57j0.10344j0j7&sourceid=chrome&ie=UTF-8>
In the majority of analyses, an alpha of **0.05** is used as the cutoff for significance. If the **p**-value is less than **0.05**, we reject the null hypothesis that there's no difference between the **means** and conclude that a significant difference **does**exist.

From <https://www.google.co.in/search?q=what+is+z-score+and+p-value&oq=what+is+z+value+and+p+&aqs=chrome.1.69i57j0.10344j0j7&sourceid=chrome&ie=UTF-8>

8. **Z score**
   The absolute **value** of the **z-score** tells you how many standard deviations you are away from the mean. If a **z-score** is equal to 0, it is on the mean. If a **Z-Score** is equal to +1, it is 1 Standard Deviation above the mean. If a **z-score** is equal to +2, it is 2 Standard Deviations above the mean.

   From <https://www.google.co.in/search?q=what+is+z-score+and+p-value&oq=what+is+z+value+and+p+&aqs=chrome.1.69i57j0.10344j0j7&sourceid=chrome&ie=UTF-8>

9. **Standard Error**
   The **standard error** indicates the likely accuracy of the sample **mean** as compared with the population **mean**. SE= Standard deviation/ sqrt(n) n=sample size

   From <https://www.google.co.in/search?hl=en-IN&ei=aq5nW7TgMMPgvATD3p2IBQ&q=std+error+interpretation&oq=std.error+inter&gs_l=psy-ab.3.0.0i22i30k1l4.886001.903179.0.905071.12.12.0.0.0.0.299.1804.0j9j2.11.0....0...1c.1.64.psy-ab..1.11.1802...0j0i67k1j0i131k1j0i30k1j0i22i10i30k1.0.I8FLo4hQmBw>

Reshedule vs fitted

Saturday, August 04, 2018 7:13 PM

Residual = Observed y-value - Predicted y-value

A residual is the difference between the observed *y*-value (from scatter plot) and the predicted *y*-value (from regression equation line).
It is the vertical **distance** from the actual plotted point to the point on the regression line.
You can think of a residual as how far the data "fall" from the regression line (sometimes referred to as "*observed error*")

From <https://mathbitsnotebook.com/Algebra1/StatisticsReg/ST2Residuals.html>

Measure of dispersion, Range and IQR

Monday, December 6, 2021 4:11 PM

Measure of dispersion:
[Measures of Dispersion in Statistics (Definition & Types) (byjus.com)]

Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which a numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.

In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

The types of absolute measures of dispersion are:

1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
2. **Variance:** Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance. Variance $(\sigma 2)=\sum(X-\mu)2/N$
3. **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = $\sqrt{\sigma}$.
4. **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
5. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

Range & IQR:
[Statistics Measures of Dispersion Range Interquartile outliers | K2 Analytics]

In statistics, the range is one of the most common measures of dispersion. It is the difference between the largest and the smallest observation in the data distribution. The range has the same unit as the data variable.
Formula: For the values of X, the range is

Range = Largest Value of X – Smallest Value of X

Since the range takes only two values (largest and smallest), the extreme values/outliers will impact the range.
Let's compute the salary range of the employees in a small startup. The salary details of all employees including the founding members are given below:

Emp. No.	1	2	3	4	5	6	7	8	9	10
Monthly Salary (k)	90	80	18	18	17	16	16	16	15	14

- The range of the salary = 90 – 14 = 76
- But, the data distribution has two extreme values (90, 80). If we chop them, the range of the salary will be 4 (18 – 14).
- When there are extreme values/outliers, the range statistic can be misleading. Hence, we often

use the interquartile range instead of a range.

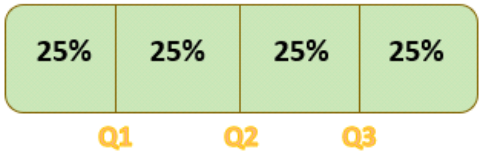
Interquartile Range (IQR) is the range of the middle 50% of the values in the data distribution. It is the difference between the third quartile(**Q3**) and the first quartile(**Q1**).

- **Formula:**

IQR = Q3 – Q1

Quartiles divide the rank-ordered data distribution into four equal parts. The values that separate parts are called the first, second, and third quartiles (*Wikipedia link*)

- **First Quartile (Q1):** It is the median of the lower half of the data distribution (25th percentile)
- **Second Quartile (Q2):** It is the median of the entire data distribution (50th percentile)
- **Third Quartile (Q3):** It is the median of the upper half of the data distribution (75th percentile)



Example

We will use the small start-up example having 10 employees as discussed earlier. The monthly salary of the employees is given in the table below. Find the quartiles and interquartile range of the salary.

Emp. No.	1	2	3	4	5	6	7	8	9	10
Monthly Salary (k)	90	80	18	18	17	16	16	16	15	14

Second Quartile

Let us first calculate the second quartile (Median).

- Sort the values in ascending order

14	15	16	16	16	17	18	18	80	90

- The number of observations, n=10 (even), therefore Q2 is mean of (n/2)th observation and ((n/2) + 1)th observation
- **Q2(median)** = **(1/2) * (5th observation + 6th observation)**
  Q2 = (16 + 17) / 2
  Q2 = 16.5

First Quartile

Now, let's calculate the first quartile (Q1)

- **Q2 is the median**. It splits the dataset into the upper and lower half of the distribution.
- **Q1** is the median of the lower half of the distribution (90,80,18,18,17). The number of observations is 5, it is an odd number. As such Q1 is the value at 3rd position, (n+1) / 2.
- **Q1** = Value at 3rd observation
  Q1 = 16

Third Quartile

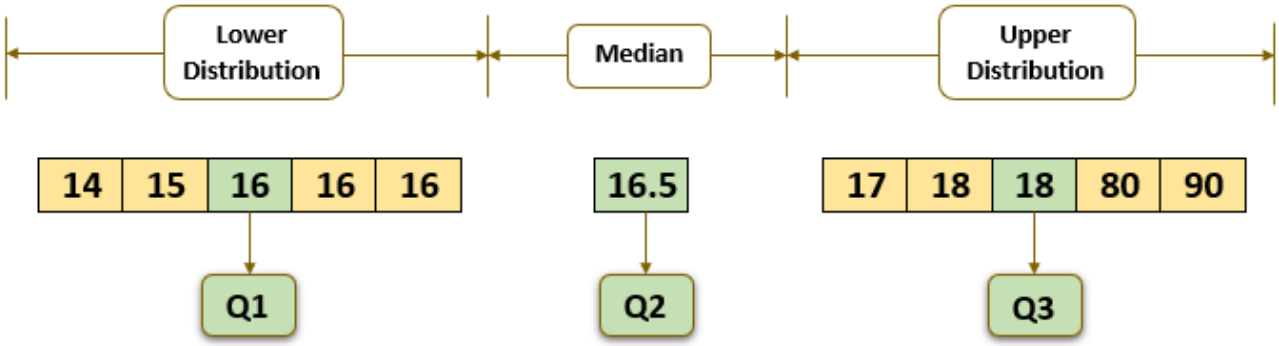
- **Q3** is the median of the upper half of the distribution (16,16,16,15,14). The number of

observations in the upper half also is 5. As such Q3 will be the value at 3rd position in the upper half of the data.

- **Therefore Q3 = 18**

Interquartile Range (IQR)

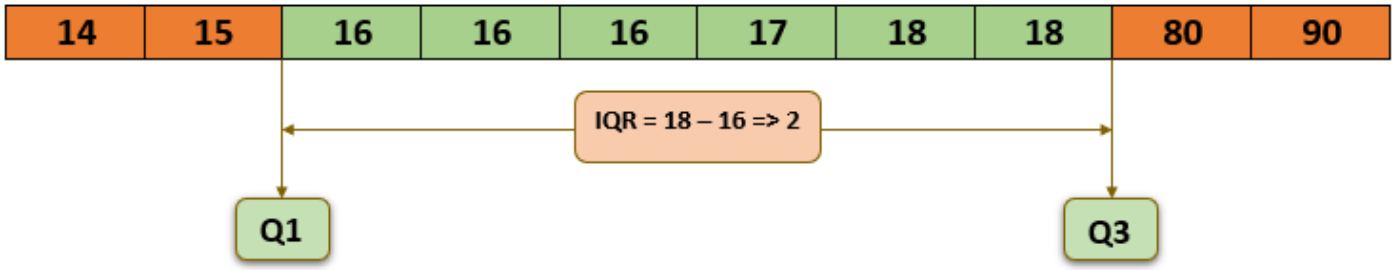
- The three quartiles that divide the data distribution into four equal parts are:
- **Q1 = 16; Q2 = 16.5; Q3 = 18;**



- **IQR = Q3 – Q1 = 18 – 16**
- **IQR = 2**

Outliers do not impact Interquartile Range statistic

- **Interquartile Range (IQR)** is the range of the middle 50% of the values in the data distribution. IQR is also called the **midspread** or the **middle 50%**. It only considers values between Q3 and Q1. It omits the top and the bottom 25% of values in the sorted data distribution.



- IQR omits the extreme values (80,90), hence it is not impacted by the outliers. When the data have outliers, IQR is the best measure of dispersion.

Outlier Detection

One of the commonly used formula to find outliers on the lower and higher side is:

- If there is any value below (Q1 – 1.5 * IQR) then it is considered to be an outlier. In statistics, (Q1 – 1.5 * IQR) is called Lower Control Limit (LCL).
- Likewise, any value above (Q3 + 1.5 * IQR) is also considered as an outlier. The computation of (Q3 + 1.5 * IQR) is called Upper Control Limit (UCL)

Empirical rule & Chebyshev rule

Tuesday, December 7, 2021 11:45 AM

The Empirical Rule
Approximately 68%68% of the data lie within one
standard deviation
of the mean, that is, in the interval with endpoints $\bar{x} \pm s \bar{x} \pm s$ for samples and with endpoints $\mu \pm \sigma \mu \pm \sigma$ for populations; if a data set has an approximately bell-shaped
relative frequency histogram
, then (Figure 2.5.22.5.2)

- approximately 95%95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2 s \bar{x} \pm 2 s$ for samples and with endpoints $\mu \pm 2 \sigma \mu \pm 2 \sigma$ for populations; and
- approximately 99.7%99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3 s \bar{x} \pm 3 s$ for samples and with endpoints $\mu \pm 3 \sigma \mu \pm 3 \sigma$ for populations.

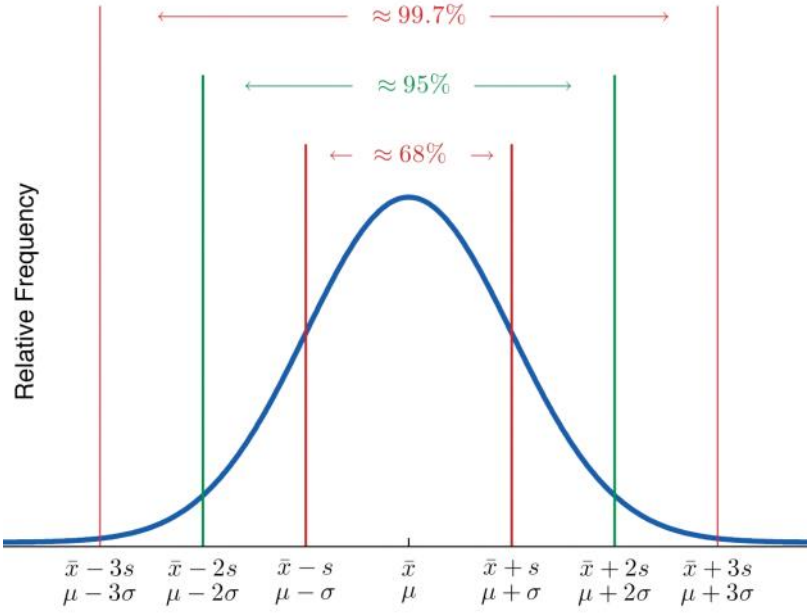


Figure 2.5.22.5.2: The Empirical Rule
Two key points in regard to the Empirical Rule are that the data distribution must be approximately bell-shaped and that the percentages are only approximately true. The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any ofthe intervals specified by the rule could be either greater or less than those given in the rule.

From <https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/02%3A_Descriptive_Statistics/2.05%3A_The_Empirical_Rule_and_Chebyshev's_Theorem>

The Empirical Rule does not apply to all data sets, only to those that are bell-shaped, and even then is stated in terms of approximations. A result that applies to every data set is known as Chebyshev's Theorem.
Chebyshev's Theorem
For any numerical data set,

- at least 3/43/4 of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2 s \bar{x} \pm 2 s$ for samples and with endpoints $\mu \pm 2 \sigma \mu \pm 2 \sigma$ for populations;
- at least 8/98/9 of the data lie within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3 s \bar{x} \pm 3 s$ for samples and with endpoints $\mu \pm 3 \sigma \mu \pm 3 \sigma$ for populations;
- at least $1 - 1/k2 1 - 1/k2$ of the data lie within kk standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm k s \bar{x} \pm k s$ for samples and with endpoints $\mu \pm k \sigma \mu \pm k \sigma$ for populations, where kk is any positive whole number that is greater than 11.

Figure 2.5.42.5.4 gives a visual illustration of Chebyshev's Theorem.

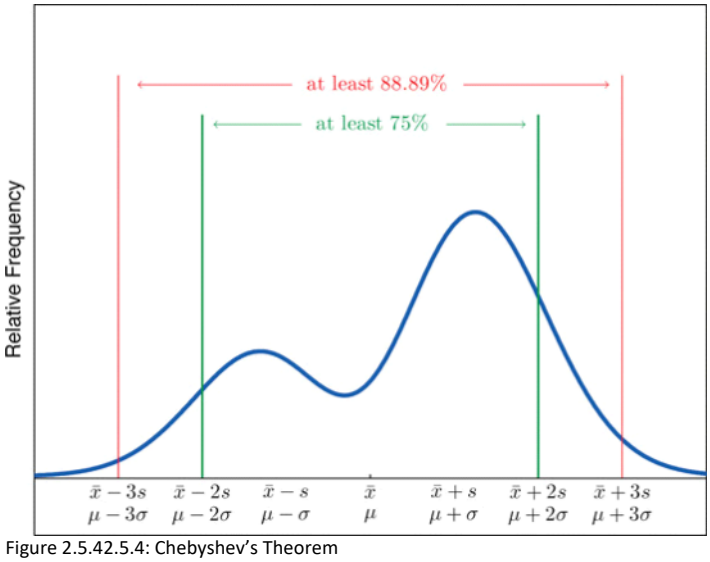


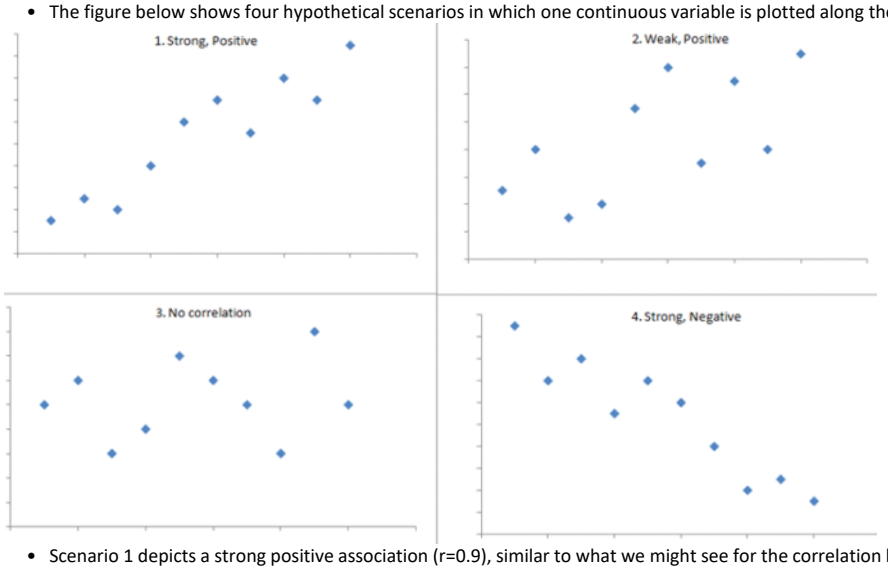
Figure 2.5.42.5.4: Chebyshev's Theorem
It is important to pay careful attention to the words "at least" at the beginning of each of the three parts of Chebyshev's Theorem. The theorem gives the *minimum* proportion of the data which must lie within a given number of standard deviations of the mean; the true proportions found within the indicated regions could be greater than what the theorem guarantees.

From <https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/02%3A_Descriptive_Statistics/2.05%3A_The_Empirical_Rule_and_Chebyshev's_Theorem>

Correlation analysis

Tuesday, December 7, 2021 11:49 AM

- In correlation analysis, we estimate a **sample correlation coefficient**, more specifically the **Pearson Product Moment correlation coefficient**.
- The sample correlation coefficient, denoted r, ranges between -1 and +1 and quantifies the direction and strength of the linear association between the two variables.
- The correlation between two variables can be positive (i.e., higher levels of one variable are associated with higher levels of the other) or negative (i.e., higher levels of one variable are associated with lower levels of the other).
- The sign of the correlation coefficient indicates the direction of the association.
- The magnitude of the correlation coefficient indicates the strength of the association.
- For example, a correlation of r = 0.9 suggests a strong, positive association between two variables, whereas a correlation of r = -0.2 suggest a weak, negative association.
- A correlation close to zero suggests no linear association between two continuous variables.
- It is important to note that there may be a non-linear association between two continuous variables, but computation of a correlation coefficient does not detect this.
- Therefore, it is always important to evaluate the data carefully before computing a correlation coefficient.
- Graphical displays are particularly useful to explore associations between variables.
- The figure below shows four hypothetical scenarios in which one continuous variable is plotted along the X-axis and the other along the Y-axis.



- Scenario 1 depicts a strong positive association (r=0.9), similar to what we might see for the correlation between infant bir th weight and birth length.
- Scenario 2 depicts a weaker association (r=0,2) that we might expect to see between age and body mass index (which tends to i ncrease with age).
- Scenario 3 might depict the lack of association (r approximately = 0) between the extent of media exposure in adolescence and age at which adolescents initiate sexual activity.
- Scenario 4 might depict the strong negative association (r= -0.9) generally observed between the number of hours of aerobic exercise per week and percent body fat.

From <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Correlation-Regression/BS704_Correlation-Regression2.html>

Five number summary Boxplot & other plots

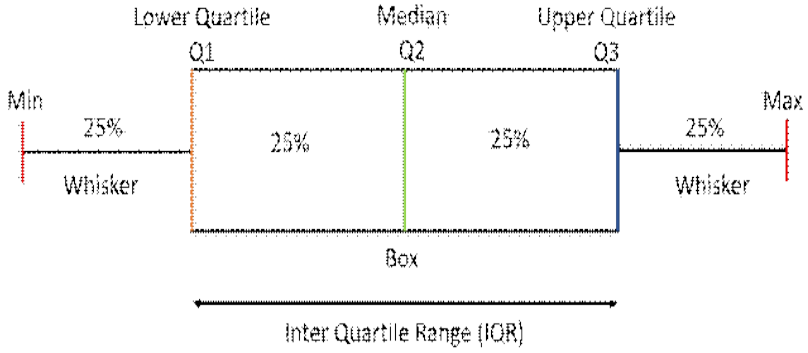
Tuesday, December 7, 2021 11:55 AM

Box Plot: It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

1) Understanding the components of a box plot

A box plot gives a five-number summary of a set of data which is-

- **Minimum** – It is the minimum value in the dataset excluding the outliers
- **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
- **Median (Q2)** – It is the mid-point of the dataset. Half of the values lie below it and half above.
- **Third Quartile (Q3)** – 75% of the data lies below the Third (Upper) Quartile.
- **Maximum** – It is the maximum value in the dataset excluding the outliers.



Note: The box plot shown in the above diagram is a perfect plot with no skewness. The plots can have skewness and the median might not be at the center of the box.

The area inside the box (50% of the data) is known as the **Inter Quartile Range.** The **IQR** is calculated as –

```
IQR = Q3-Q1
```

Outliers are the data points **below and above** the **lower and upper limit**. The lower and upper limit is calculated as –

```
Lower Limit = Q1 - 1.5*IQR
Upper Limit = Q3 + 1.5*IQR
```

The values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit.

Probability & distribution

Tuesday, December 7, 2021 1:01 PM

What is the Probability Distribution?

Probability distribution could be defined as the table or equations showing respective probabilities of different possible outcomes of a defined event or scenario. In simple words, its calculation shows the possible outcome of an event with the relative possibility of occurrence or non-occurrence as required.

Probability Distribution Formula

The probability of occurring event can be calculated by using the below formula;

Probability of Event = No of Possibility of Event / No of Total Possibility

Example #1

Let's suppose a coin was tossed twice, and we have to show the probability distribution of showing heads.

Solution

In the given an example, possible outcomes could be (H, H), (H, T), (T, H), (T, T)

Then possible no. of heads selected will be – 0 or 1 or 2, and the probability of such event could be calculated by using the following formula:

Probability of selecting 0 Head = No of Possibility of Event / No of Total Possibility

= 1/4

Probability of selecting 1 Head = No of Possibility of Event / No of Total Possibility

= 2/4

= 1/2

Probability of selecting 2 heads =No of Possibility of Event / No of Total Possibility

= 1/4

Explanation: In the given an example, the event was 'No. of heads'. And the number of heads that can occur is either 0 or 1 or 2, which would be termed as possible outcomes, and the respective possibility could be 0.25, 0.5, 0.25 of the possible outcomes.

Example #2

In an interview hall, there were 4 people present consisting of 2 men and 2 women after

being tested by the interviewers. But the concerned company had only 2 vacancies to fill. So the interviewer decided to select 2 candidates from the people present in the hall. What will be the probability distribution of 'selecting at least one woman.'

Solution

In the given case, the number of possibilities of selecting candidate could be,

(W1, W2), (W1, M1), (W1, M2), (W2, M1), (W2, M2), (M1, M2)

As per the requirement, let's denote the event 'number of women' as X, then the possible values of X could be;

X = 1 or 2

Calculation of Probability of an event

So, the probability of selecting 0 women = no of the possibility of selecting 1 women / total possibilities

=1/6
Similarly,

Probability of selecting X women = no of the possibility of selecting X women / total possibilities

So, the probability of selecting 1 woman = no of the possibility of selecting 1 women / total possibilities
=4/6
=2/3
Similarly,

Probability of selecting 2 women =no of the possibility of selecting 2 women / total possibilities
=1/6

Now, as per the question, the probability of selecting at least 1 woman will be

= Probability of selecting 1 woman + Probability of selecting 2 women
= 2/3+1/6

=5/6
So, the probability distribution for selecting women will be shown as;

Explanation: In this scenario, the management decided to fill up the 2 vacancies through interviews, and during the interview, they chose 4 people. For the final selection, they decide to select randomly, and the number of women selected could be either 0 or 1 or 2. Possibility of an event where no women would be selected is & the possibility of an event

where only 1 woman will be selected amounted to, whereas the possibility of selection of both women is.

So, through the use of Probability distribution, the trend of employment, trend of hiring, selection of candidates, and other nature could be summarised and studied upon.

Example #3

In a similar type of situation, let's assume a situation where a manufacturing company named ABC Inc. was engaged in the manufacturing of tube lights. One day the Operation Manager decided to randomly evaluate the effectiveness of production by evaluating the percentage of Damaged stocks produced within 1 hour. Let's say, within 1 hour, 10 tube lights were produced, out of which 2 were damaged. The manager decided to pick 3 of the tube lights randomly. Prepare the probability distribution of selecting damaged goods.

Solution

In the given example, the random variable is the 'number of damaged tube lights selected.' Let's denote the event as 'X.'

Then, the possible values of X are (0,1,2)

So, the probability could be calculated by using the formula;

Probability of selecting X = no of possibilities of selecting X / total possibilities

Then,

Probability of selecting 0 damaged lights = probability of selecting good light in 1st round X probability of selecting good light in 2nd round X probability of selecting good light in 3rd round

P (0) = P(G) X P(G) X P(G)
= 8/10*7/9*6/8
= 7/15
Similarly, Probability of selecting only 1 damage light = [P(G) X P(G) X P(D)] X 3

(multiplied by 3 because the damaged light can be selected in 3 ways, i.e., either in 1st round or 2nd or 3rd round)

So,

P (1) = (8/10*7/9*2/8)*3
=7/15
Similarly, Probability of selecting 2 damage lights = [P(G) X P(D) X P(D)] X 3

(multiplied by 3 because the good light can be selected in 3 ways, i.e., either in 1st round or 2nd or 3rd round)

So,

P (2) = (8/10*2/9*1/8)*3
= **1/15**
So the probability of selecting at least 1 Damaged lights = Probability of selecting 1 Damage + Probability of selecting 2 Damage

= P (1) + P (2)
=7/15+1/15
=8/15
So, the probability distribution for selecting damage lights could be shown as;

Explanation: The operation Manager of the Business organization wanted to evaluate the effectiveness of the process through the random selection of goods and evaluating the chances of production of damaged goods.

Through this example, we can see that the industry can also use Probability distribution for evaluating the effectiveness of its processes and the ongoing trends.

Relevance and Uses

A probability distribution is basically used for recording the possibility of occurrence or non-occurrence of a certain event. From a business point of view, it can also be used for predicting or estimating the possible future returns or profitability of the business. In modern-day business, the probability distribution calculation is used for sales forecasting, risk evaluation, finding and evaluating the obsolete part of any business or process, etc.

From <https://www.wallstreetmojo.com/probability-distribution/>

Bayes Theorem

Tuesday, December 14, 2021 5:36 PM

Bayes' Theorem is a way of finding a probability when we know certain other probabilities.

The formula is:

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Which tells us: how often A happens *given that B happens*, written **P(A|B)**,
When we know: how often B happens *given that A happens*, written **P(B|A)**
and how likely A is on its own, written **P(A)**
and how likely B is on its own, written **P(B)**

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

P(Fire|Smoke) means how often there is fire when we can see smoke
P(Smoke|Fire) means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know "backwards" P(Smoke|Fire)

Example:

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$P(Fire|Smoke) = \frac{P(Fire)\,P(Smoke|Fire)}{P(Smoke)}$$

$$= \frac{1\% \times 90\%}{10\%}$$

$$= 9\%$$

So it is still worth checking out any smoke to be sure.

Example: Picnic Day

You are planning a picnic today, but the morning is cloudy

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.

The chance of Rain given Cloud is written P(Rain|Cloud)

So let's put that in the formula:

$$P(\text{Rain}|\text{Cloud}) = \frac{P(\text{Rain})\ P(\text{Cloud}|\text{Rain})}{P(\text{Cloud})}$$

- P(Rain) is Probability of Rain = 10%
- P(Cloud|Rain) is Probability of Cloud, given that Rain happens = 50%
- P(Cloud) is Probability of Cloud = 40%

$$P(\text{Rain}|\text{Cloud}) = \frac{0.1 \times 0.5}{0.4} = .125$$

Or a 12.5% chance of rain. Not too bad, let's have a picnic!

Just 4 Numbers

Imagine 100 people at a party, and you tally how many wear pink or not, and if a man or not, and get these numbers:

	Pink	notPink
Man	5	35
notMan	20	40

Bayes' Theorem is based off just those 4 numbers!

Let us do some totals:

	Pink	notPink	
Man	5	35	40
notMan	20	40	60
	25	75	100

And calculate some probabilities:

- the probability of being a man is P(Man) = $\frac{40}{100}$ = 0.4
- the probability of wearing pink is P(Pink) = $\frac{25}{100}$ = 0.25
- the probability that a man wears pink is P(Pink|Man) = $\frac{5}{40}$ = 0.125
- the probability that a person wearing pink is a man **P(Man|Pink) = ...**

And then the puppy arrives! Such a cute puppy.

But all your data is **ripped up**! Only 3 values survive:

- **P(Man) = 0.4,**
- **P(Pink) = 0.25 and**
- **P(Pink|Man) = 0.125**

Can you discover **P(Man|Pink)** ?

Imagine a pink-wearing guest leaves money behind ... was it a man? We can answer this question using Bayes' Theorem:

$$P(\text{Man}|\text{Pink}) = \frac{P(\text{Man})\ P(\text{Pink}|\text{Man})}{P(\text{Pink})}$$

$$P(\text{Man}|\text{Pink}) = \frac{0.4 \times 0.125}{0.25} = 0.2$$

Note: if we still had the raw data we could calculate directly $\frac{5}{25} = 0.2$

Being General

Why does it work?

Let us replace the numbers with letters:

	B	notB	
A	s	t	s+t
notA	u	v	u+v
	s+u	t+v	s+t+u+v

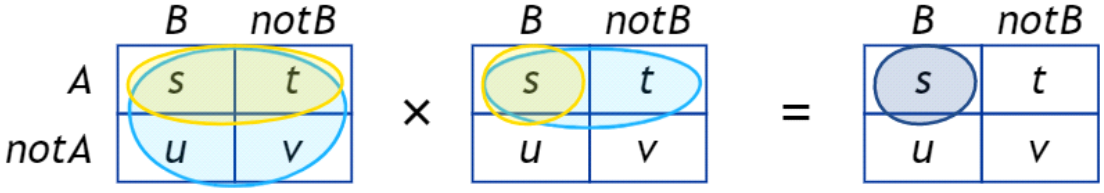
Now let us look at **probabilities**. So we take some ratios:

- the overall probability of "A" is P(A) = $\frac{s+t}{s+t+u+v}$
- the probability of "B given A" is P(B|A) = $\frac{s}{s+t}$

And then multiply them together like this:

$$P(A) \times P(B|A) = P(A)\,P(B|A)$$

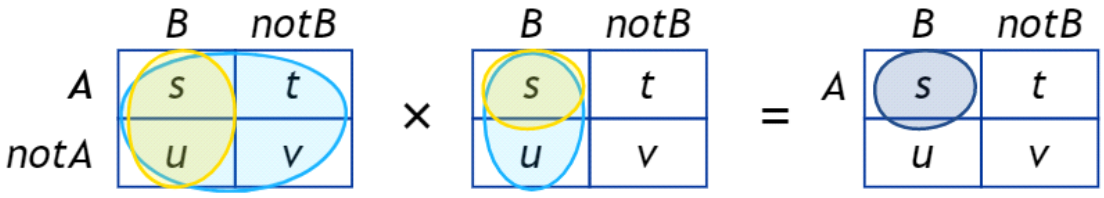
$$\frac{s+t}{s+t+u+v} \times \frac{s}{s+t} = \frac{s}{s+t+u+v}$$



Now let us do that again but use **P(B)** and **P(A|B)**:

$$P(B) \times P(A|B) = P(B)\,P(A|B)$$

$$\frac{s+u}{s+t+u+v} \times \frac{s}{s+u} = \frac{s}{s+t+u+v}$$



Both ways get the **same result** of $\frac{s}{s+t+u+v}$

So we can see that:

$$P(B)\,P(A|B) = P(A)\,P(B|A)$$

Nice and symmetrical isn't it?

It actually *has* to be symmetrical as we can swap rows and columns and get the same top-left corner.

And it is also **Bayes Formula** ... just divide both sides by P(B):

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Remembering

First think "AB AB AB" then remember to group it like: "AB = A BA / B"

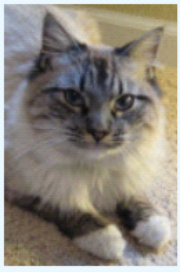
$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Cat Allergy?

One of the famous uses for Bayes Theorem is False Positives and False Negatives.

For those we have two possible cases for "A", such as **Pass**/**Fail** (or Yes/No etc)

Example: Allergy or Not?



Hunter says she is itchy. There is a test for Allergy to Cats, but this test is not always right:

- For people that **really do** have the allergy, the test says "Yes" **80%** of the time
- For people that **do not** have the allergy, the test says "Yes" **10%** of the time ("false positive")

If 1% of the population have the allergy, and **Hunter's test says "Yes"**, what are the chances that Hunter really has the allergy?

We want to know the chance of having the allergy when test says "Yes", written **P(Allergy|Yes)**

Let's get our formula:

$$P(\text{Allergy}|\text{Yes}) = \frac{P(\text{Allergy})\ P(\text{Yes}|\text{Allergy})}{P(\text{Yes})}$$

- P(Allergy) is Probability of Allergy = 1%
- P(Yes|Allergy) is Probability of test saying "Yes" for people with allergy = 80%
- P(Yes) is Probability of test saying "Yes" (to anyone) = ??%

Oh no! We **don't know** what the **general** chance of the test saying "Yes" is ...

... but we can calculate it by adding up those **with**, and those **without** the allergy:

- 1% have the allergy, and the test says "Yes" to 80% of them
- 99% do **not** have the allergy and the test says "Yes" to 10% of them

Let's add that up:

$$P(\text{Yes}) = 1\% \times 80\% + 99\% \times 10\% = 10.7\%$$

Which means that about 10.7% of the population will get a "Yes" result.

So now we can complete our formula:

$$P(\text{Allergy}|\text{Yes}) = \frac{1\% \times 80\%}{10.7\%} = 7.48\%$$

$$P(\text{Allergy}|\text{Yes}) = \text{about } \mathbf{7\%}$$

This is the same result we got on False Positives and False Negatives.

In fact we can write a special version of the Bayes' formula just for things like this:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)}$$

"A" With Three (or more) Cases

We just saw "A" with two cases (A and not A), which we took care of in the bottom line.

When "A" has 3 or more cases we include them all in the bottom line:

$$P(A1|B) = \frac{P(A1)P(B|A1)}{P(A1)P(B|A1) + P(A2)P(B|A2) + P(A3)P(B|A3) + \ldots\text{etc}}$$

Example: The Art Competition has entries from three painters: Pam, Pia and Pablo

- Pam put in 15 paintings, 4% of her works have won First Prize.
- Pia put in 5 paintings, 6% of her works have won First Prize.
- Pablo put in 10 paintings, 3% of his works have won First Prize.

What is the chance that Pam will win First Prize?

$$P(\text{Pam}|\text{First}) = \frac{P(\text{Pam})P(\text{First}|\text{Pam})}{P(\text{Pam})P(\text{First}|\text{Pam}) + P(\text{Pia})P(\text{First}|\text{Pia}) + P(\text{Pablo})P(\text{First}|\text{Pablo})}$$

Put in the values:

$$P(\text{Pam}|\text{First}) = \frac{(15/30) \times 4\%}{(15/30) \times 4\% + (5/30) \times 6\% + (10/30) \times 3\%}$$

Multiply all by 30 (makes calculation easier):

$$P(\text{Pam}|\text{First}) = \frac{15 \times 4\%}{15 \times 4\% + 5 \times 6\% + 10 \times 3\%}$$

$$= \frac{0.6}{0.6 + 0.3 + 0.3}$$

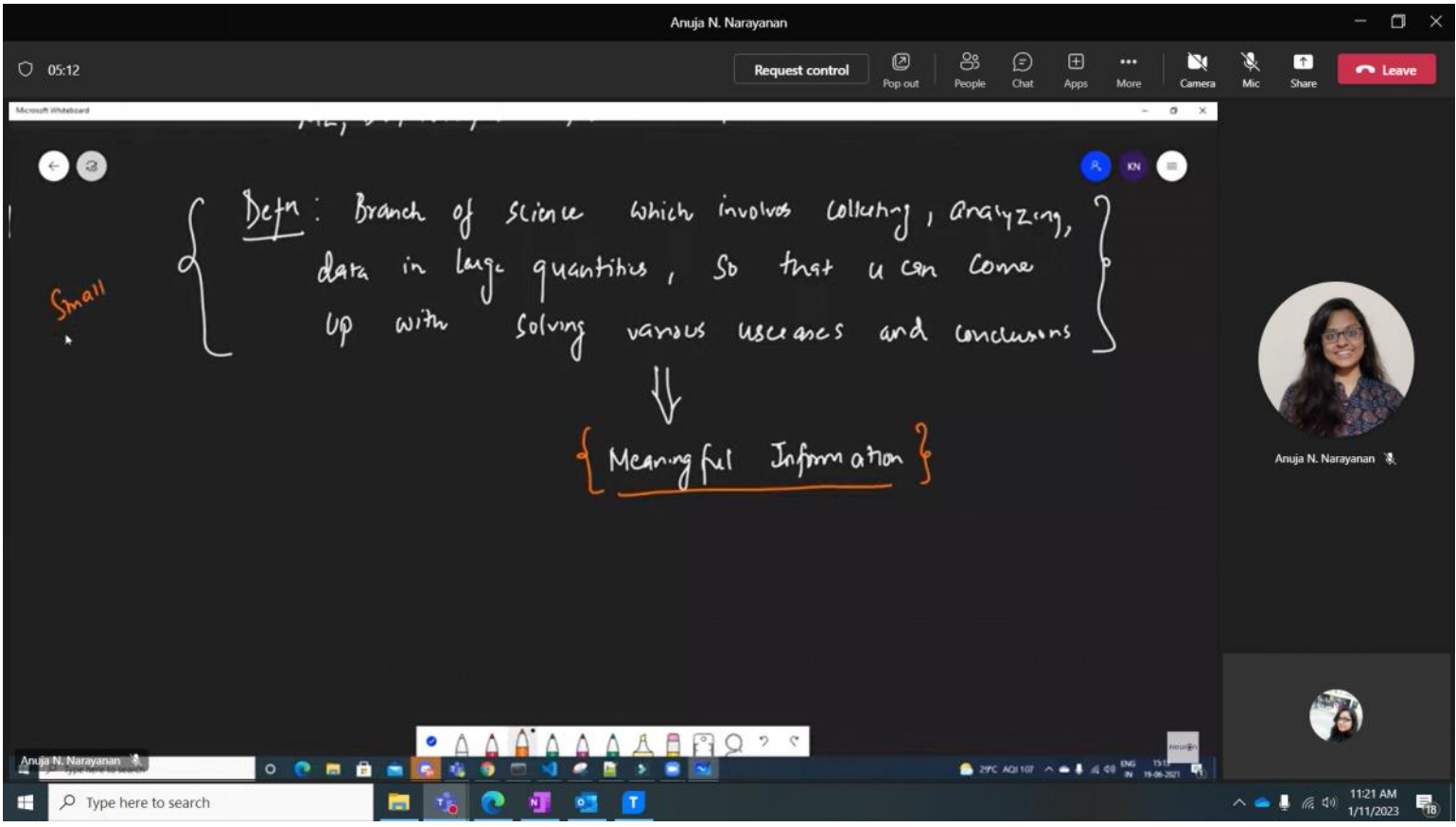
$$= 50\%$$

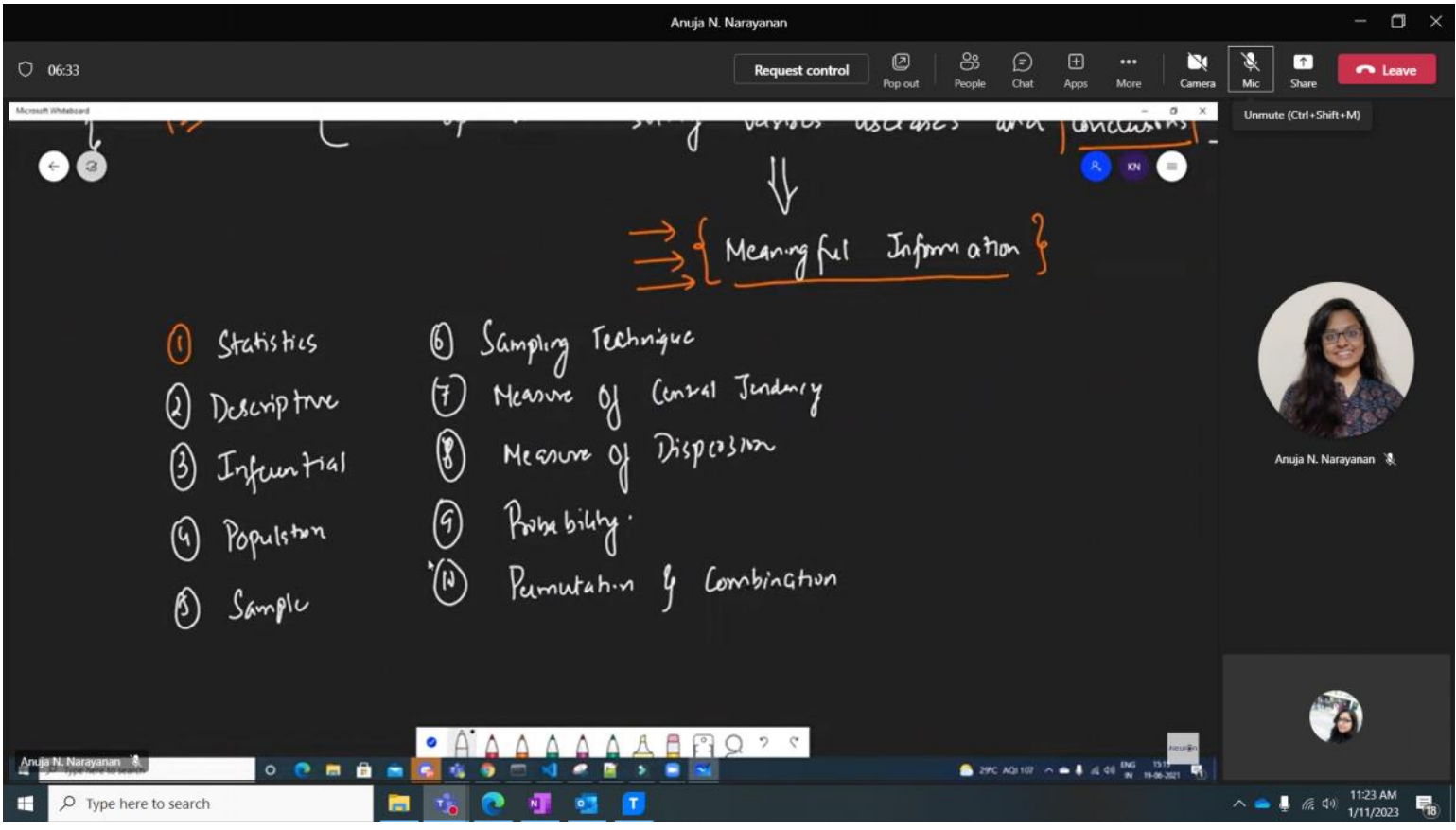
A good chance!

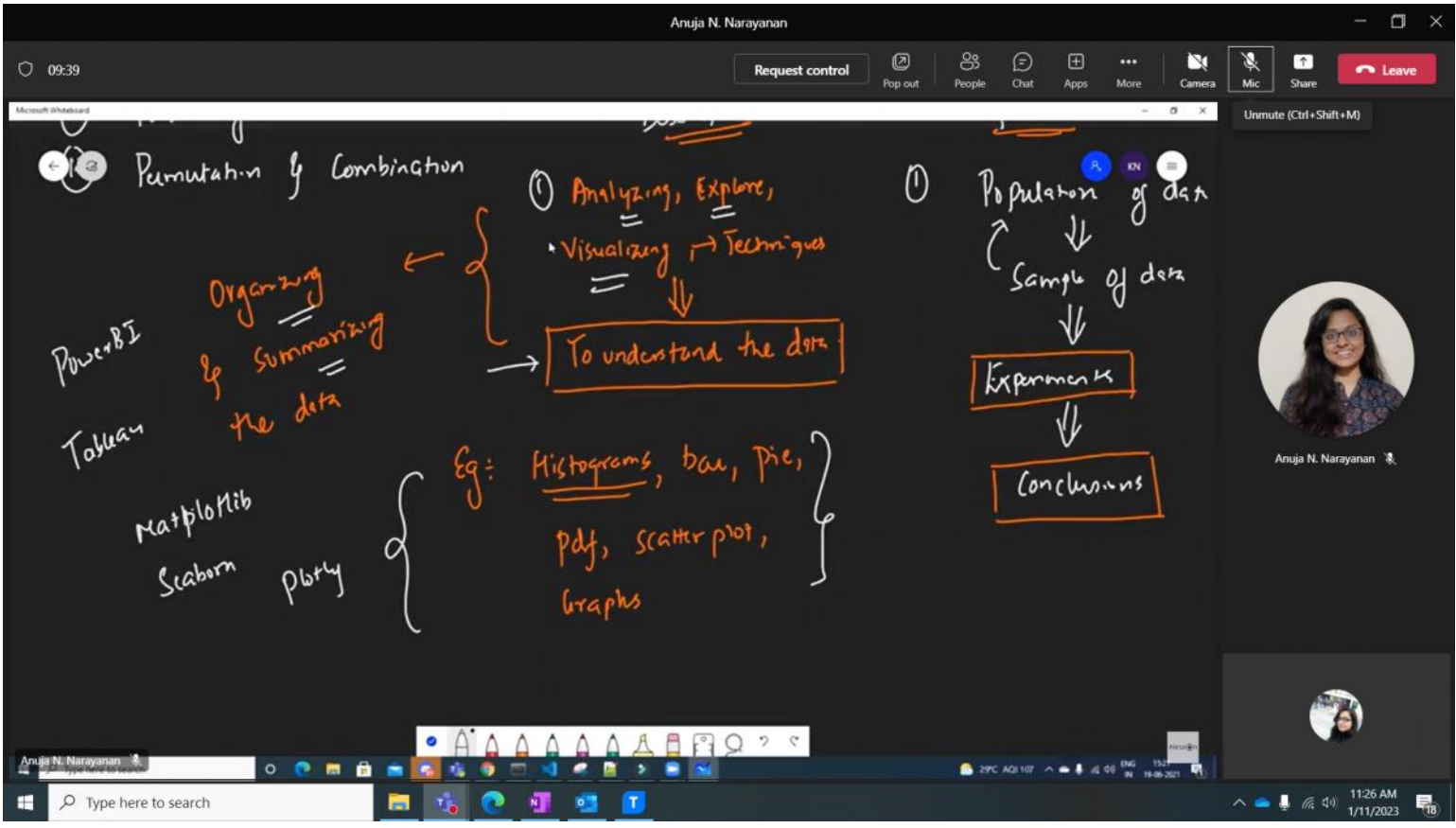
Pam isn't the most successful artist, but she did put in lots of entries.

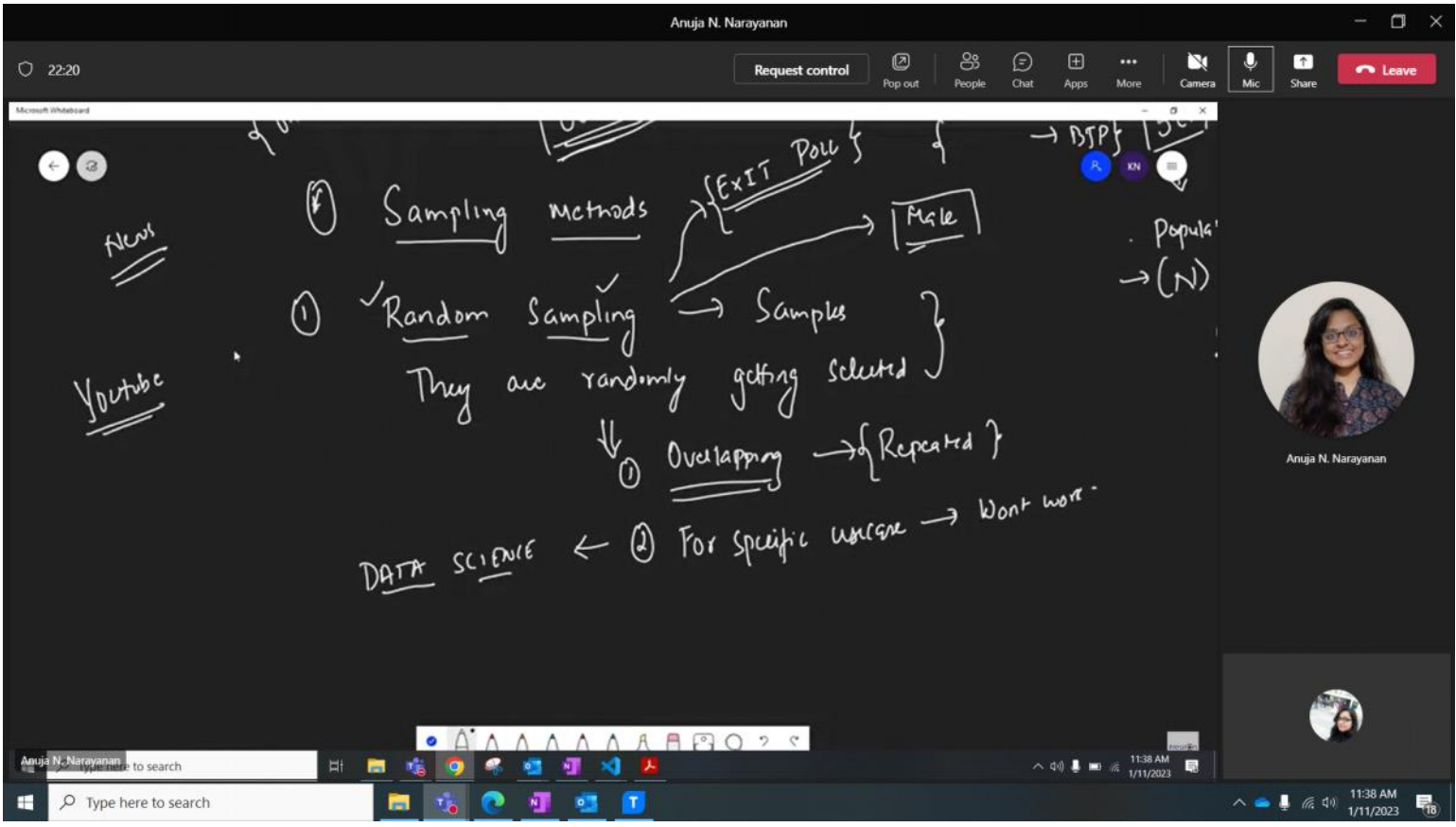
KN(lec-1)

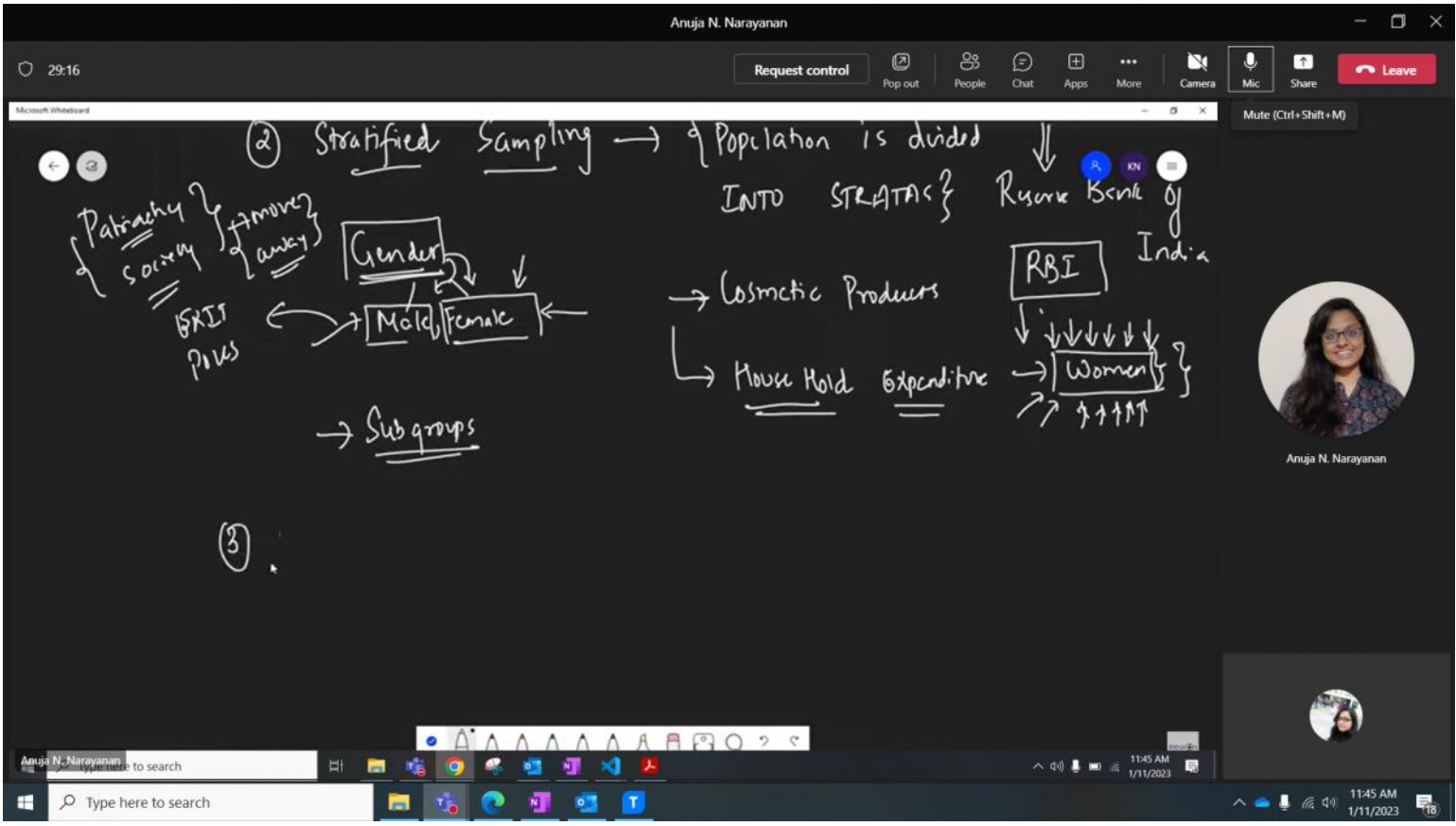
Wednesday, December 15, 2021 3:40 PM

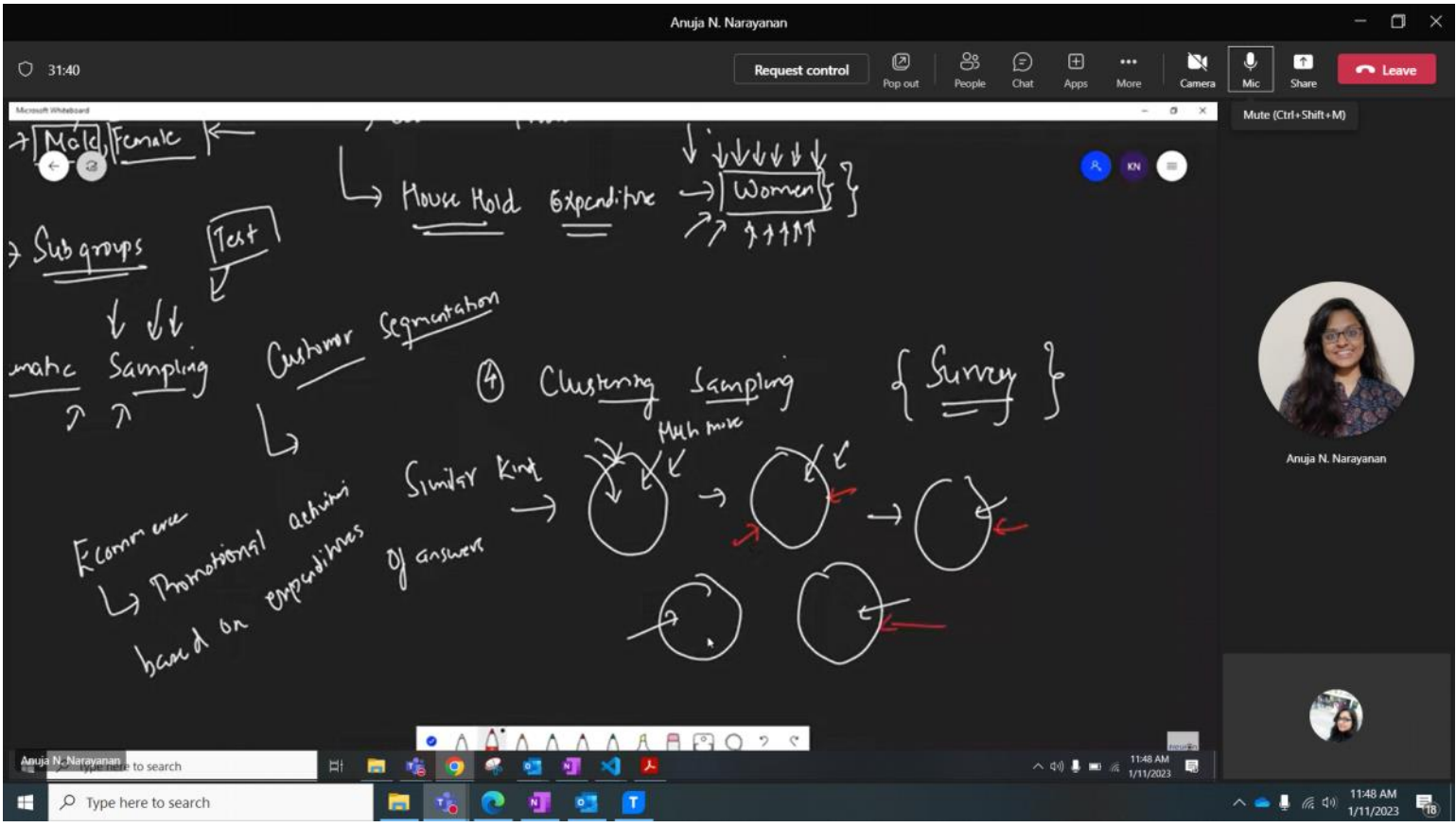


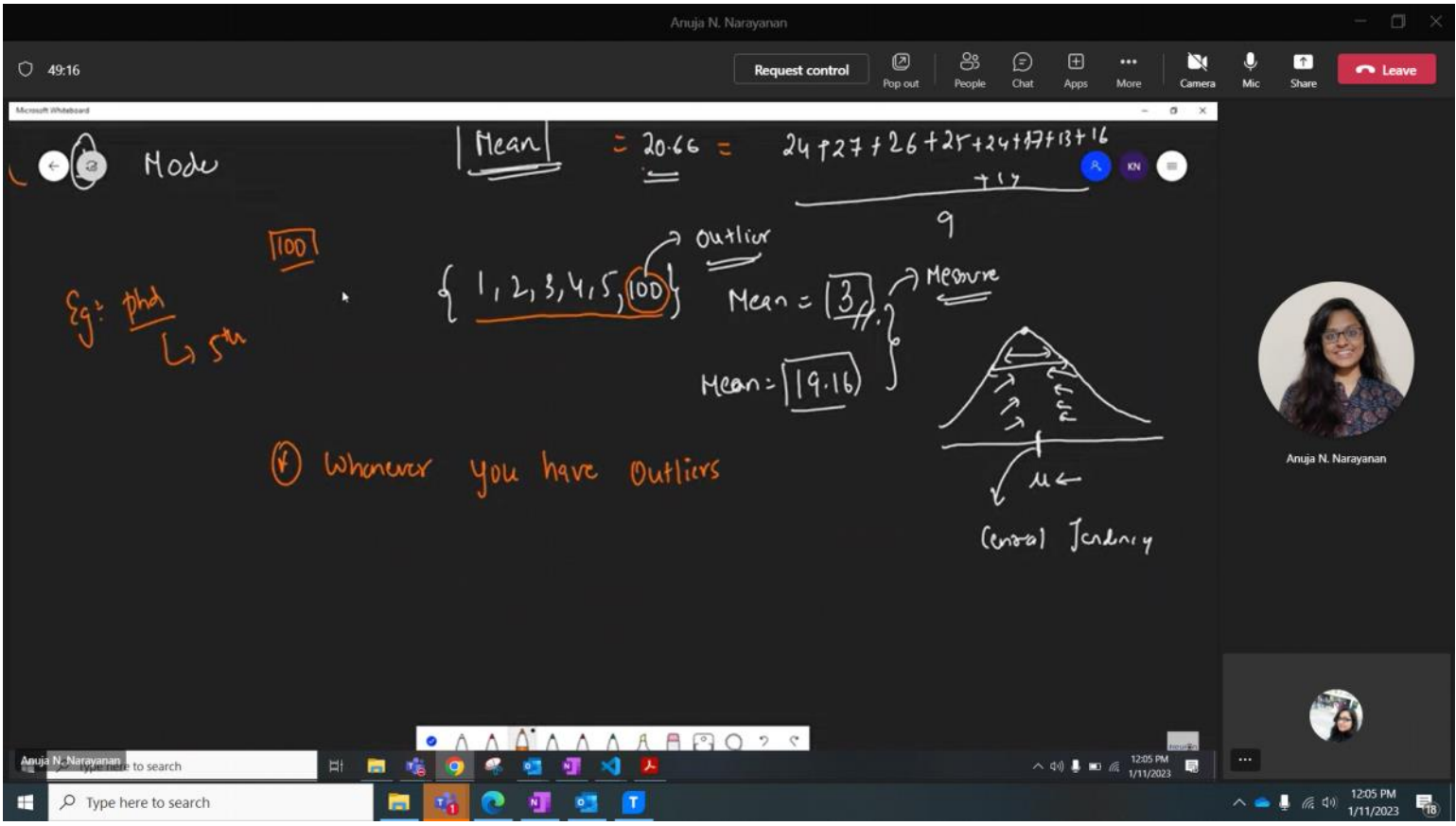


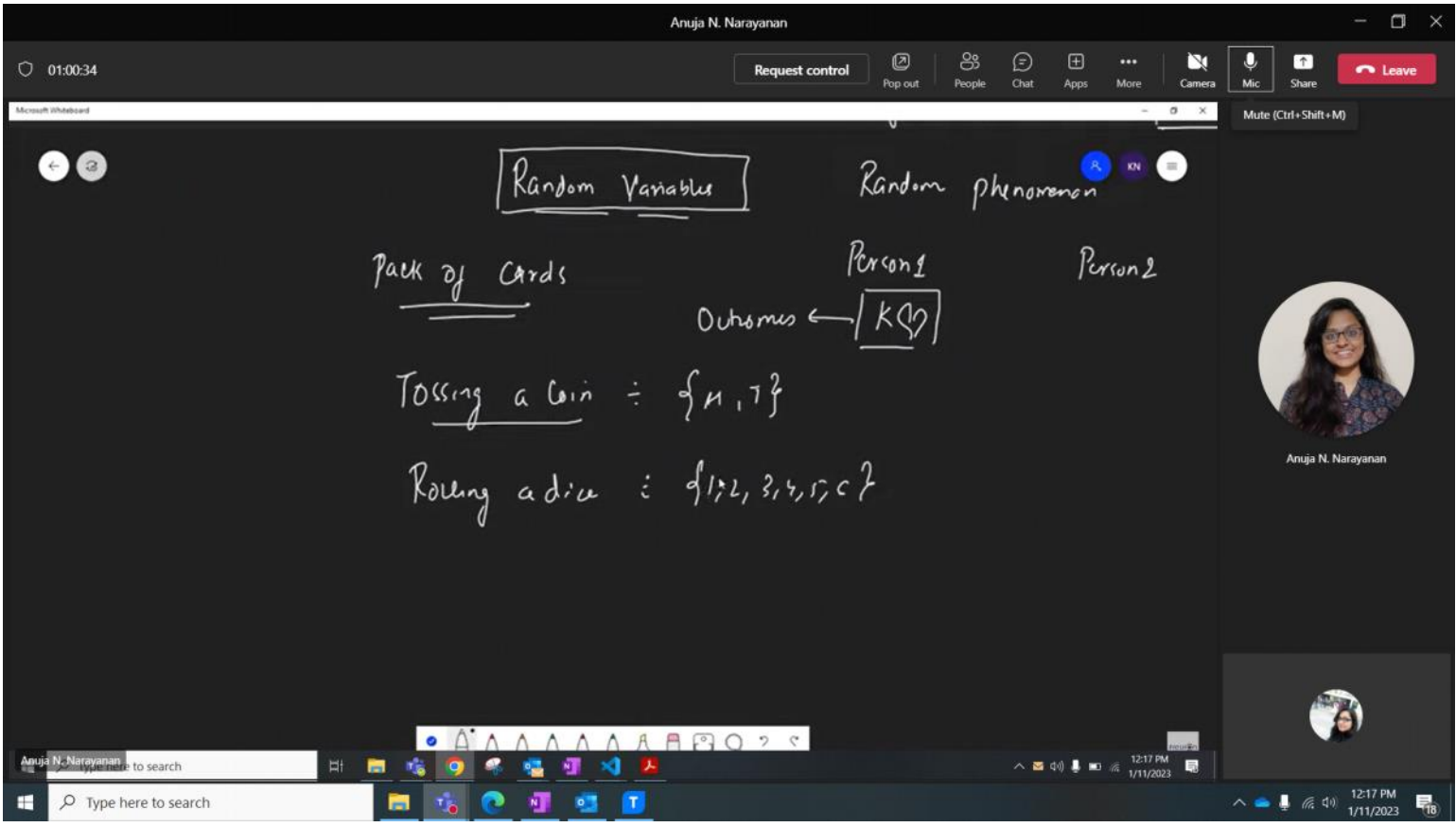

Anuja N. Narayanan
09:39
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Unmute (Ctrl+Shift+M)
Permutation & Combination
① Analyzing, Explore,
Visualizing → Techniques
To understand the data
Organizing & Summarizing the data
PowerBI
Tableau
MatplotLib
Seaborn
plotly
Eg: Histograms, bar, pie, pdf, scatter plot, Graphs
① Population of data
Sample of data
Experiments
Conclusions
Anuja N. Narayanan
Type here to search
11:26 AM
1/11/2023

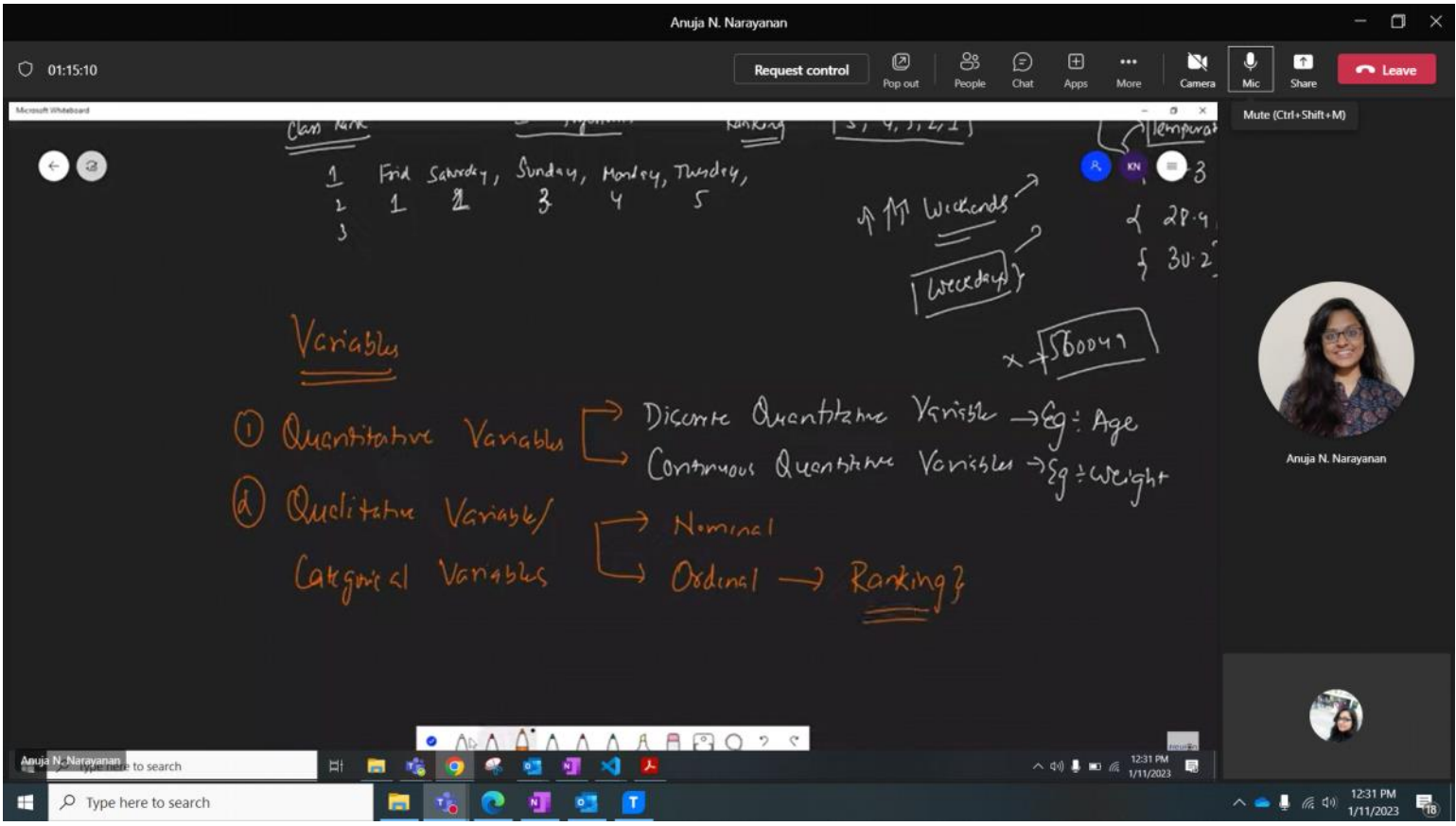

Anuja N. Narayanan
22:20
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
News
Youtube
① Sampling Methods
Exit Poll
Male
① Random Sampling → Samples
They are randomly getting selected
① Overlapping → Repeated
DATA SCIENCE ← ② For specific usecase → Wont work
→ DSTP
→ (N)
Anuja N. Narayanan
Type here to search
11:38 AM
1/11/2023

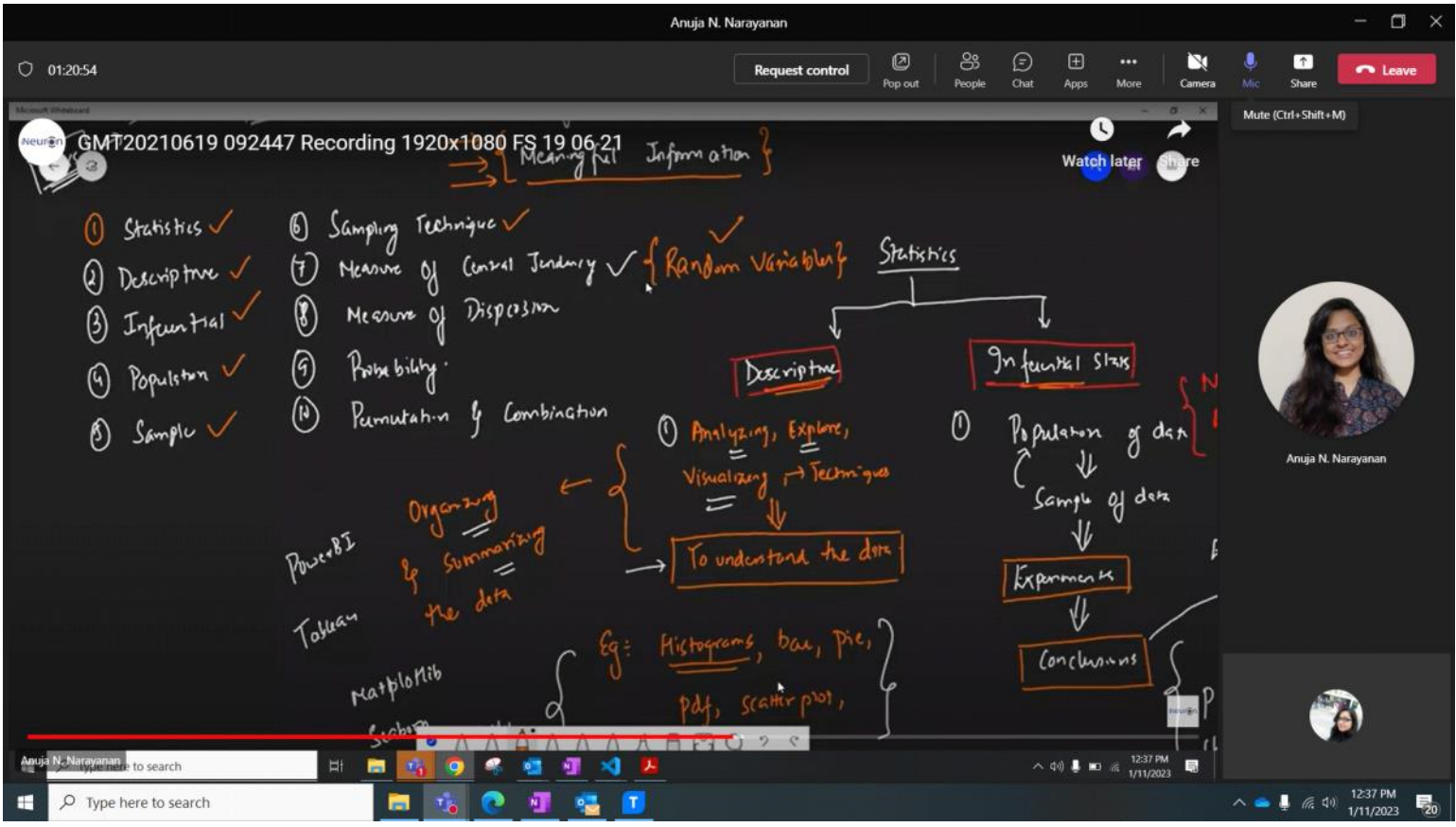

Anuja N. Narayanan
29:16
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Mute (Ctrl+Shift+M)
② Stratified Sampling → Population is divided INTO STRATAS
Reserve Bank of India
RBI
Patriarchy society
more risky
Gender
Male
Female
EXIT Polls
→ Cosmetic Products
→ House Hold Expenditure
Women
→ Sub groups
③
Anuja N. Narayanan
Type here to search
11:45 AM
1/11/2023

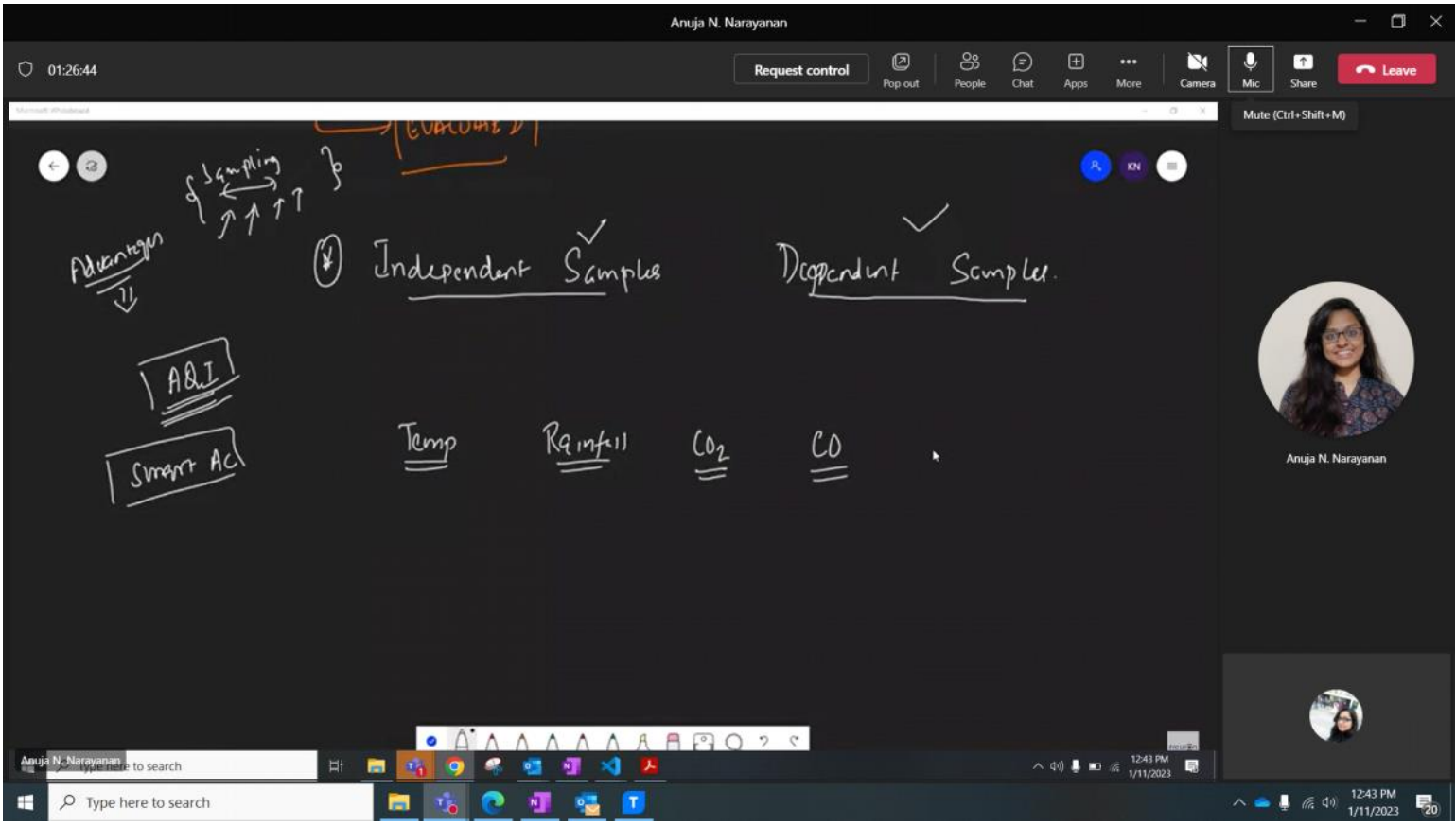

Anuja N. Narayanan
31:40
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Mute (Ctrl+Shift+M)
Male
Female
→ House Hold Expenditure
Women
→ Sub groups
Test
matic Sampling
Customer Segmentation
④ Clustering Sampling
{ Survey }
Much more
Similar Kind of answer
E commerce
→ Promotional activities based on expenditures
Anuja N. Narayanan
Type here to search
11:48 AM
1/11/2023


Anuja N. Narayanan
49:16
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Mode
Mean = 20.66 = 24+27+26+25+24+17+13+16 / 9
Outlier
100
Eg: phd
5th
{1,2,3,4,5,100}
Mean = 3
Mean = 19.16
Measure
Whenever you have Outliers
Central Tendency
Anuja N. Narayanan
Type here to search
12:05 PM
1/11/2023


Anuja N. Narayanan
01:00:34
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Mute (Ctrl+Shift+M)
Random Variable
Random phenomenon
Pack of Cards
Person 1
Person 2
Outcomes ← K
Tossing a Coin : {H, T}
Rolling a dice : {1, 2, 3, 4, 5, 6}
Anuja N. Narayanan
Type here to search
12:17 PM
1/11/2023


Anuja N. Narayanan
01:15:10
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Mute (Ctrl+Shift+M)
Variables
① Quantitative Variables
Discrete Quantitative Variable → Eg: Age
Continuous Quantitative Variables → Eg: Weight
② Qualitative Variable/ Categorical Variables
Nominal
Ordinal → Ranking
Anuja N. Narayanan
Type here to search
12:31 PM
1/11/2023


Anuja N. Narayanan
01:20:54
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Mute (Ctrl+Shift+M)
GMT20210619 092447 Recording 1920x1080 FS 19 06 21
Watch later
Share
Meaningful Information
① Statistics
② Descriptive
③ Inferential
④ Population
⑤ Sample
⑥ Sampling Technique
⑦ Measure of Central Tendency
⑧ Measure of Dispersion
⑨ Probability
⑩ Permutation & Combination
Random Variables
Statistics
Descriptive
Inferential Stats
① Analyzing, Explore, Visualizing → Techniques
To understand the data
Organizing & Summarizing the data
PowerBI
Tableau
MatplotLib
Eg: Histograms, bar, Pie, pdf, scatter plot,
① Population of data
Sample of data
Experiments
Conclusions
Anuja N. Narayanan
Type here to search
12:37 PM
1/11/2023


Anuja N. Narayanan
01:26:44
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Mute (Ctrl+Shift+M)
Sampling
Advantages
Independent Samples
Dependent Samples
AQI
Smart AC
Temp
Rainfall
CO_2
CO
Anuja N. Narayanan
Type here to search
12:43 PM
1/11/2023

Sampling techniques

Friday, May 20, 2022 9:34 AM

[8 Types of Sampling Techniques. Understanding Sampling Methods (Visuals… | by Prakhar Mishra | Towards Data Science]

1. Random Sampling
2. Stratified Sampling
3. Cluster Sampling
4. Systematic Sampling
5. Multistage Sampling

Probabilistic

6. Convenience Sampling
7. Voluntary Sampling
8. Snowball Sampling

Non Probabilistic

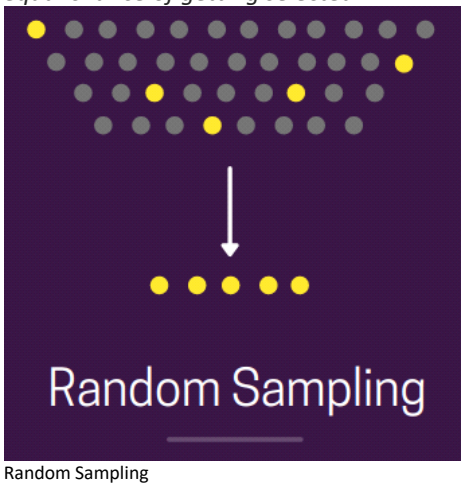
Sampling is the process of selecting a subset*(a predetermined number of observations)* from a larger population. It's a pretty common technique wherein, we run experiments and draw conclusions about the population, without the need of having to study the entire population. In this blog, we will go through two types of sampling methods:

1. ***Probability Sampling*** *—Here we choose a sample based on the theory of probability.*
2. ***Non-Probability Sampling*** *— Here we choose a sample based on non-random criteria, and not every member of the population has a chance of being included.*

From <https://towardsdatascience.com/8-types-of-sampling-techniques-b21adcdd2124>

Random Sampling

Under Random sampling, every element of the population has an equal probability of getting selected. *Below fig. shows the pictorial view of the same — All the points collectively represent the entire population wherein every point has an equal chance of getting selected.*



Random Sampling

You can implement it using python as shown below —

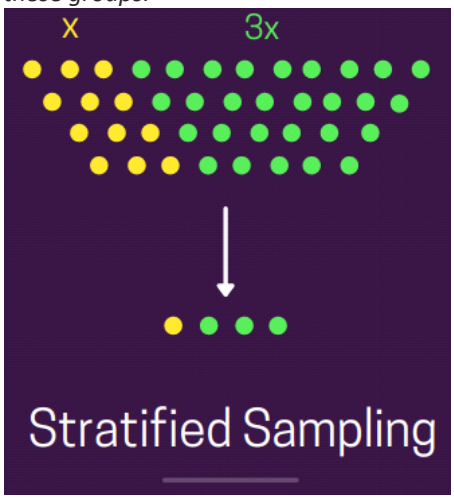
```
import randompopulation = 100

data = range(population)print(random.sample(data,5))
```

> 4, 19, 82, 45, 41

Stratified Sampling

Under stratified sampling, we **group the entire population into subpopulations** by some common property. *For example — Class labels in a typical ML classification task.* We then randomly sample from those groups individually, such that the **groups are still maintained in the same ratio** as they were in the entire population*. Below fig. shows a pictorial view of the same — We have two groups with a count ratio of x and 4x based on the colour, we randomly sample from yellow and green sets separately and represent the final set in the same ratio of*

these groups.



Stratified Sampling

You can implement it very easily using python sklearn lib. as shown below —

```
from sklearn.model_selection import train_test_split
```

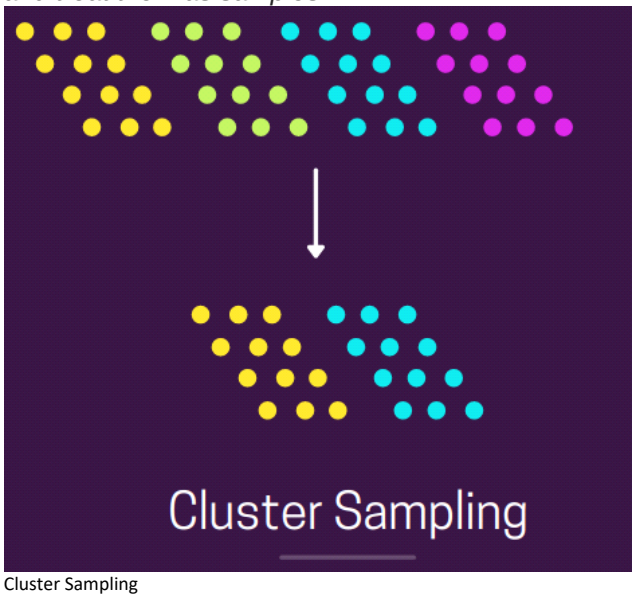
```
stratified_sample, _ = train_test_split(population, test_size=0.9, stratify=population[['label']])
```

```
print (stratified_sample)
```

You can also implement it without the lib., read this.

Cluster Sampling

In Cluster sampling, we **divide the entire population into subgroups**, wherein, each of those subgroups has similar characteristics to that of the population when considered in totality. Also, instead of sampling individuals, we **randomly select the entire subgroups**. *As can be seen in the below fig. that we had 4 clusters with similar properties (size and shape), we randomly select two clusters and treat them as samples*.



Cluster Sampling

Real-Life example — Class of 120 students divided into groups of 12 for a common

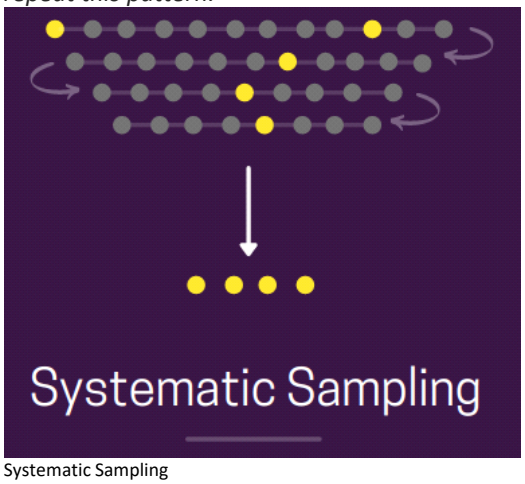
class project. Clustering parameters like *(Designation, Class, Topic)* are all similar over here as well.

You can implement it using python as shown below —

```
import numpy as npclusters=5

pop_size = 100

sample_clusters=2#assigning cluster ids sequentially from 1 to 5 on gap of 20

cluster_ids = np.repeat([range(1,clusters+1)], pop_size/clusters)cluster_to_select =
random.sample(set(cluster_ids), sample_clusters)indexes = [i for i, x in enumerate(cluster_ids) if x in
cluster_to_select]cluster_associated_elements = [el for idx, el in enumerate(range(1, 101)) if idx in
indexes]print (cluster_associated_elements)
```

Systematic Sampling

Systematic sampling is about sampling items from the population at **regular predefined intervals***(basically fixed and periodic intervals)*. *For example — Every 5th element, 21st element and so on.* This sampling method tends to be more effective than the vanilla random sampling method in general. *Below fig. shows a pictorial view of the same — We sample every 9th and 7th element in order and then repeat this pattern.*



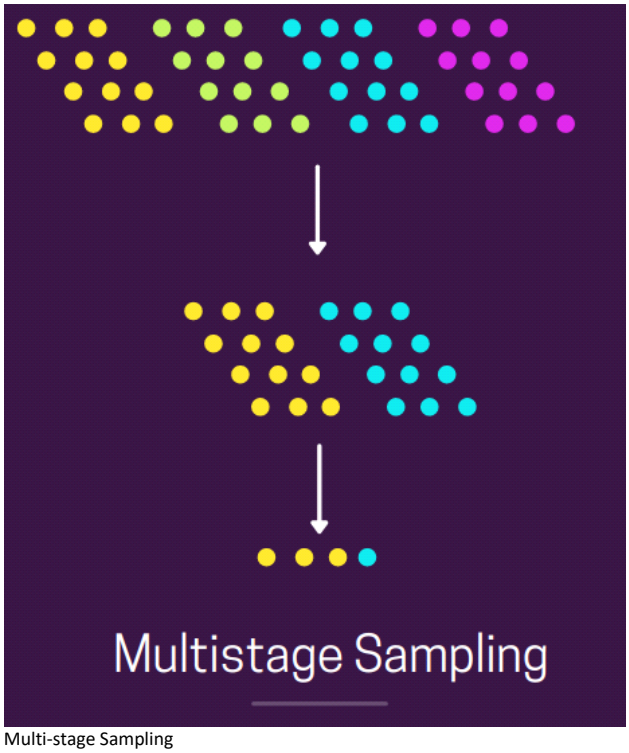
Systematic Sampling

You can implement it using python as shown below —

```
population = 100

step = 5sample = [element for element in range(1, population, step)]

print (sample)
```

Multistage sampling

Under Multistage sampling, we **stack multiple sampling methods** one after the other. For example, at the first stage, cluster sampling can be used to choose clusters from the population and then we can perform random sampling to choose elements from each cluster to form the final set. *Below fig. shows a pictorial view of the same —*



Multi-stage Sampling

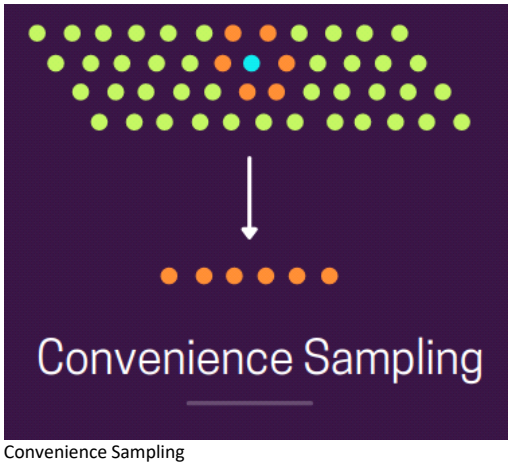
You can implement it using python as shown below —

```
import numpy as npclusters=5

pop_size = 100

sample_clusters=2

sample_size=5#assigning cluster ids sequentially from 1 to 5 on gap of 20

cluster_ids = np.repeat([range(1,clusters+1)], pop_size/clusters)cluster_to_select = random.sample(set(cluster_ids), sample_clusters)indexes = [i for i, x in enumerate(cluster_ids) if x in cluster_to_select]cluster_associated_elements = [el for idx, el in enumerate(range(1, 101)) if idx in indexes]print (random.sample(cluster_associated_elements, sample_size))
```

Non-Probability Sampling

Convenience Sampling

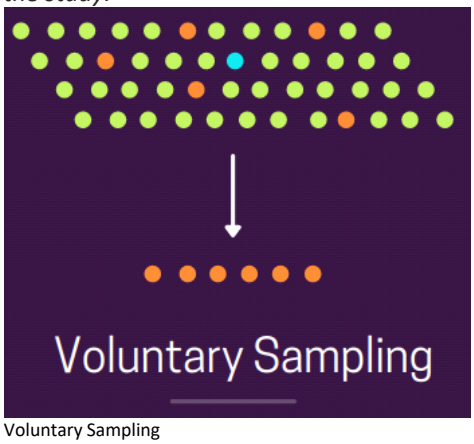
Under convenience sampling, the researcher includes only those **individuals who are most accessible and available to participate in the study**. *Below fig. shows the pictorial view of the same — Blue dot is the researcher and orange dots are the most accessible set of people in orange's vicinity.*



Convenience Sampling

Voluntary Sampling

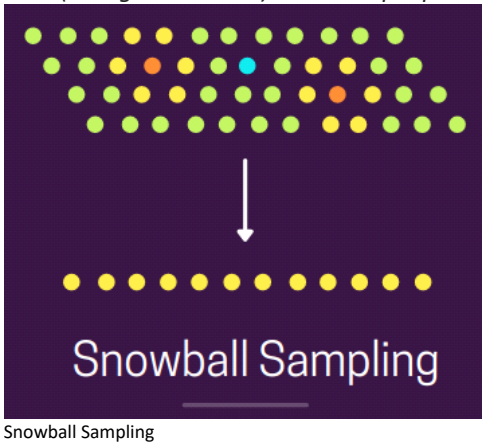
Under Voluntary sampling, **interested people usually take part by themselves** by filling in some sort of survey forms. A good example of this is the youtube survey about “Have you seen any of these ads”, which has been recently shown a lot. Here, the **researcher who is conducting the survey has no right to choose anyone**. *Below fig. shows the pictorial view of the same — Blue dot is the researcher, orange one’s are those who voluntarily agreed to take part in the study.*



Voluntary Sampling

Snowball Sampling

Under Snowball sampling, th**e final set is chosen via other participants**, i.e. The researcher asks other known contacts to find people who would like to participate in the study. *Below fig. shows the pictorial view of the same — Blue dot is the researcher, orange ones are known contacts(of the researcher), and yellow ones (orange’s contacts) are other people that got ready to participate in the study.*

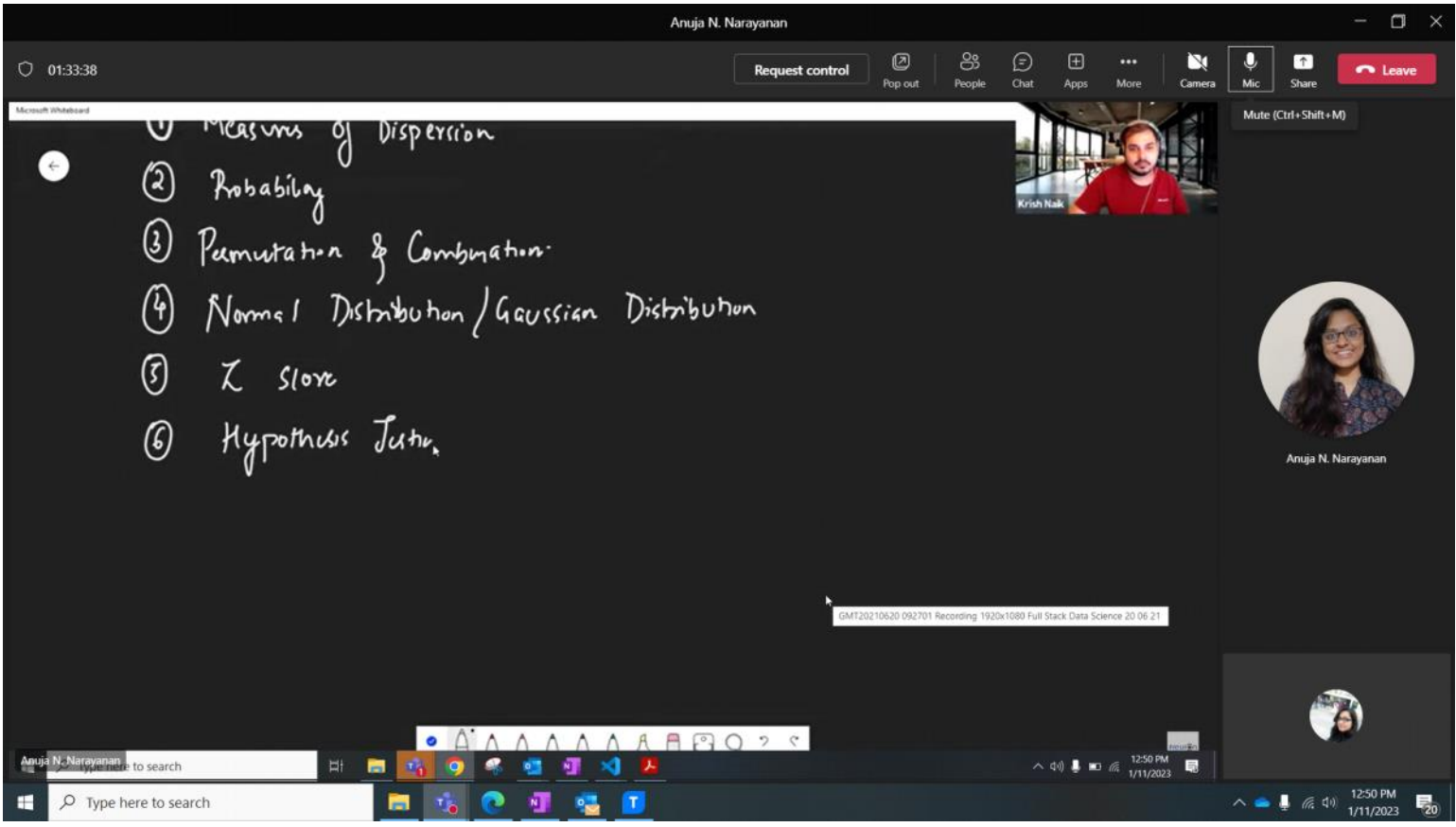


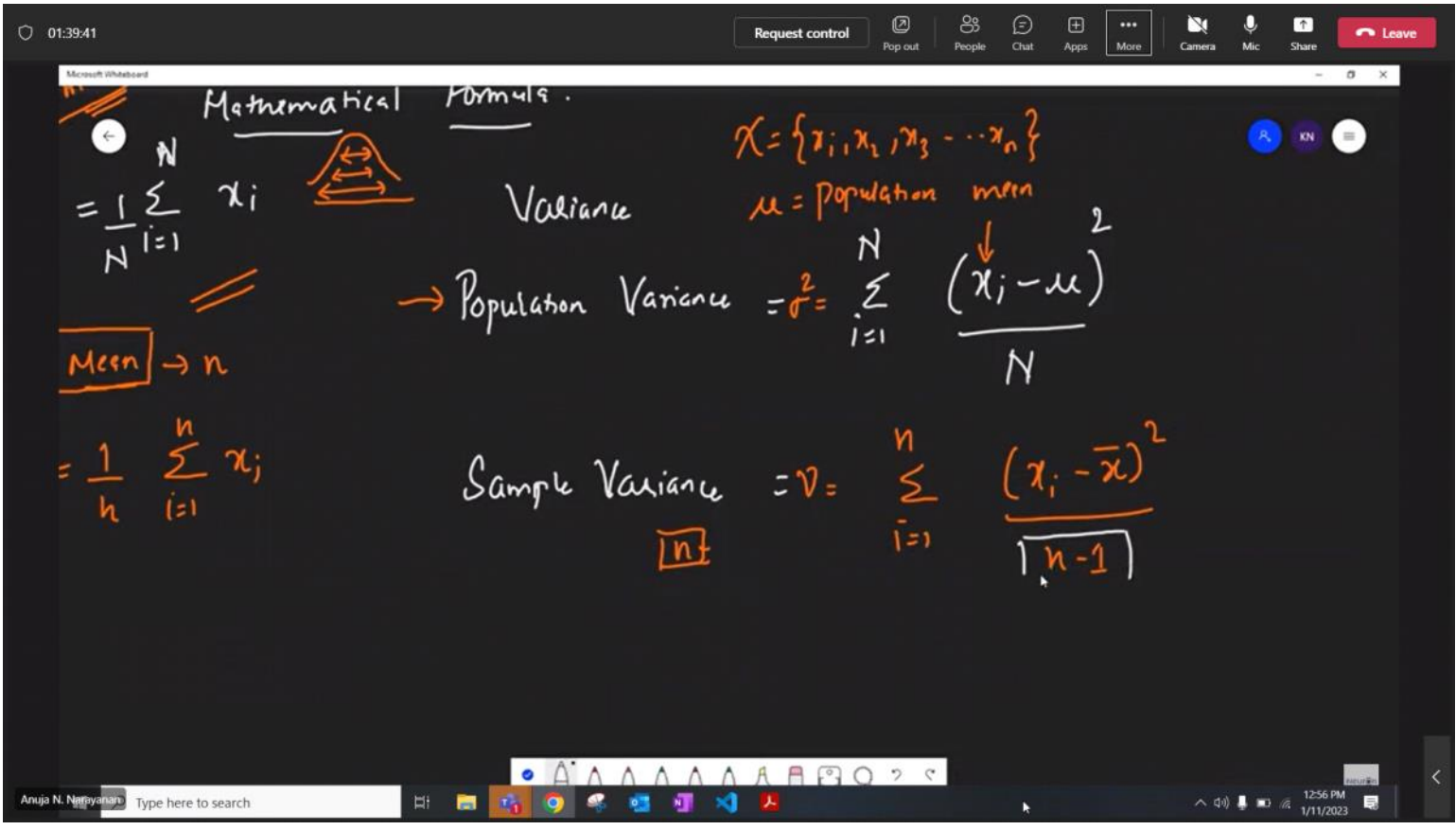
Snowball Sampling

From <https://towardsdatascience.com/8-types-of-sampling-techniques-b21adcdd2124>

KN(lec-2)

Wednesday, January 11, 2023 12:50 PM





01:41:19 Request control Pop out People Chat Apps More Camera Mic Share Leave

Microsoft Whiteboard

Formula.

$X = \{x_1, x_2, x_3, \dots x_n\}$

μ = Population mean

Variance

Population Variance $= \sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$

{ Standard Deviation }

$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}}$

Sample Variance $= v = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

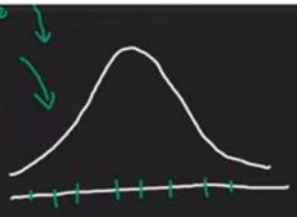
$v = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$

Anuja N. Narayanan Type here to search 12:57 PM 1/11/2023

02:09:46 Request control Pop out People Chat Apps More Camera Mic Share Leave

Microsoft Whiteboard Mute (Ctrl+Shift+M)

Bell Curve



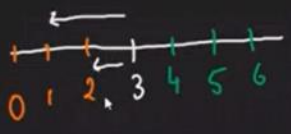
$X = \{1, 2, 3, 4, 5\}$ → 4

$\mu = 3$

$\sigma = 1$ {Eg}

Calculation

$Z\text{-score} = \frac{x_i - \mu}{\sigma}$



$= \frac{1-3}{1} = -2$

$= \frac{2-3}{1} = -1$

S.D = Small

Anuja N. Narayanan Type here to search 1:26 PM 1/11/2023

02:21:03

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Mute (Ctrl+Shift+M)

GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21

Watch later | Share

Same Scale — India Cricket Teams England

4 Test

2020

Avg score = 181

$\sigma = 12$

Environment → { Score }

{ India score in the final = 187

$Z = \frac{x_i - \mu}{\sigma} = \frac{187 - 181}{12} = 6/12 = 0.5$

187

2021

Avg score = 182

$\sigma = 5$

Raining

India final = 185

$Z = \frac{185 - 182}{5} = 3/5 = 0.6$

2:29:17

Anuja N. Narayanan | Type here to search | 1:37 PM 1/11/2023

02:35:56

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21

Watch later | Share

EXP : Taking out a card from deck

Non mutual Exclusive Event

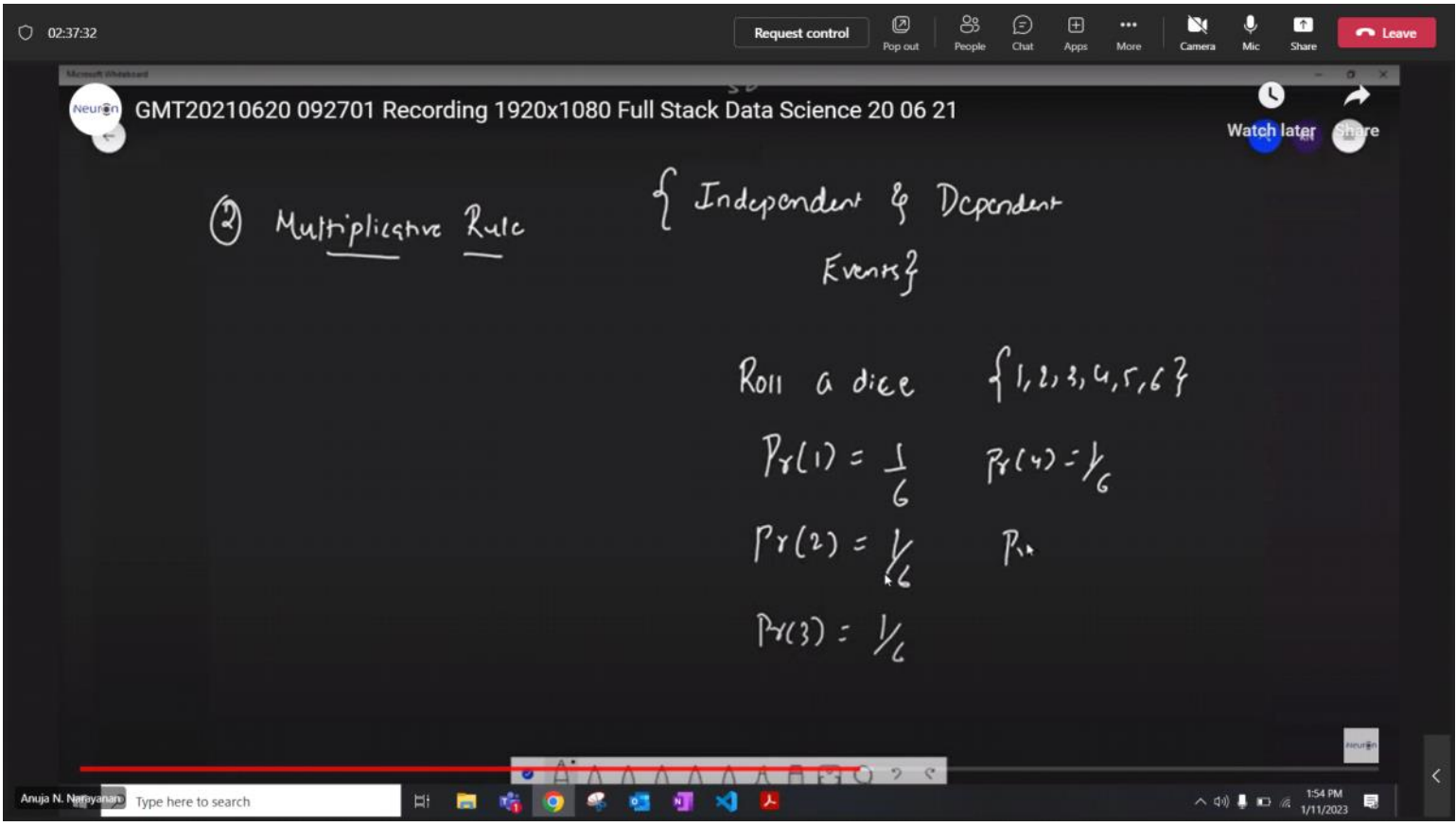
$Pr(A \text{ or } B) = Pr(A) + Pr(B) - Pr(A \cap B)$

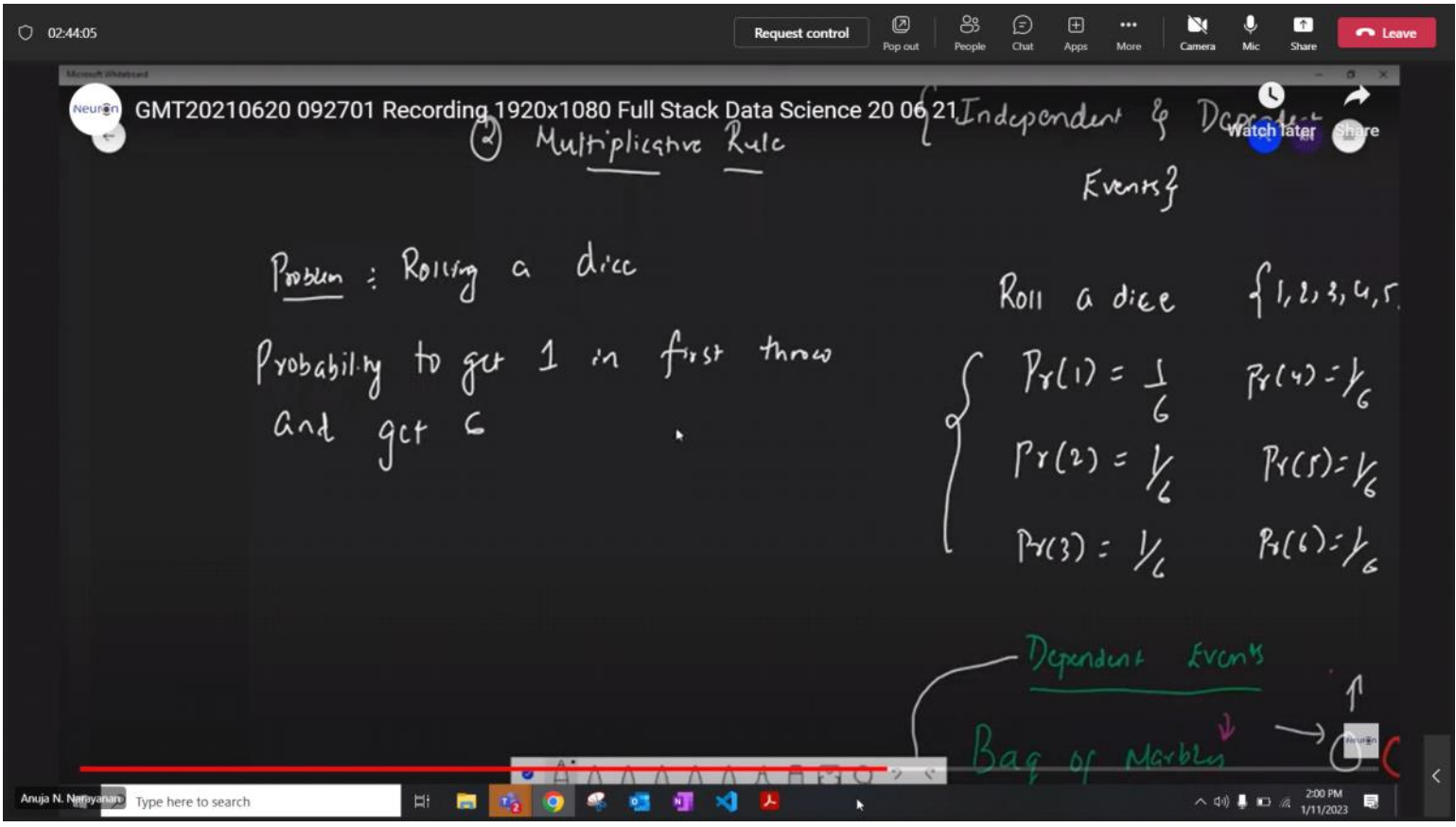
$Pr(K \text{ or } ♡) = Pr(K) + Pr(♡) - P(K \cap ♡)$

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$

$= \frac{16}{52}$

Anuja N. Narayanan | Type here to search | 1:52 PM 1/11/2023


02:37:32
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21
Watch later
Share
(2) Multiplicative Rule
{ Independent & Dependent Events }
Roll a dice {1,2,3,4,5,6}
Pr(1) = 1/6
Pr(2) = 1/6
Pr(3) = 1/6
Pr(4) = 1/6
Anuja N. Narayanan
Type here to search
1:54 PM
1/11/2023


02:44:05
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21
Watch later
Share
(2) Multiplicative Rule
{ Independent & Dependent Events }
Problem : Rolling a dice
Probability to get 1 in first throw and get 6
Roll a dice {1,2,3,4,5,
Pr(1) = 1/6
Pr(2) = 1/6
Pr(3) = 1/6
Pr(4) = 1/6
Pr(5) = 1/6
Pr(6) = 1/6
Dependent Events
Bag of Marbles
Anuja N. Narayanan
Type here to search
2:00 PM
1/11/2023

02:44:13 Request control Pop out People Chat Apps More Camera Mic Share Leave

GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21

Watch later Share

Probability to get 1 in first throw and get 6 in second throw

$Pr(1 \text{ and } 6) = Pr(1) * Pr(6)$

$Pr(1) = 1/6$ $Pr(2) = 1/6$ $Pr(3) = 1/6$ $Pr(4) = 1/6$ $Pr(5) = 1/6$ $Pr(6) = 1/6$

Dependent Events

Bag of Marbles

$Pr(W) = 1/5$ $Pr(R)$.

2:04:51

Anuja N. Narayanan Type here to search 2:00 PM 1/11/2023

02:44:26 Request control Pop out People Chat Apps More Camera Mic Share Leave

GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21

Watch later Share

Dependent : Probability of taking out white and then the Red Marble

Event A

$Pr(W) = 1/5$

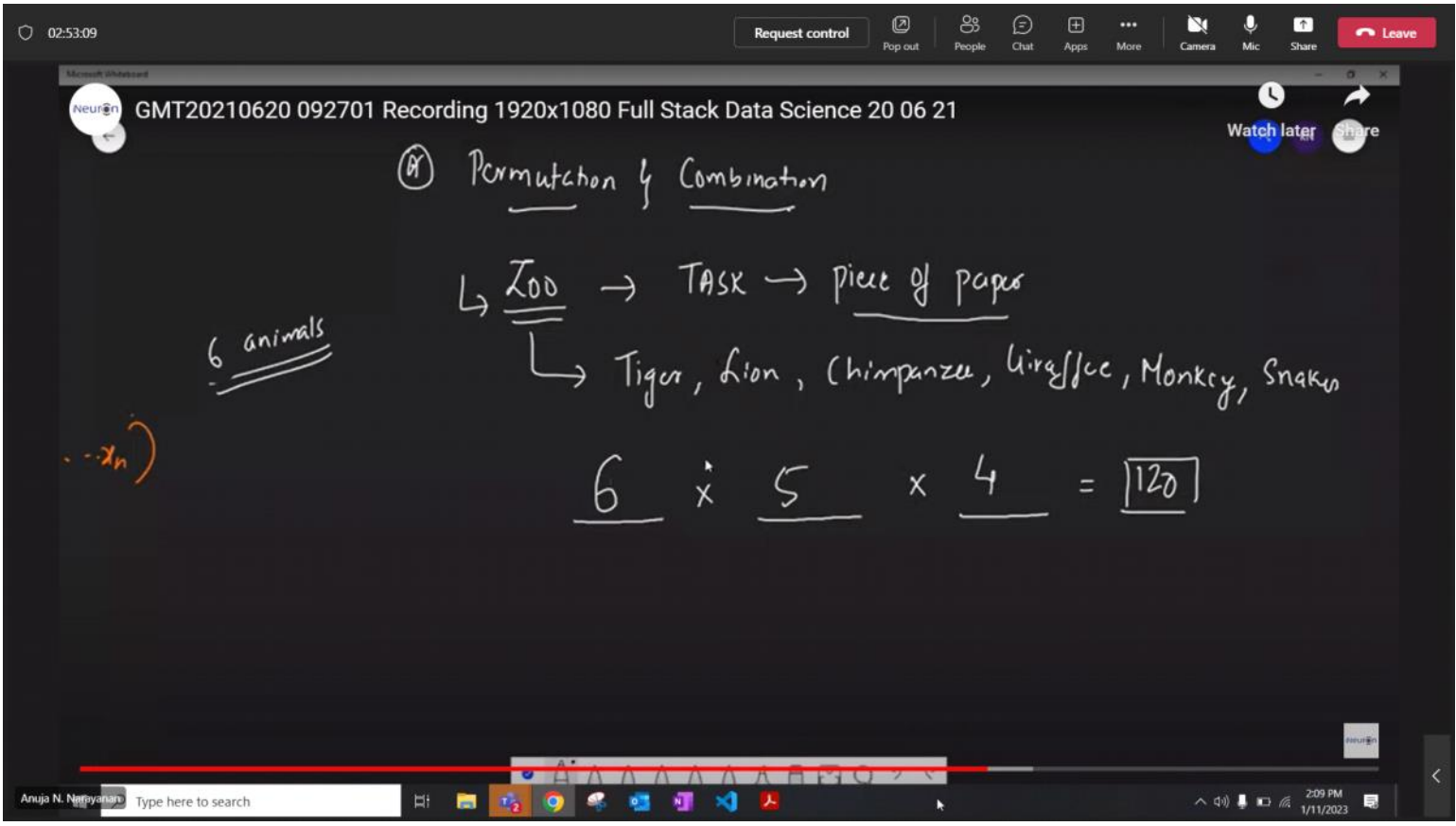
$Pr(R/W) = 1/4$

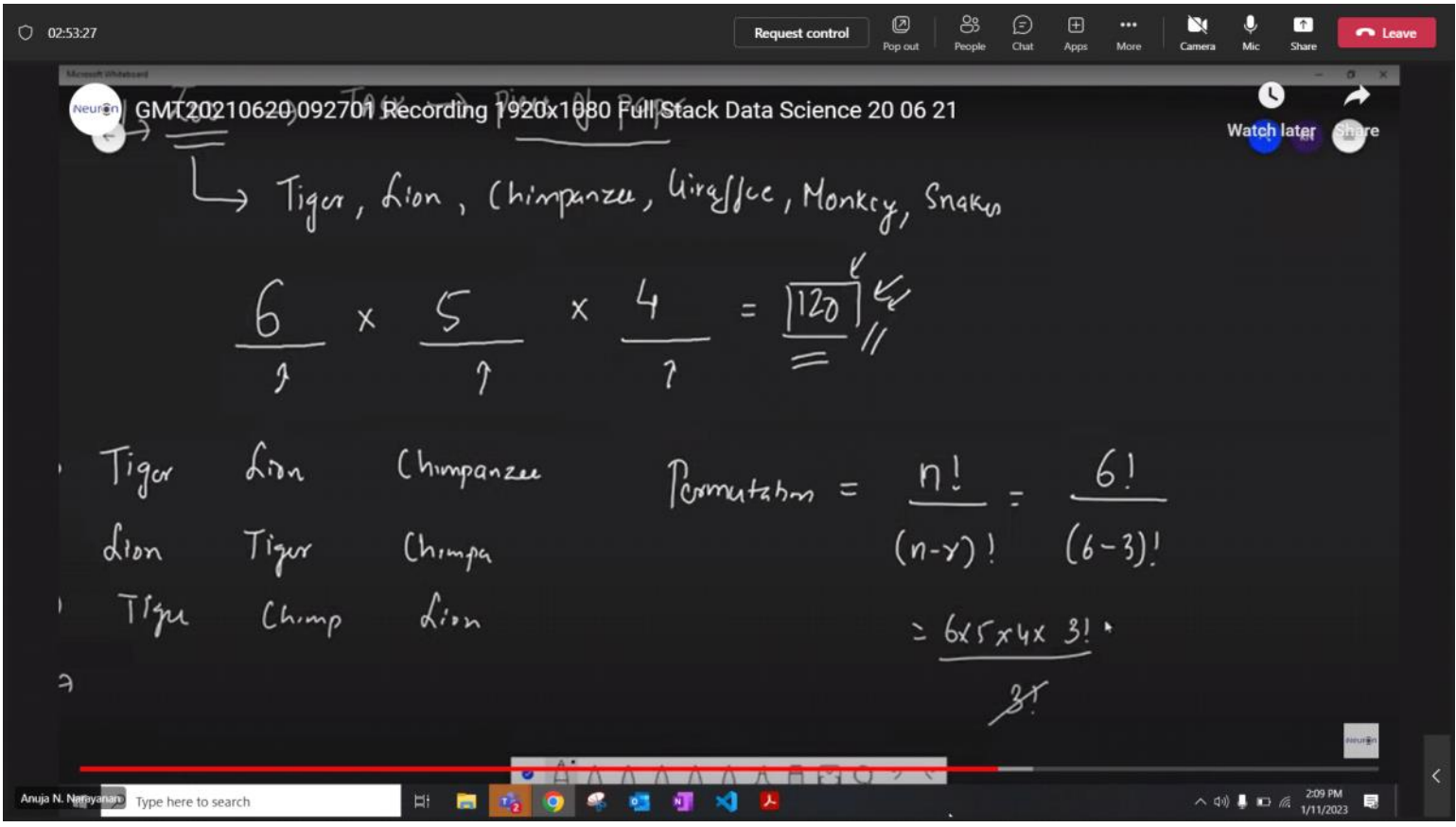
Dependent Event

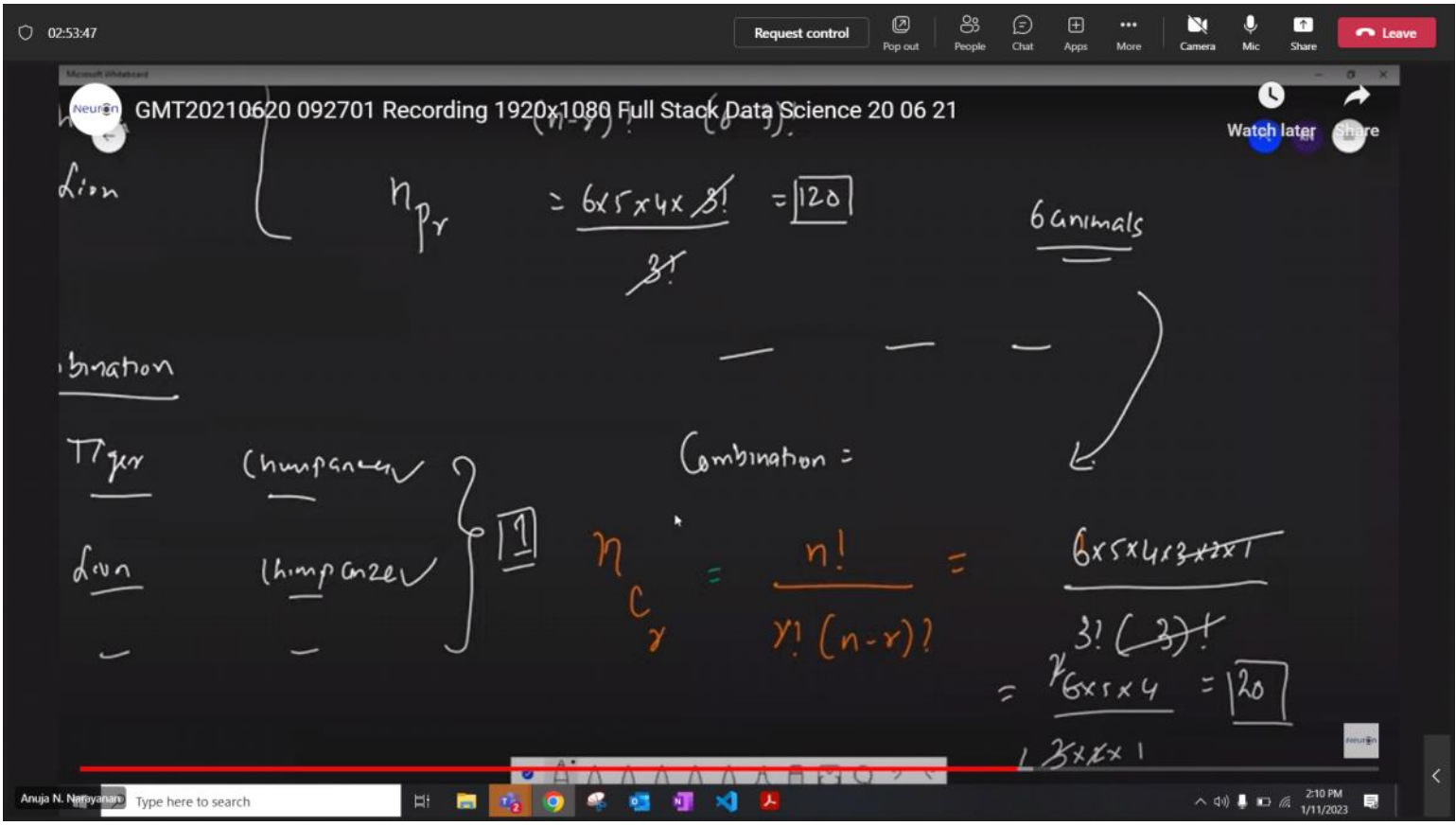
$Pr(W \text{ and } Red) = Pr(W) * Pr(R/W)$

$= 1/5 * 1/4 = 1/20$

Anuja N. Narayanan Type here to search 2:00 PM 1/11/2023

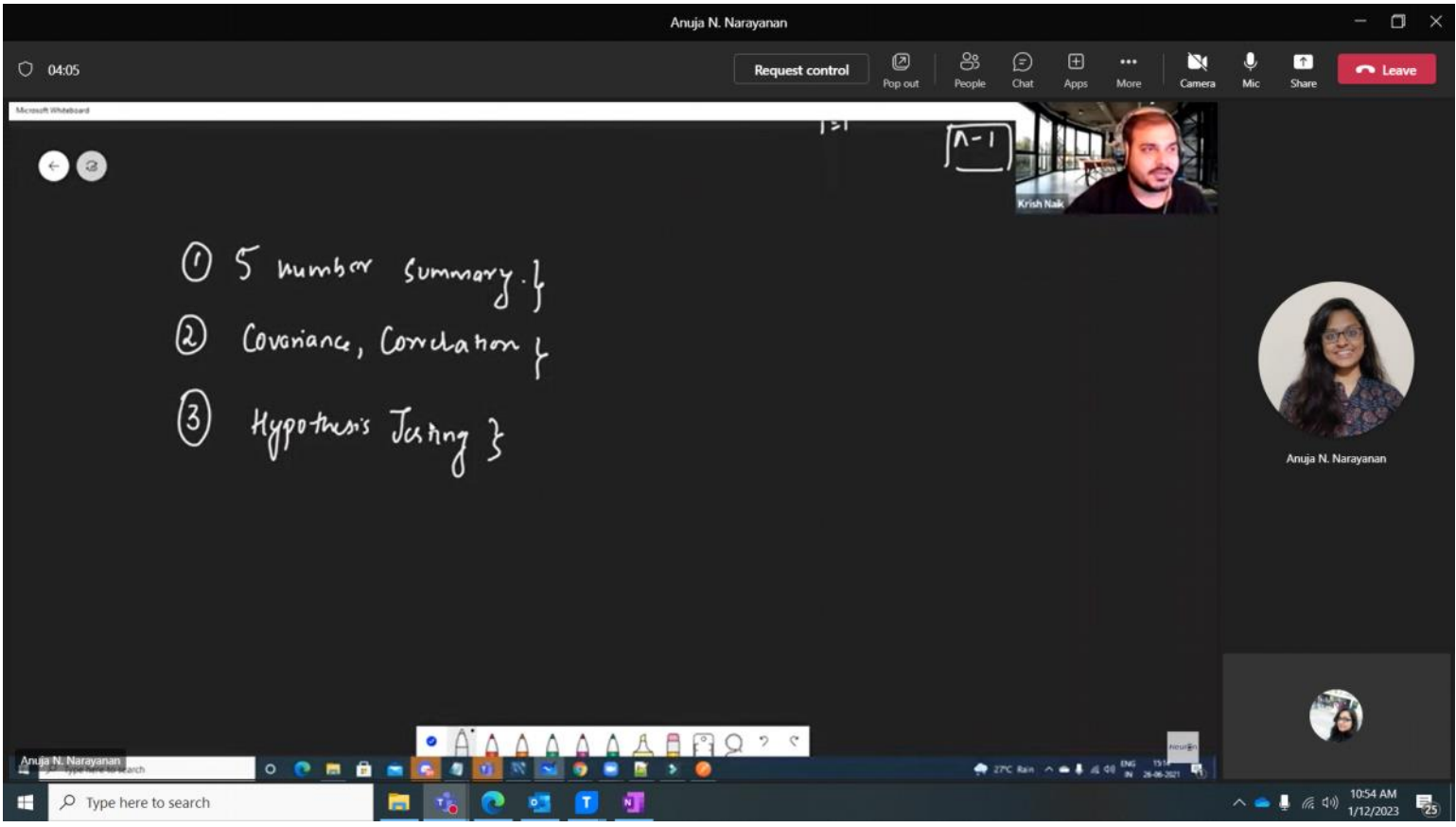

02:53:09
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21
Watch later
Share
(8) Permutation & Combination
Zoo → TASK → Piece of paper
6 animals
Tiger, Lion, Chimpanzee, Giraffe, Monkey, Snakes
6 x 5 x 4 = 120
Type here to search
2:09 PM
1/11/2023

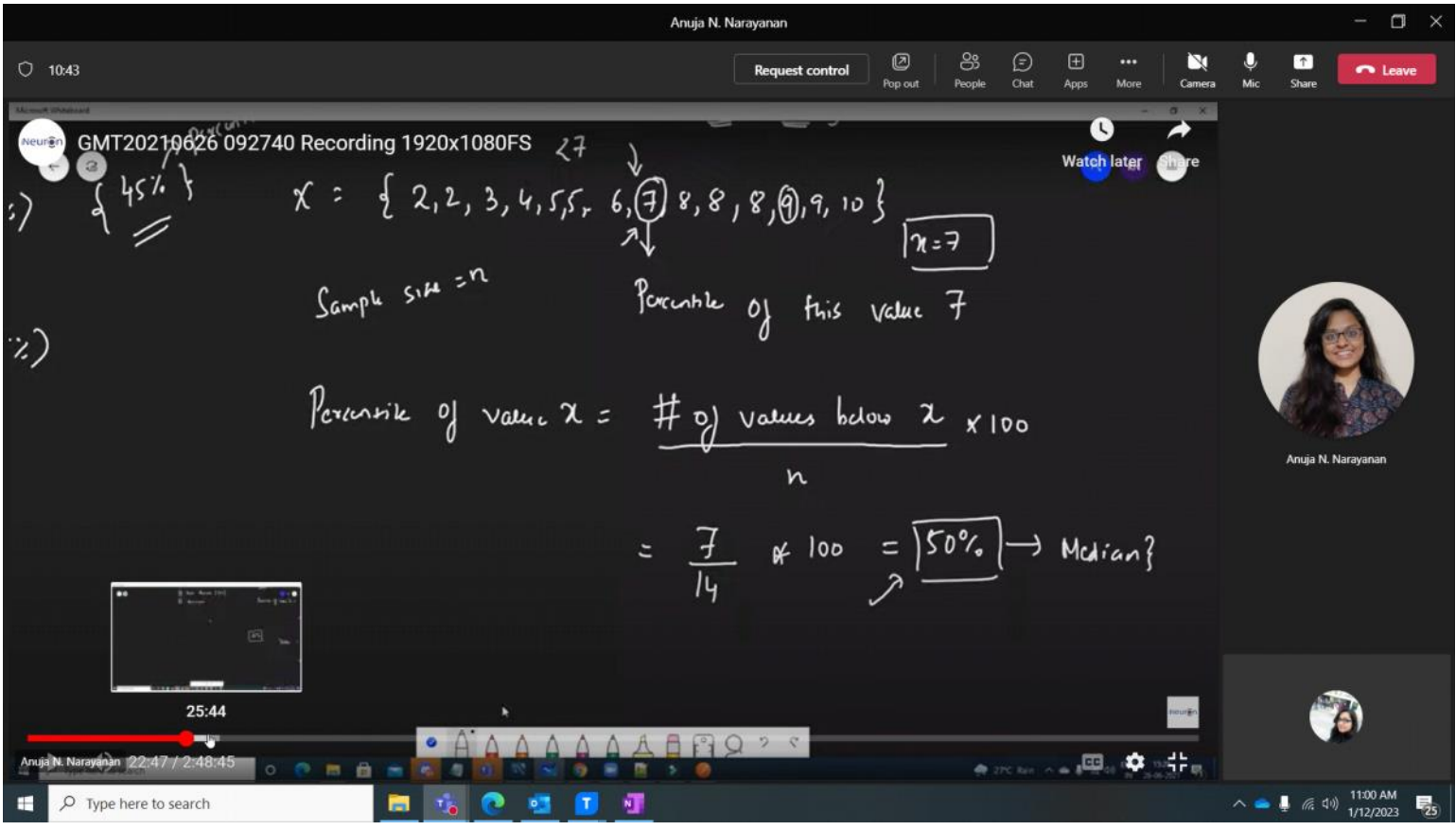

02:53:27
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21
Watch later
Share
Tiger, Lion, Chimpanzee, Giraffe, Monkey, Snakes
6 x 5 x 4 = 120
Tiger Lion Chimpanzee
Lion Tiger Chimpa
Tiger Chimp Lion
Permutation = n! / (n-r)! = 6! / (6-3)!
= 6x5x4x 3! / 3!
Type here to search
2:09 PM
1/11/2023

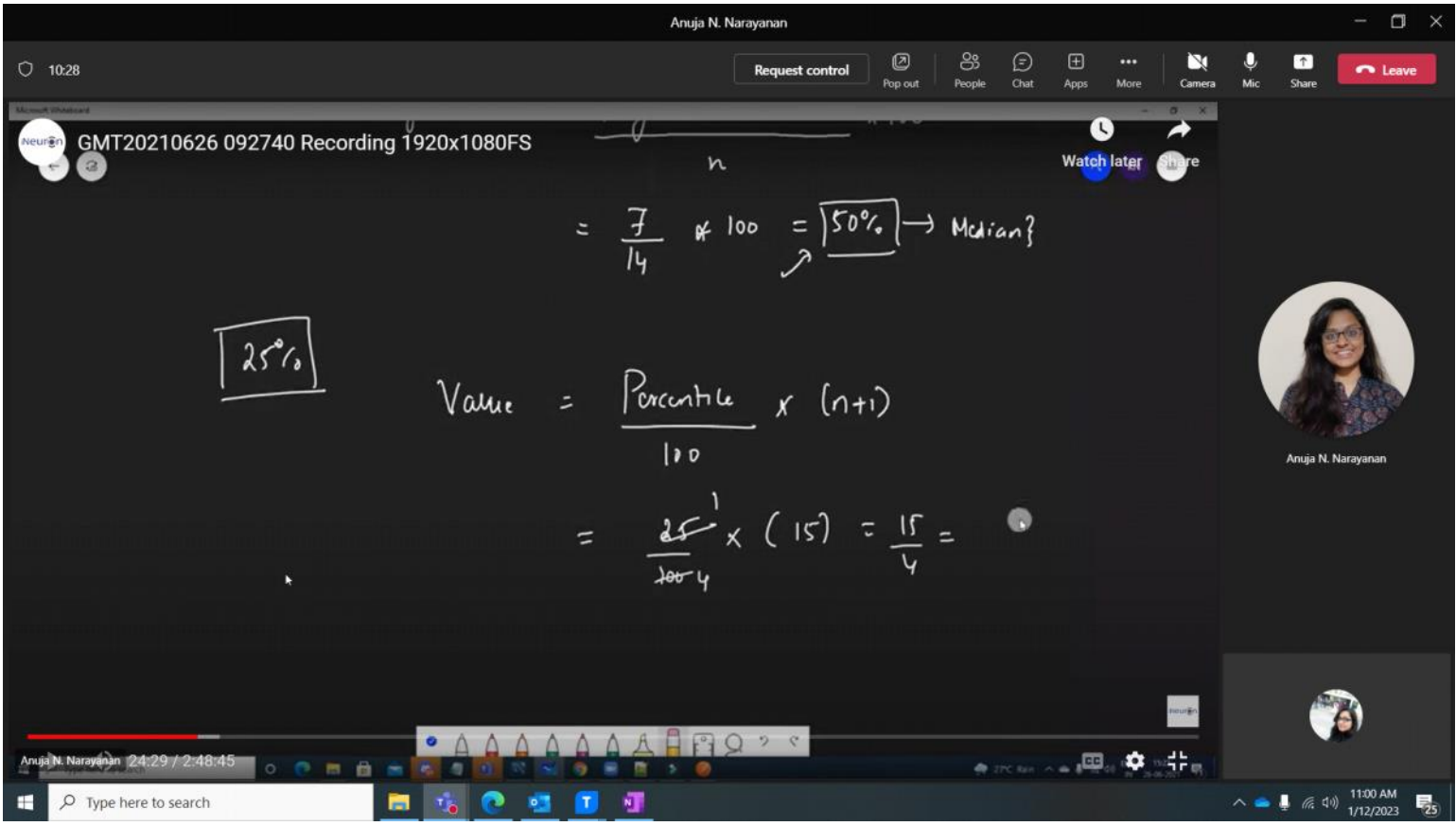

02:53:47
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210620 092701 Recording 1920x1080 Full Stack Data Science 20 06 21
Watch later
Share
6 animals
Combination =
Anuja N. Narayanan
Type here to search
2:10 PM
1/11/2023

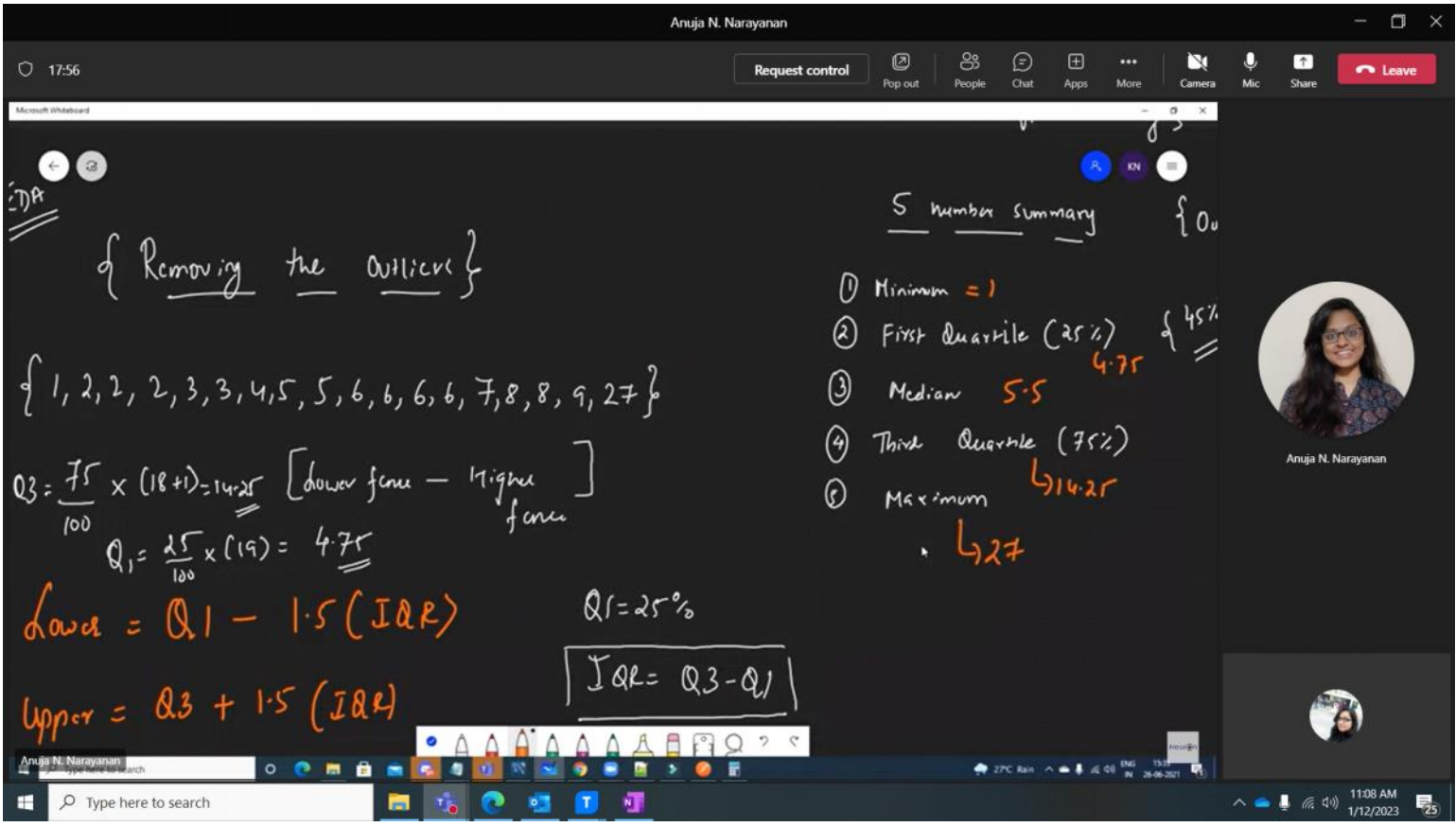
KN(lec-3)

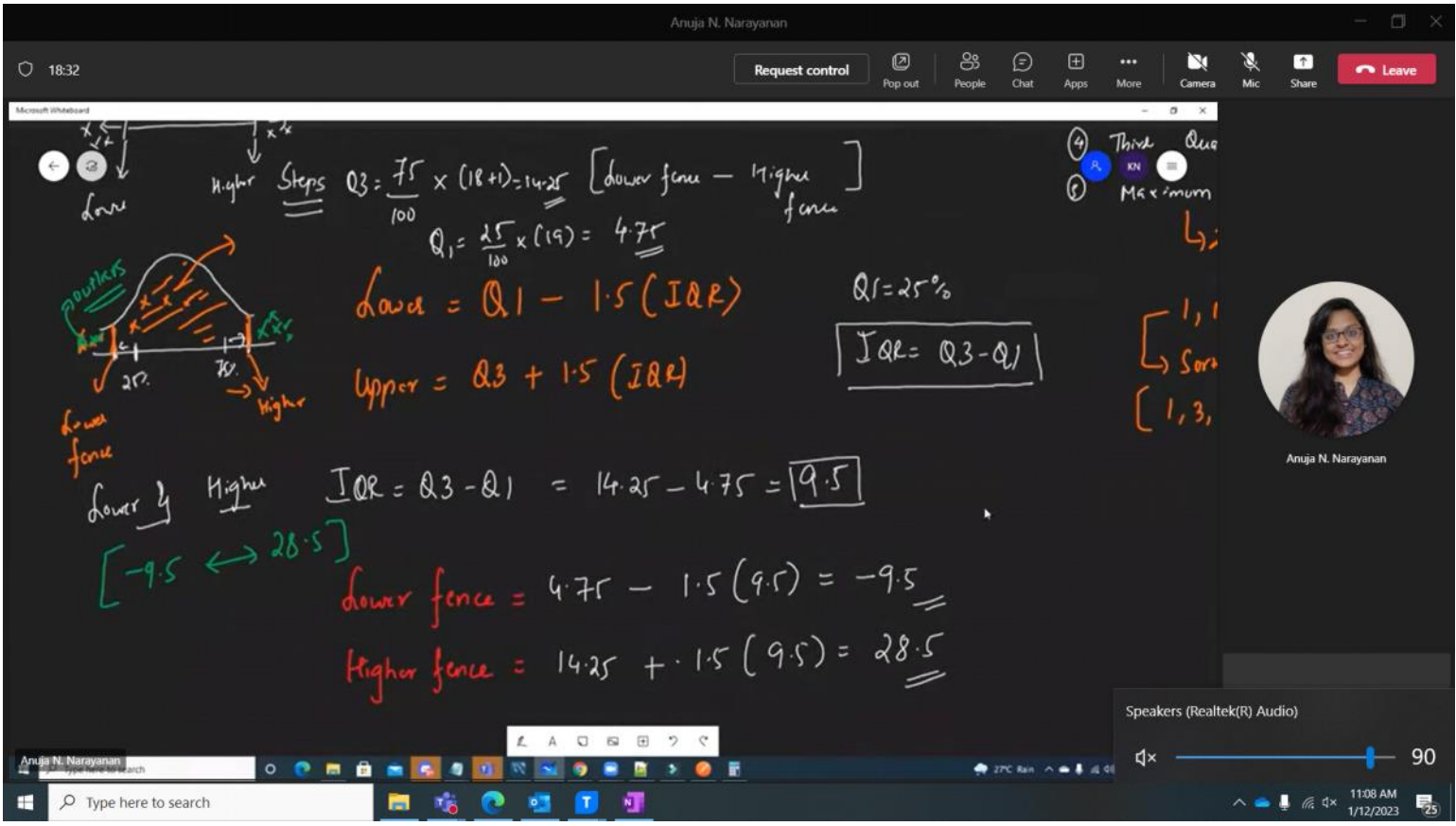
Thursday, January 12, 2023 10:52 AM

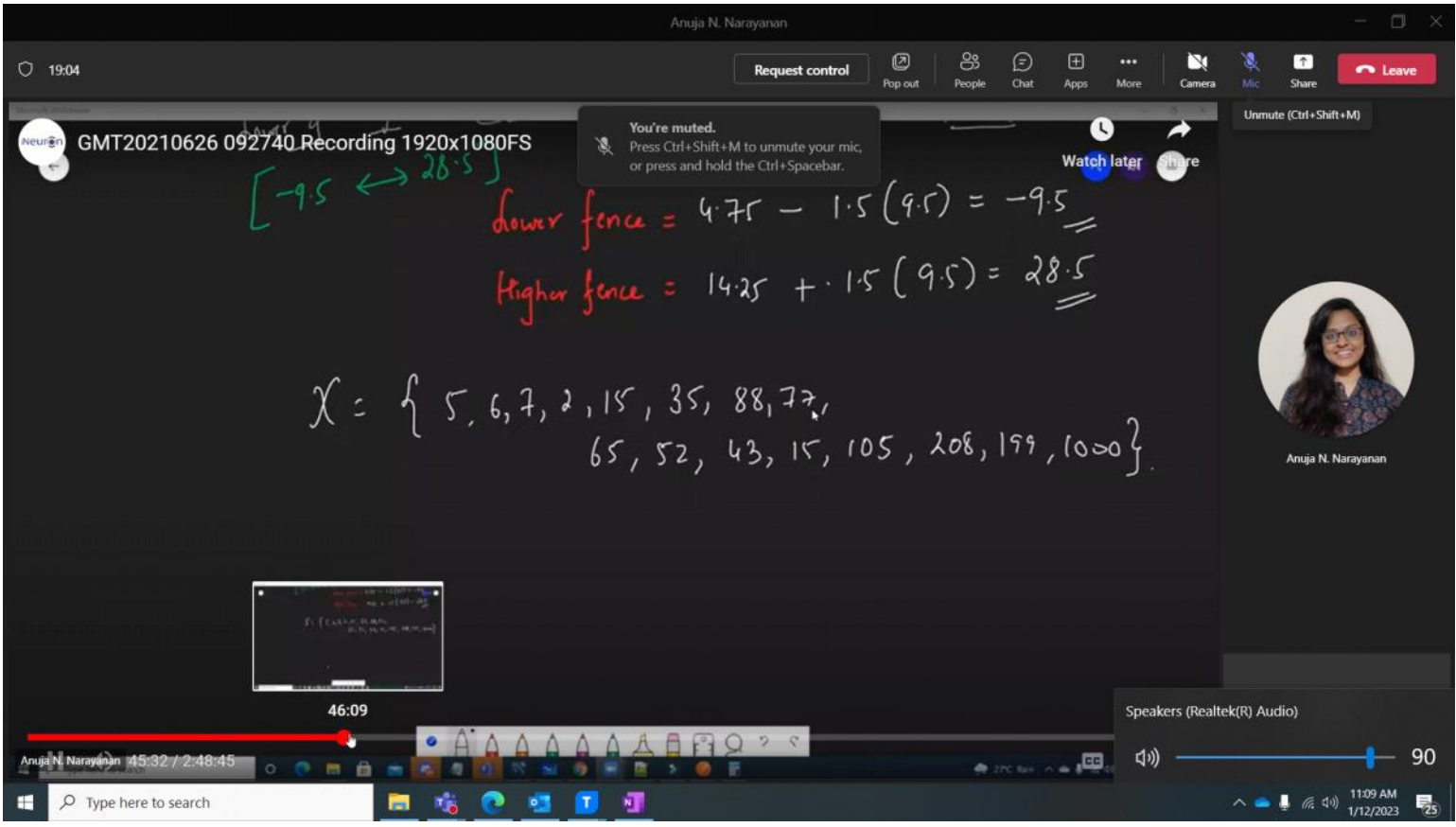


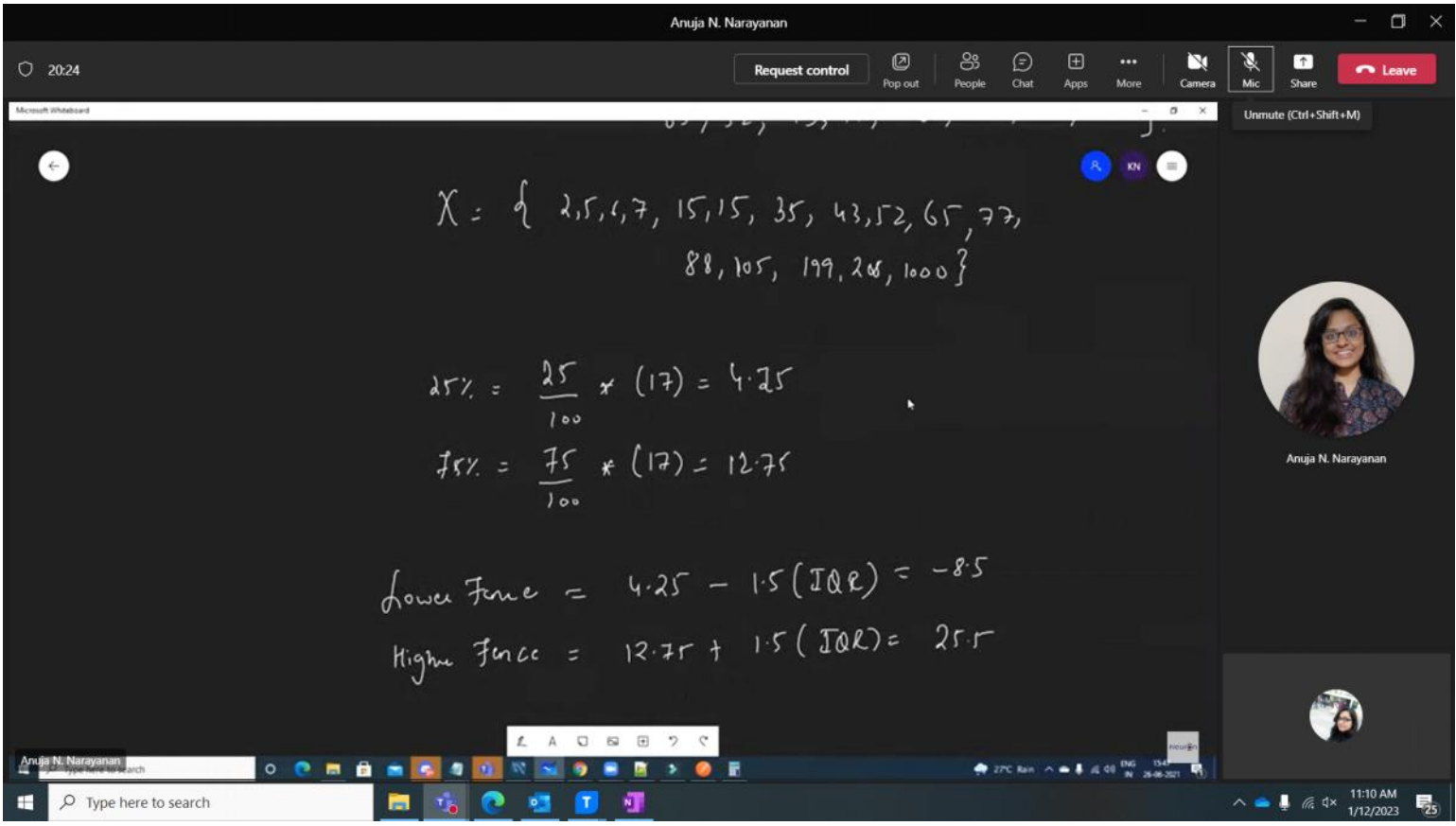


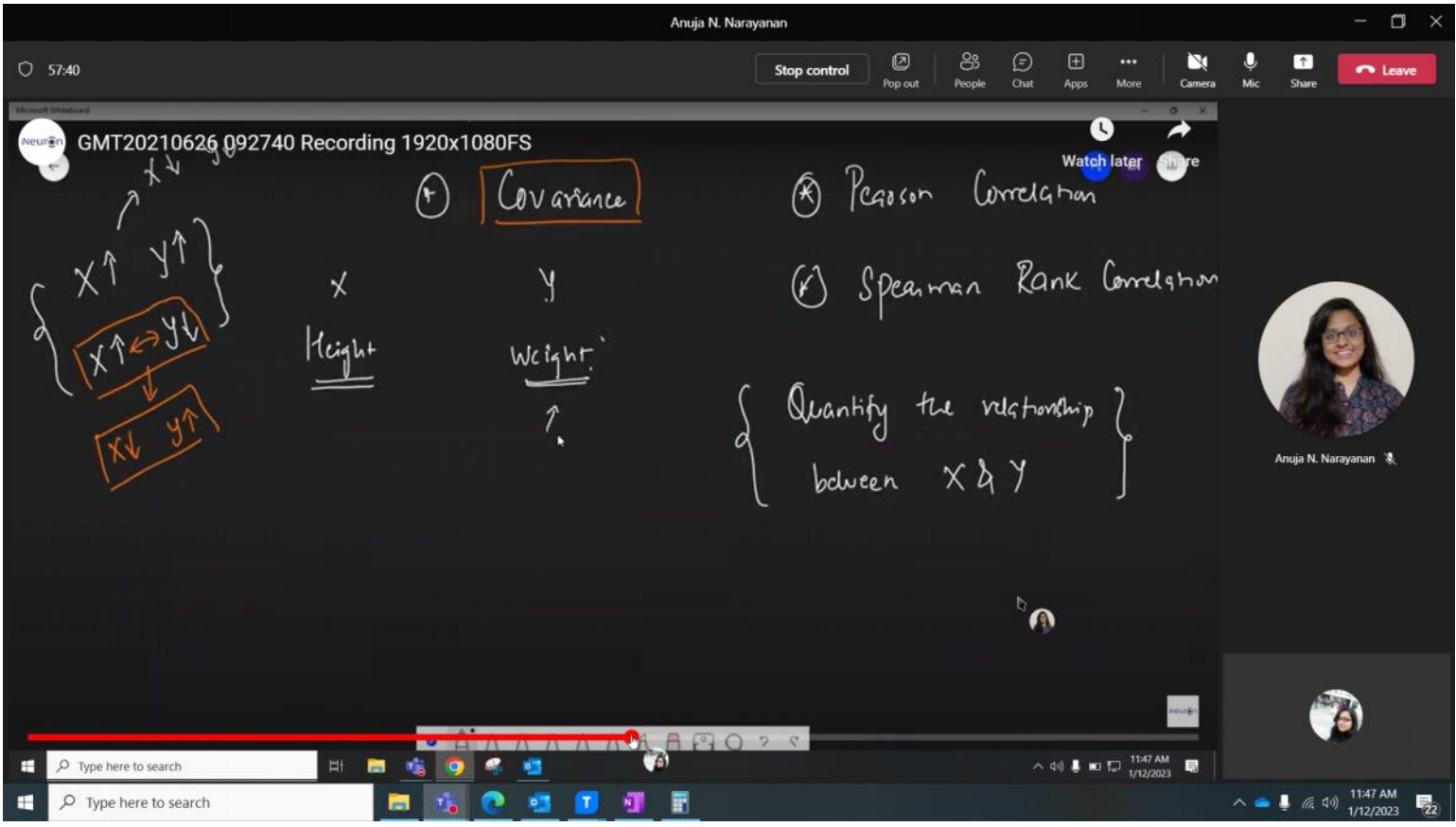

Anuja N. Narayanan
10:28
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
n
= 7/14 × 100 = 50% → Median?
25%
Value = Percentile/100 × (n+1)
= 25/100 × (15) = 15/4 =
Anuja N. Narayanan
Anuja N. Narayanan 24:29 / 2:48:45
Type here to search
11:00 AM
1/12/2023

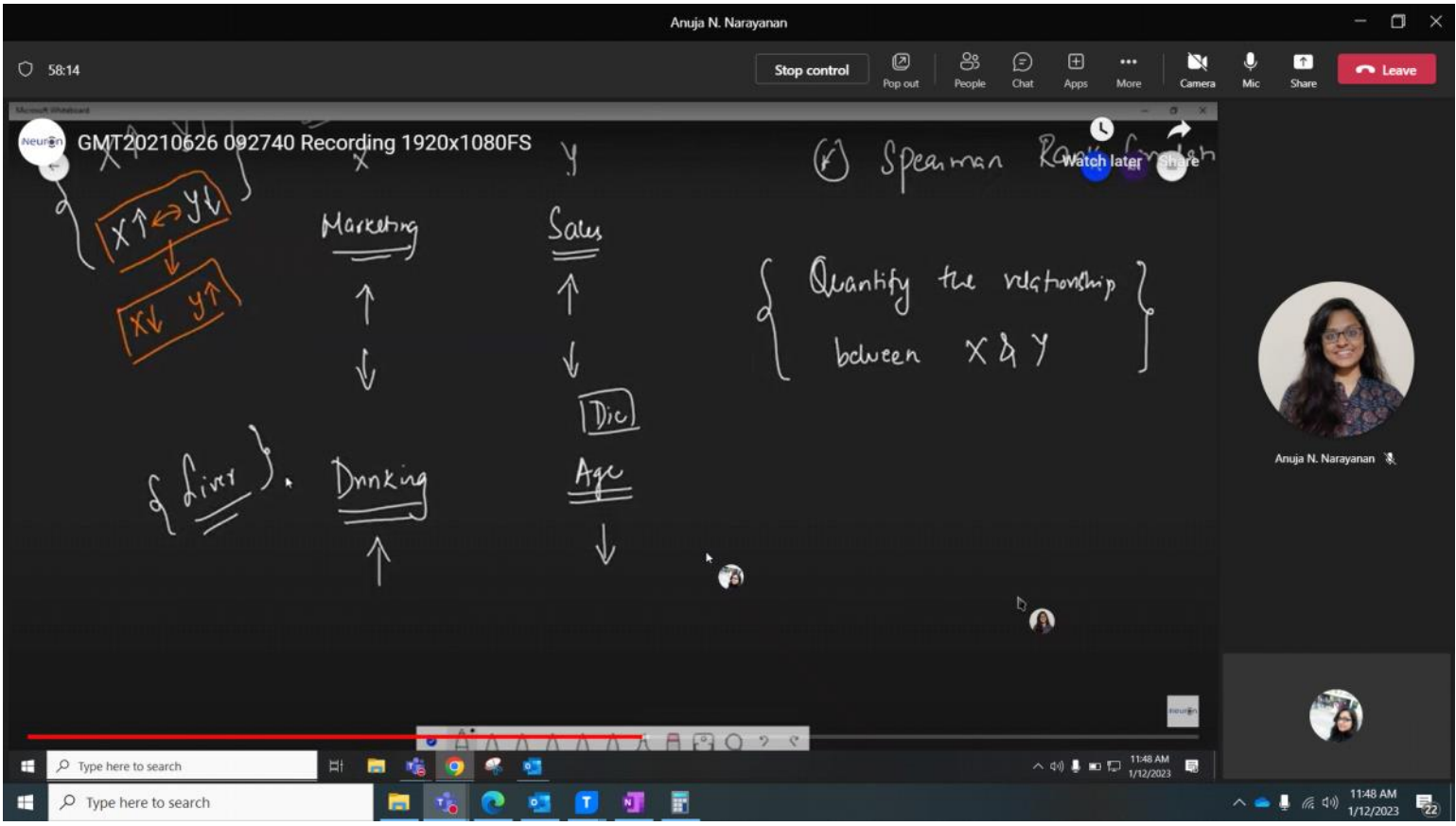

Anuja N. Narayanan
17:56
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
EDA
{ Removing the outliers }
{ 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 }
Q3 = 75/100 × (18+1) = 14.25 [lower fence — higher fence]
Q1 = 25/100 × (19) = 4.75
Lower = Q1 − 1.5 (IQR)
Upper = Q3 + 1.5 (IQR)
Q1 = 25%
IQR = Q3 − Q1
5 number summary
1) Minimum = 1
2) First Quartile (25%) 4.75
3) Median 5.5
4) Third Quartile (75%) 14.25
5) Maximum 27
Anuja N. Narayanan
Type here to search
11:08 AM
1/12/2023

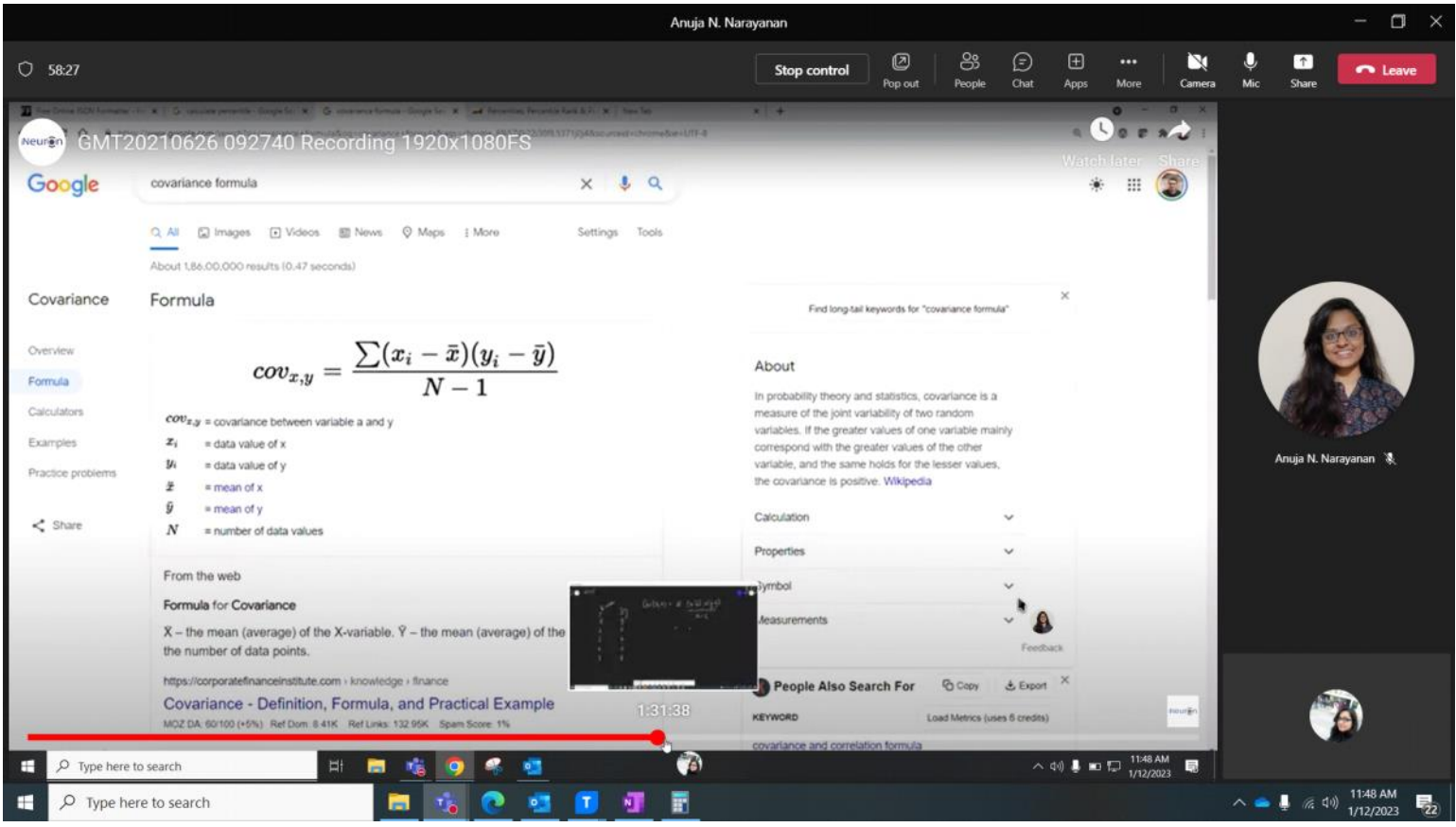

Anuja N. Narayanan
18:32
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Lower
Higher
Steps
$Q_3 = \frac{75}{100} \times (18+1) = 14.25$
[lower fence — higher fence]
$Q_1 = \frac{25}{100} \times (19) = 4.75$
outliers
25
75
Lower fence
Higher
Lower = Q1 − 1.5 (IQR)
Upper = Q3 + 1.5 (IQR)
Q1 = 25%
IQR = Q3 − Q1
Lower & Higher
IQR = Q3 − Q1 = 14.25 − 4.75 = 9.5
[−9.5 ⟷ 28.5]
lower fence = 4.75 − 1.5 (9.5) = −9.5
Higher fence = 14.25 + 1.5 (9.5) = 28.5
Third Quartile
Maximum
Anuja N. Narayanan
Speakers (Realtek(R) Audio)
90
Type here to search
11:08 AM
1/12/2023

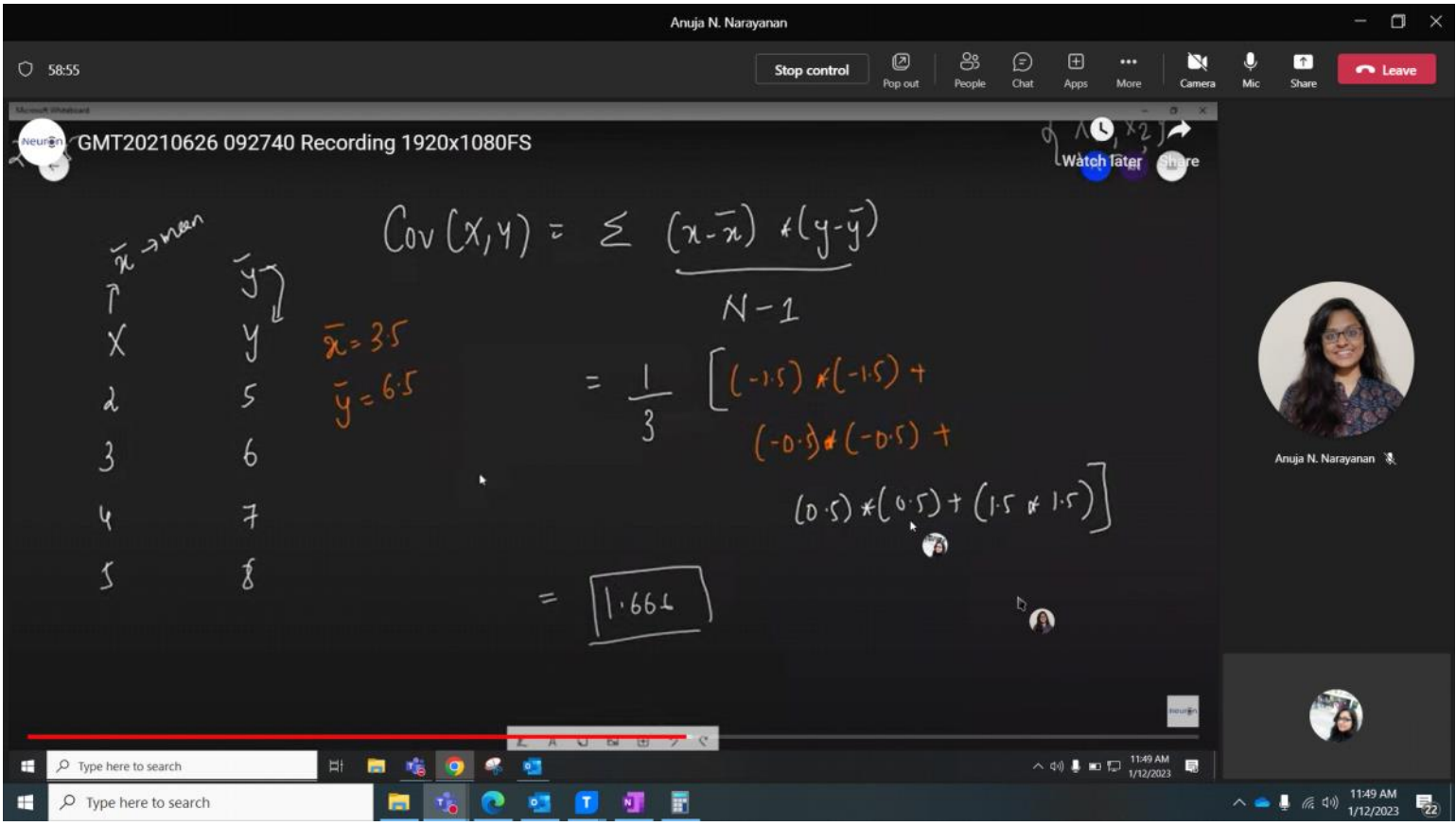

Anuja N. Narayanan
19:04
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
You're muted.
Press Ctrl+Shift+M to unmute your mic, or press and hold the Ctrl+Spacebar.
Unmute (Ctrl+Shift+M)
Watch later
Share
[−9.5 ⟷ 28.5]
lower fence = 4.75 − 1.5 (9.5) = −9.5
Higher fence = 14.25 + 1.5 (9.5) = 28.5
X = { 5, 6, 7, 2, 15, 35, 88, 77,
65, 52, 43, 15, 105, 208, 199, 1000 }
46:09
Anuja N. Narayanan 45:32 / 2:48:45
Anuja N. Narayanan
Speakers (Realtek(R) Audio)
90
Type here to search
11:09 AM
1/12/2023

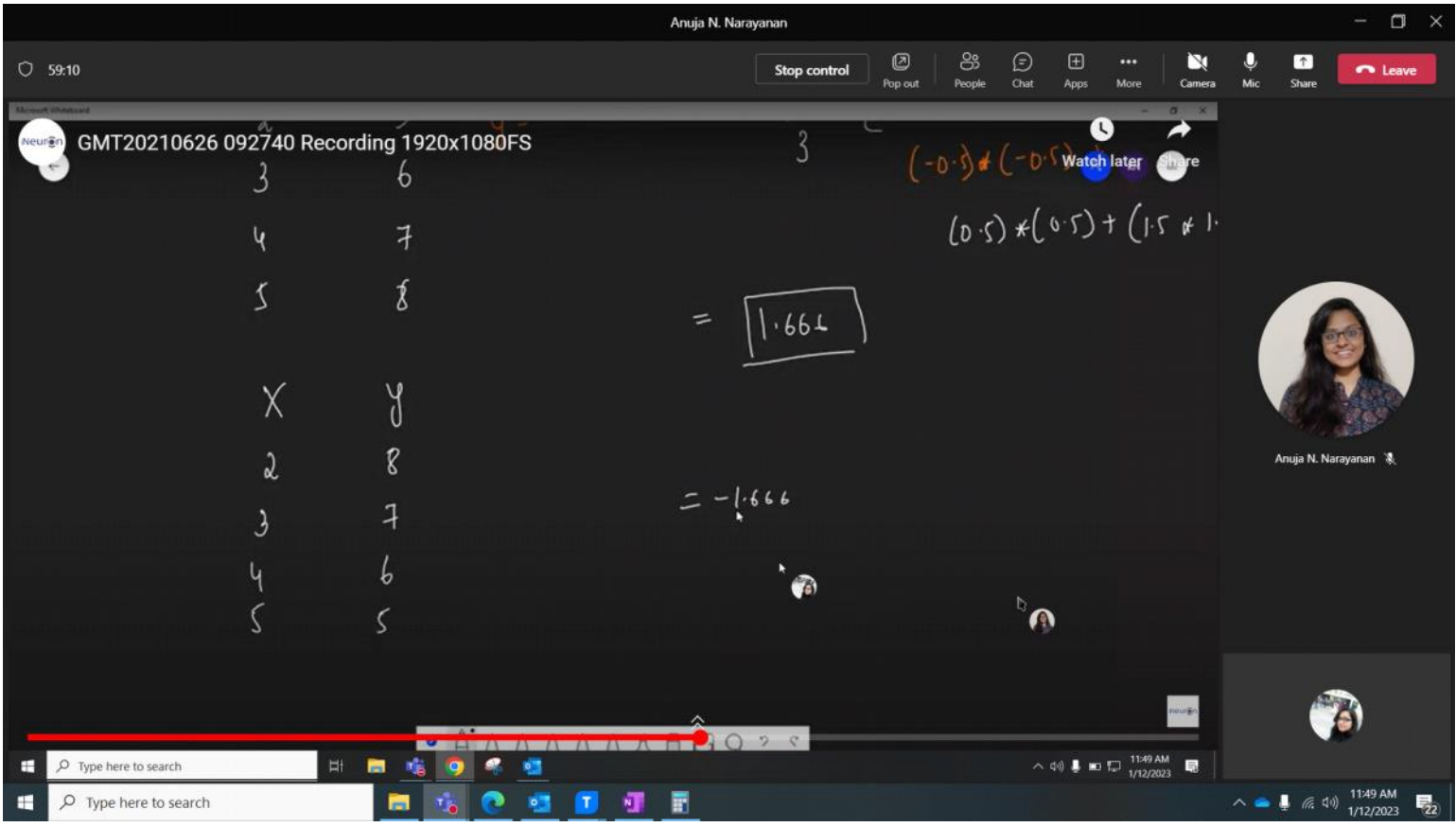

Anuja N. Narayanan
20:24
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Unmute (Ctrl+Shift+M)
X = { 2,5,6,7, 15,15, 35, 43,52, 65, 77,
88, 105, 199, 208, 1000 }
25% = 25/100 * (17) = 4.25
75% = 75/100 * (17) = 12.75
Lower Fence = 4.25 − 1.5 (IQR) = −8.5
Higher Fence = 12.75 + 1.5 (IQR) = 25.5
Anuja N. Narayanan
Type here to search
11:10 AM
1/12/2023

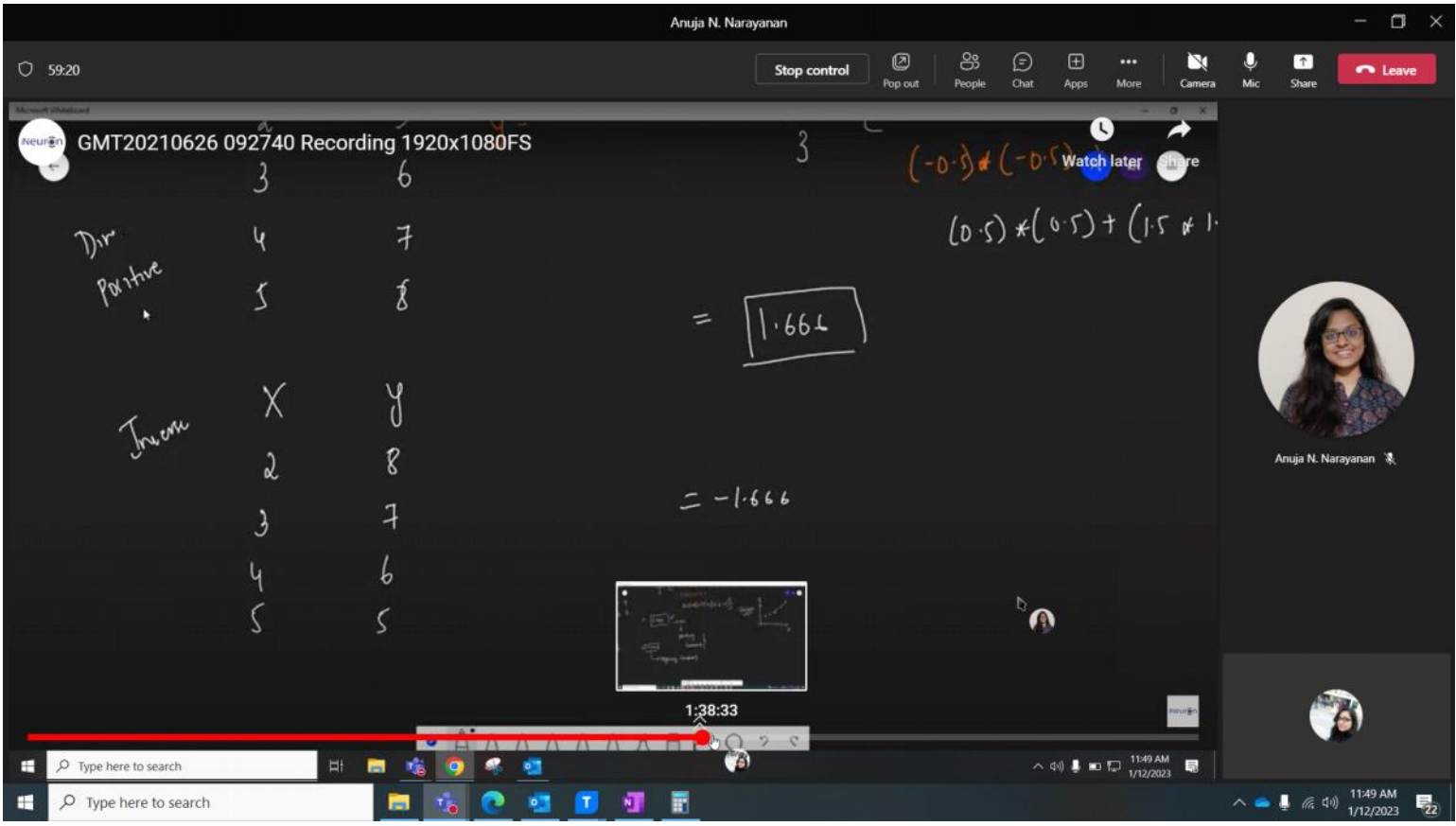

Anuja N. Narayanan
57:40
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
① Covariance
② Pearson Correlation
③ Spearman Rank Correlation
X
Height
Y
Weight
Quantify the relationship between X & Y
Anuja N. Narayanan
Type here to search
11:47 AM
1/12/2023
Type here to search
11:47 AM
1/12/2023

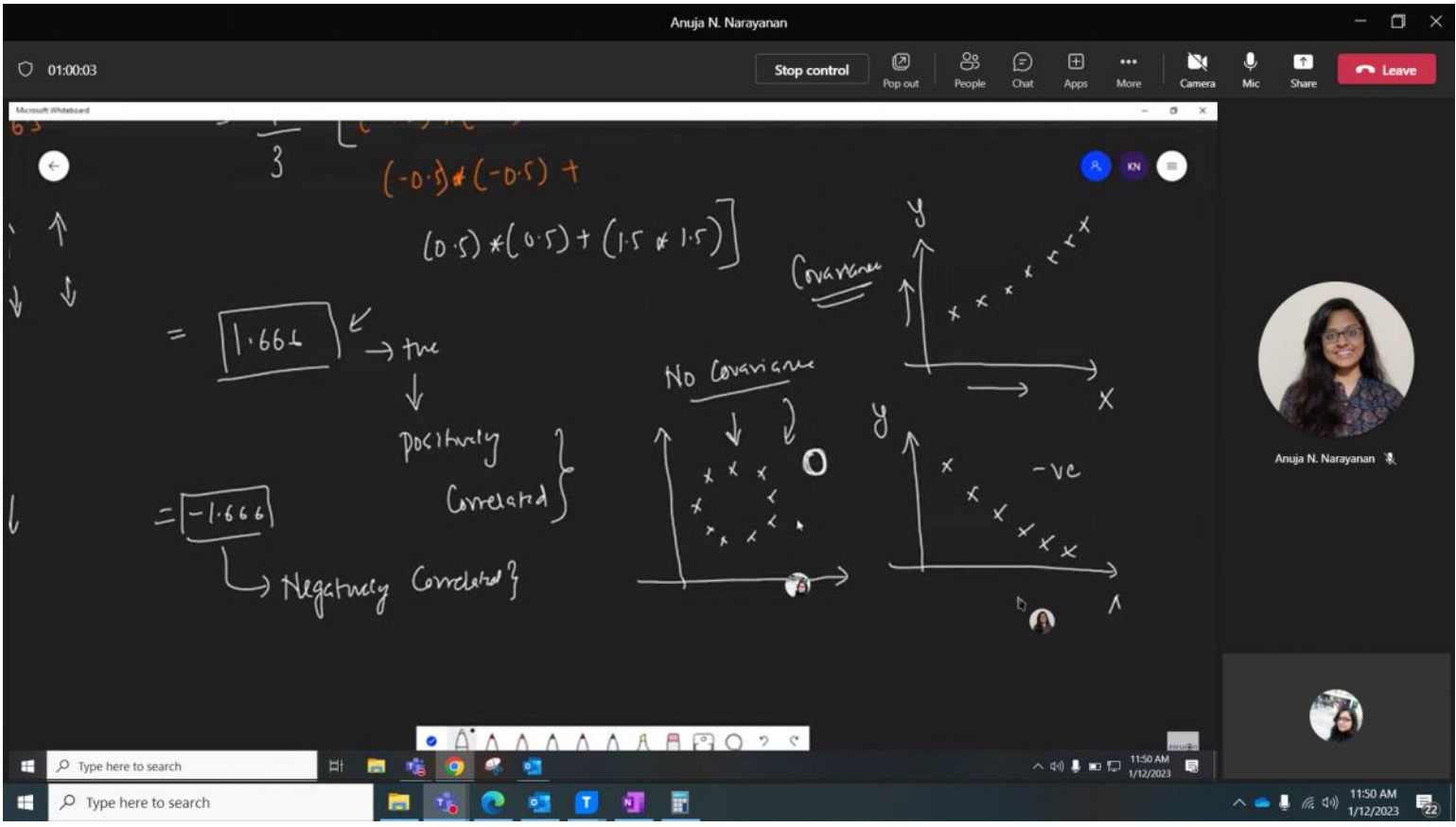

Anuja N. Narayanan
58:14
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
Marketing
Sales
Quantify the relationship between X & Y
Dic
Drinking
Age
Anuja N. Narayanan
Type here to search
11:48 AM
1/12/2023

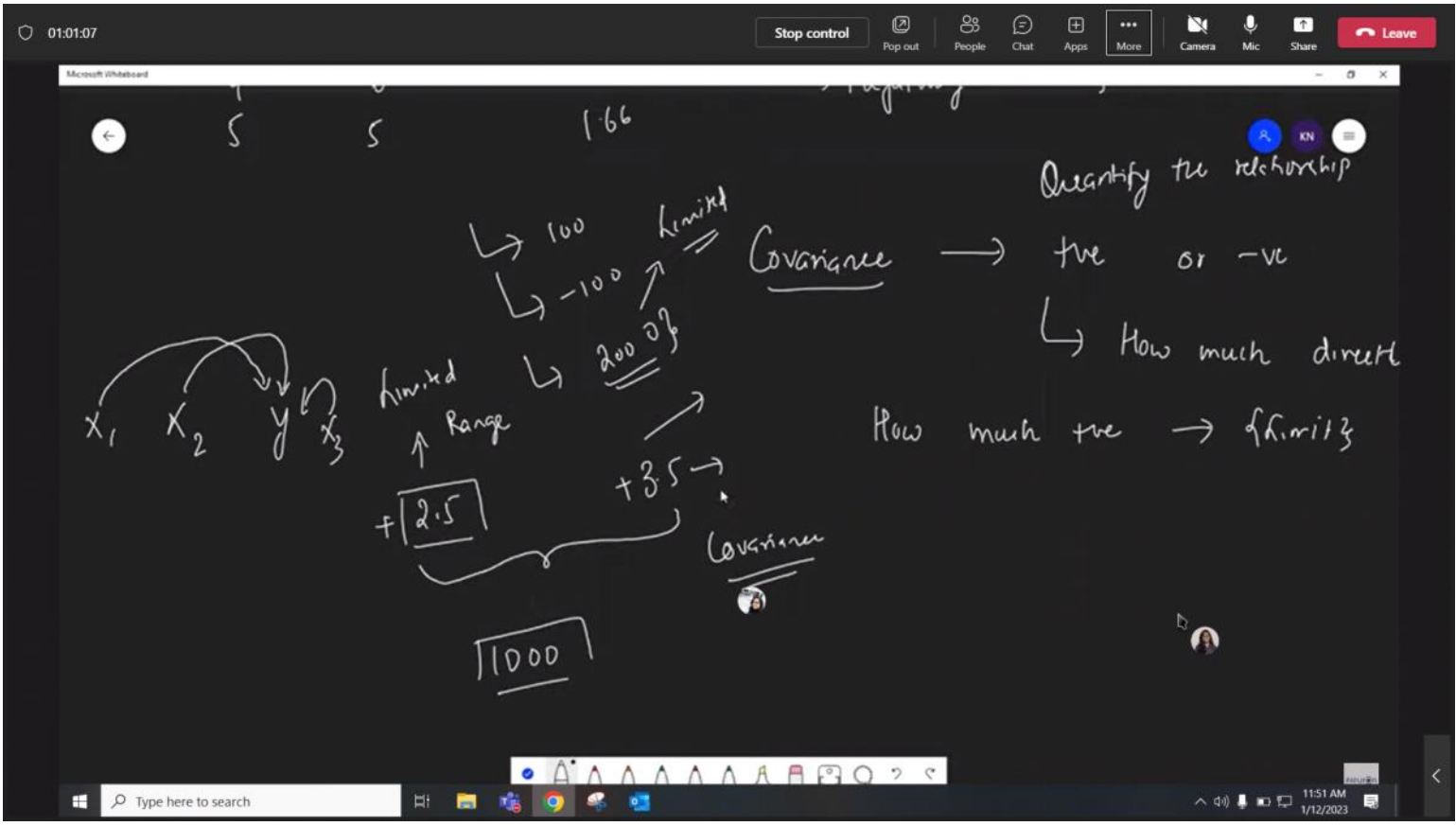

Anuja N. Narayanan
58:27
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
Google
covariance formula
About 1,86,00,000 results (0.47 seconds)
Covariance Formula
Overview
Formula
Calculators
Examples
Practice problems
Share
$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$
$cov_{x,y}$ = covariance between variable a and y
x_i = data value of x
y_i = data value of y
$\bar{x}$ = mean of x
$\bar{y}$ = mean of y
N = number of data values
From the web
Formula for Covariance
X̄ – the mean (average) of the X-variable. Ȳ – the mean (average) of the the number of data points.
https://corporatefinanceinstitute.com › knowledge › finance
Covariance - Definition, Formula, and Practical Example
About
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. Wikipedia
Calculation
Properties
Symbol
Measurements
Feedback
People Also Search For
Copy
Export
KEYWORD
Load Metrics (uses 6 credits)
covariance and correlation formula
1:31:38
Anuja N. Narayanan
Type here to search
11:48 AM
1/12/2023

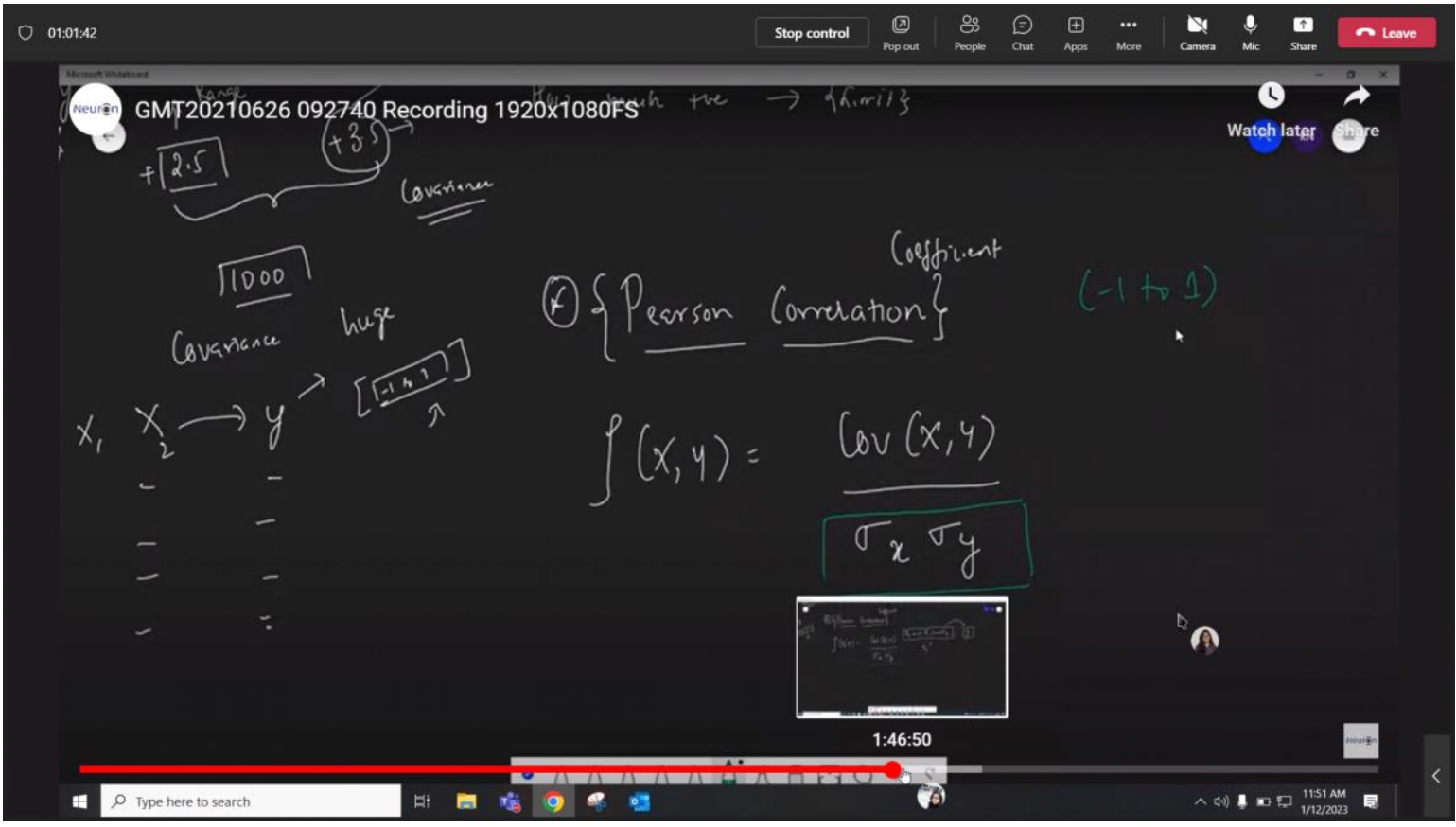

Anuja N. Narayanan
58:55
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
$\bar{x}$ → mean
$\bar{y}$
X y
2 5
3 6
4 7
5 8
$\bar{x} = 3.5$
$\bar{y} = 6.5$
$Cov(X,Y) = \frac{\Sigma (x-\bar{x}) * (y-\bar{y})}{N-1}$
$= \frac{1}{3} [(-1.5) * (-1.5) +$
$(-0.5) * (-0.5) +$
$(0.5) * (0.5) + (1.5 * 1.5)]$
$= 1.66$
Anuja N. Narayanan
Type here to search
11:49 AM
1/12/2023

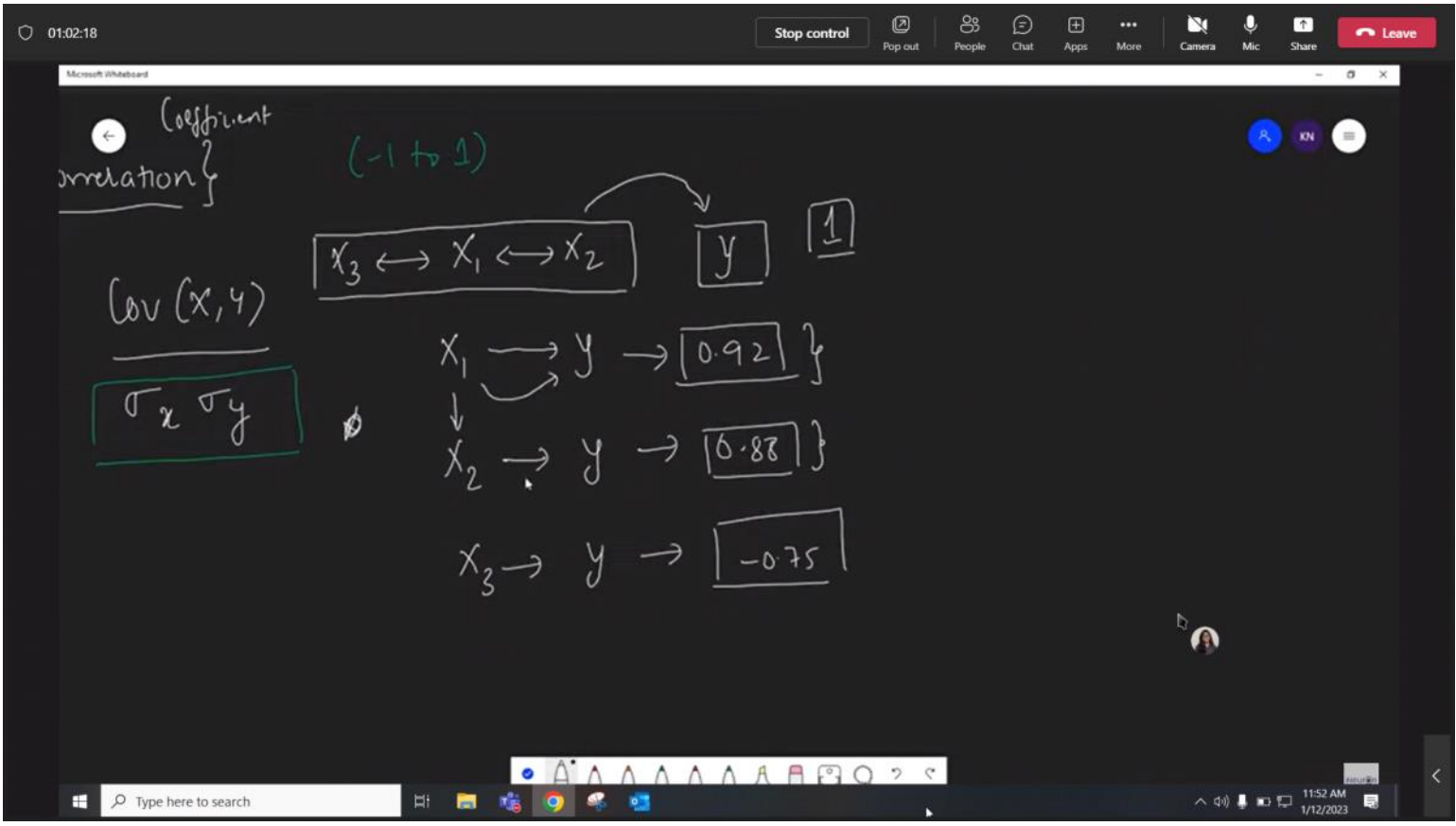

Anuja N. Narayanan
59:10
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
3 6
4 7
5 8
$= 1.66$
X y
2 8
3 7
4 6
5 5
$= -1.666$
Anuja N. Narayanan
Type here to search
11:49 AM
1/12/2023

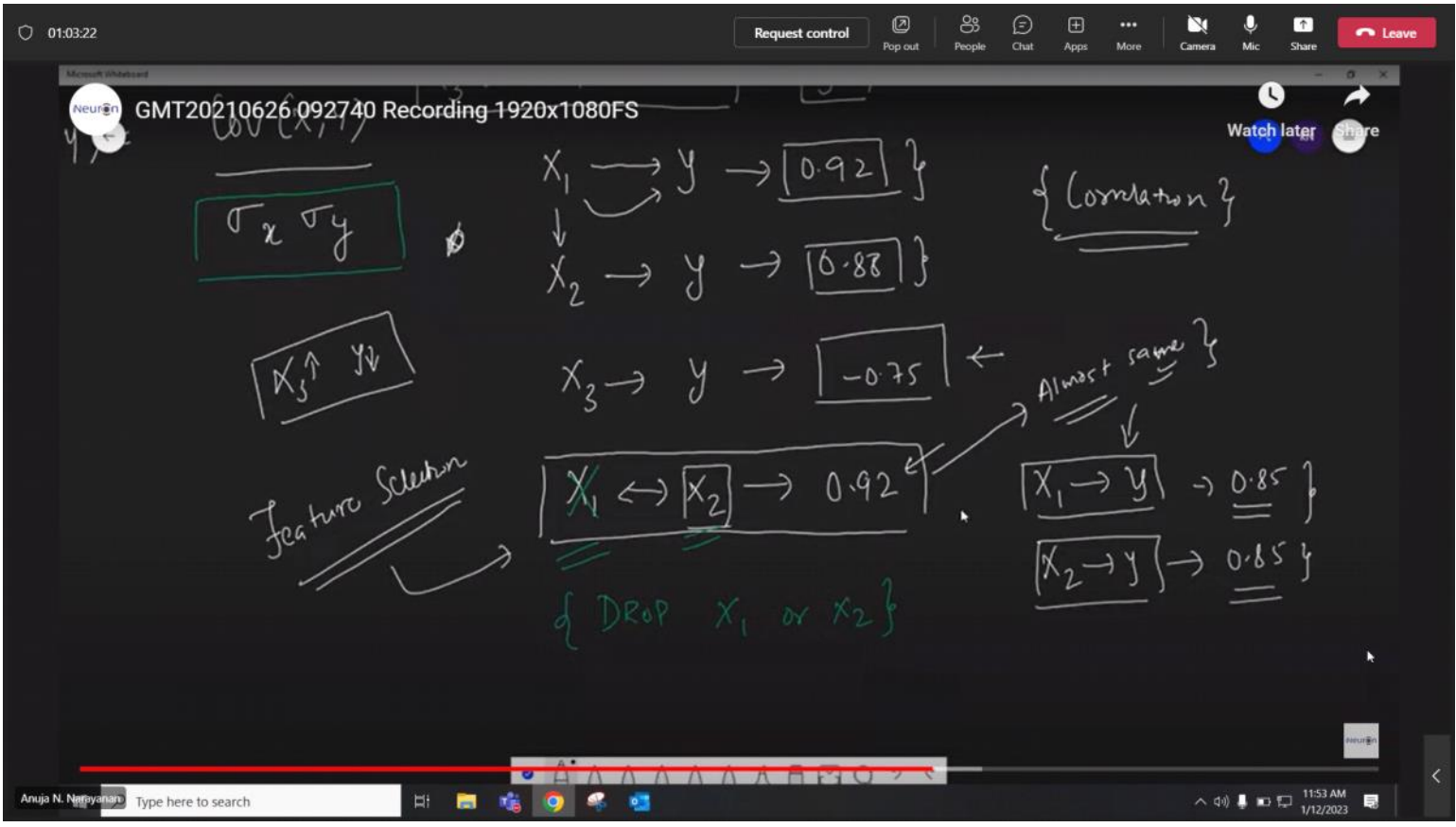

Anuja N. Narayanan
59:20
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
Dir
Positive
Inverse
= 1.661
= -1.666
1:38:33
Anuja N. Narayanan
Type here to search
11:49 AM
1/12/2023

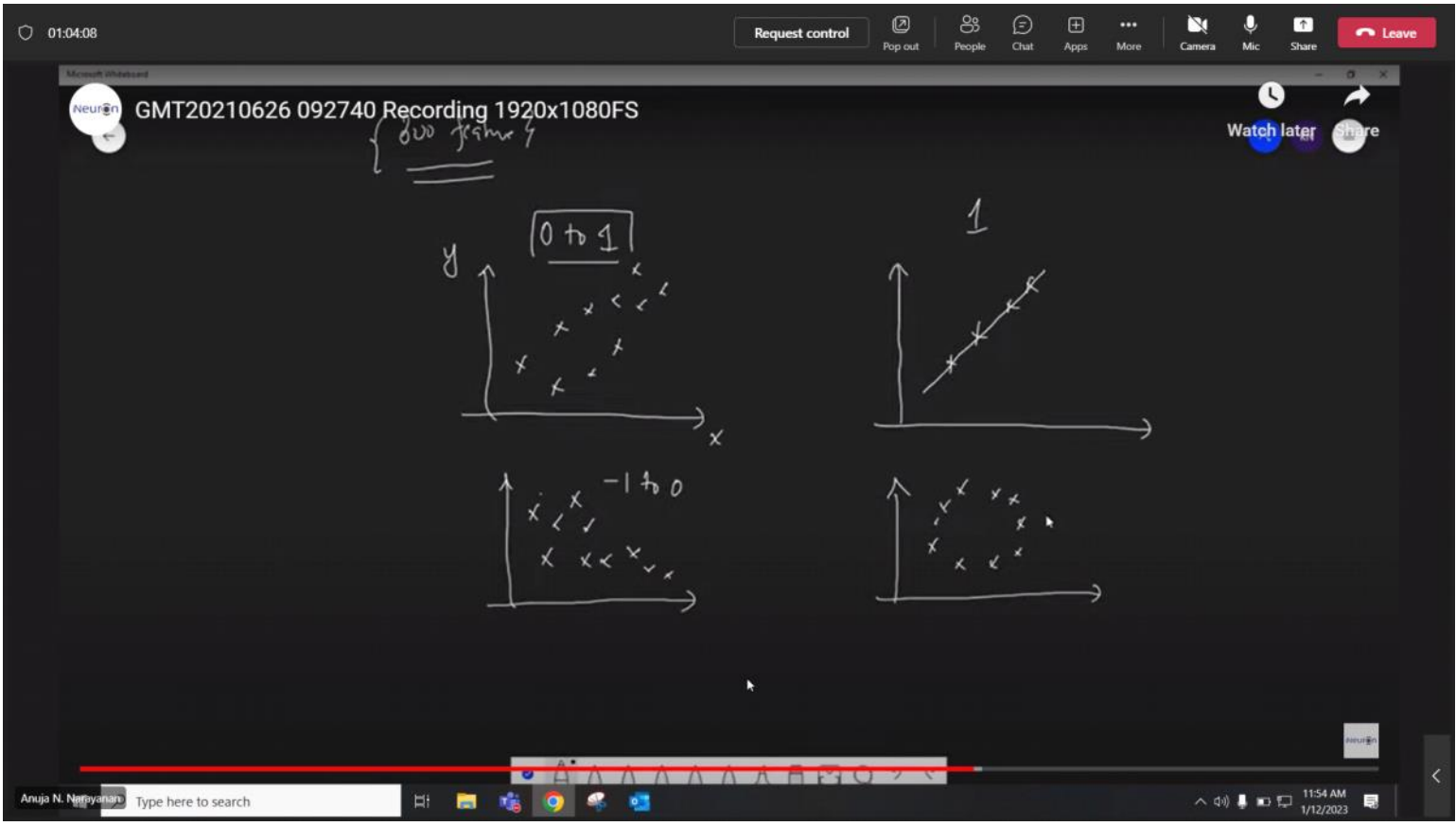

Anuja N. Narayanan
01:00:03
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
= 1.661
→ true
positively
Correlated
= -1.666
→ Negatively Correlated
Covariance
No Covariance
-ve
Anuja N. Narayanan
Type here to search
11:50 AM
1/12/2023

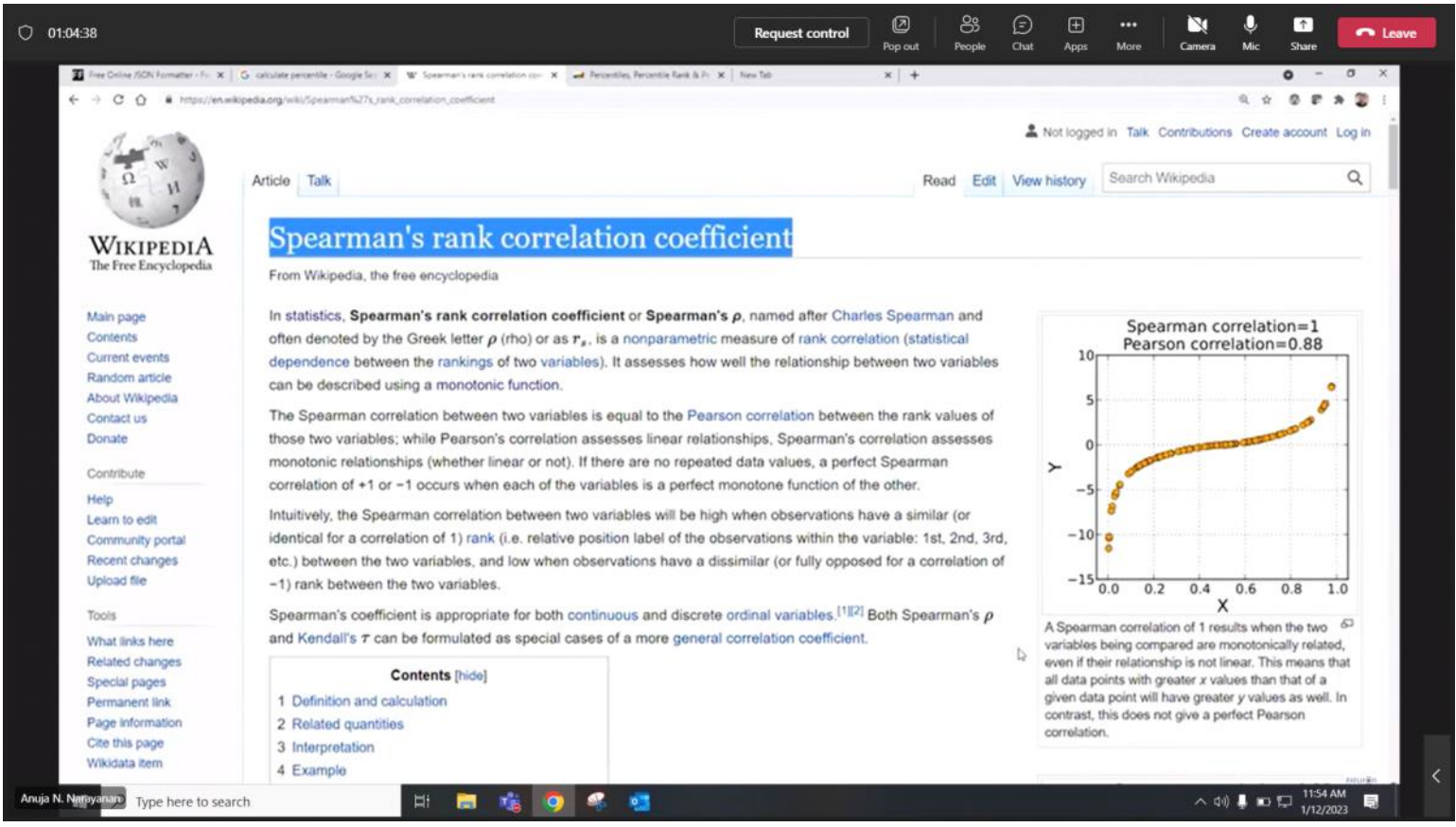

01:01:07
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Quantify the relationship
Covariance → +ve or -ve
How much direction
How much +ve → Limits
Limited
Range
+2.5
+3.5
Covariance
1000
Type here to search
11:51 AM
1/12/2023


01:01:42
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
+2.5
+3.5
Covariance
1000
Covariance huge
Coefficient
Pearson Correlation
(-1 to 1)
Cov(X,Y)
σx σy
1:46:50
Type here to search
11:51 AM
1/12/2023


01:02:18
Stop control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Coefficient
correlation
(-1 to 1)
$X_3 \longleftrightarrow X_1 \longleftrightarrow X_2$
y
1
Cov (X, Y)
$\sigma_x \sigma_y$
$X_1 \longrightarrow y \rightarrow 0.92$
$X_2 \rightarrow y \rightarrow 0.88$
$X_3 \rightarrow y \rightarrow -0.75$
Type here to search
11:52 AM
1/12/2023


01:03:22
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
$\sigma_x \sigma_y$
$X_1 \rightarrow y \rightarrow 0.92$
$X_2 \rightarrow y \rightarrow 0.88$
{ Correlation }
$X_3 \uparrow$ $y \downarrow$
$X_3 \rightarrow y \rightarrow -0.75$
Almost same
Feature Selection
$X_1 \longleftrightarrow X_2 \rightarrow 0.92$
$X_1 \rightarrow y \rightarrow 0.85$
$X_2 \rightarrow y \rightarrow 0.85$
{ DROP X_1 or X_2 }
Anuja N. Narayanan
Type here to search
11:53 AM
1/12/2023


01:04:08
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
0 to 1
1
-1 to 0
Anuja N. Narayanan
Type here to search
11:54 AM
1/12/2023


01:04:38
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Not logged in Talk Contributions Create account Log in
Article Talk
Read Edit View history
Search Wikipedia
WIKIPEDIA
The Free Encyclopedia
Spearman's rank correlation coefficient
From Wikipedia, the free encyclopedia
Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate
Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file
Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item
In statistics, Spearman's rank correlation coefficient or Spearman's ρ, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.
The Spearman correlation between two variables is equal to the Pearson correlation 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.
Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.
Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.[1][2] Both Spearman's ρ and Kendall's τ can be formulated as special cases of a more general correlation coefficient.
Contents [hide]
1 Definition and calculation
2 Related quantities
3 Interpretation
4 Example
Spearman correlation=1
Pearson correlation=0.88
A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well. In contrast, this does not give a perfect Pearson correlation.
Anuja N. Narayanan
Type here to search
11:54 AM
1/12/2023

01:05:06 Request control Pop out People Chat Apps More Camera Mic Share Leave

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

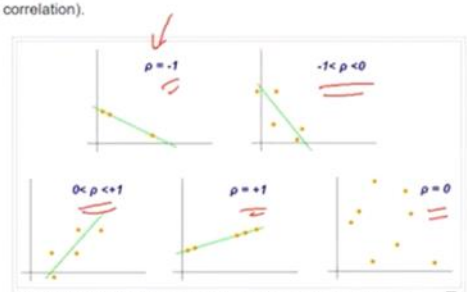
Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

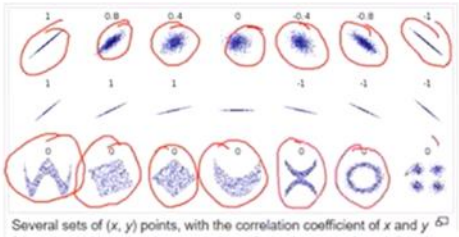
Languages
العربية
Deutsch
Español
فارسی
한국어
Italiano
日本語
Português

PPMCC, the **bivariate correlation**,[1] or colloquially simply as **the correlation coefficient**[2]) is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Contents [hide]



Examples of scatter diagrams with different values of correlation coefficient (ρ)



Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the strength and direction

Anuja N. Narayanan Type here to search 11:55 AM 1/12/2023

01:06:23 Request control Pop out People Chat Apps More Camera Mic Share Leave

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Spearman's rank correlation coefficient

From Wikipedia, the free encyclopedia

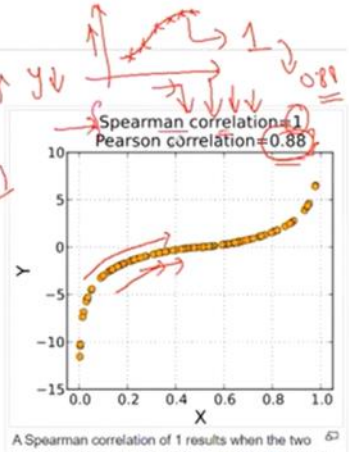
In statistics, **Spearman's rank correlation coefficient** or **Spearman's ρ**, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

The Spearman correlation between two variables is equal to the Pearson correlation 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.

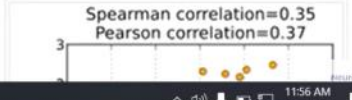
Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.[1][2] Both Spearman's ρ and Kendall's τ can be formulated as special cases of a more general correlation coefficient.

Contents [hide]



A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well. In contrast, this does not give a perfect Pearson correlation.



Anuja N. Narayanan Type here to search 11:56 AM 1/12/2023

01:18:43

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

X	y	r_x	r_y
5	100	2	4
4	500	3	2
7	200	1	3
3	1000	4	1

$$\rho_{r(x,y)} = \frac{Cov(r_x, r_y)}{\sigma_{r_x}\,\sigma_{r_y}}$$

Anuja N. Narayanan | Type here to search | 12:08 PM 1/12/2023

01:19:46

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Firstly, evaluate d_i^2. To do so use the following steps, reflected in the table below.

1. Sort the data by the first column (X_i). Create a new column x_i and assign it the ranked values 1, 2, 3, ..., n.
2. Next, sort the data by the second column (Y_i). Create a fourth column y_i and similarly assign it the ranked values 1, 2, 3, ..., n.
3. Create a fifth column d_i to hold the differences between the two rank columns (x_i and y_i).
4. Create one final column d_i^2 to hold the value of column d_i squared.

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
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

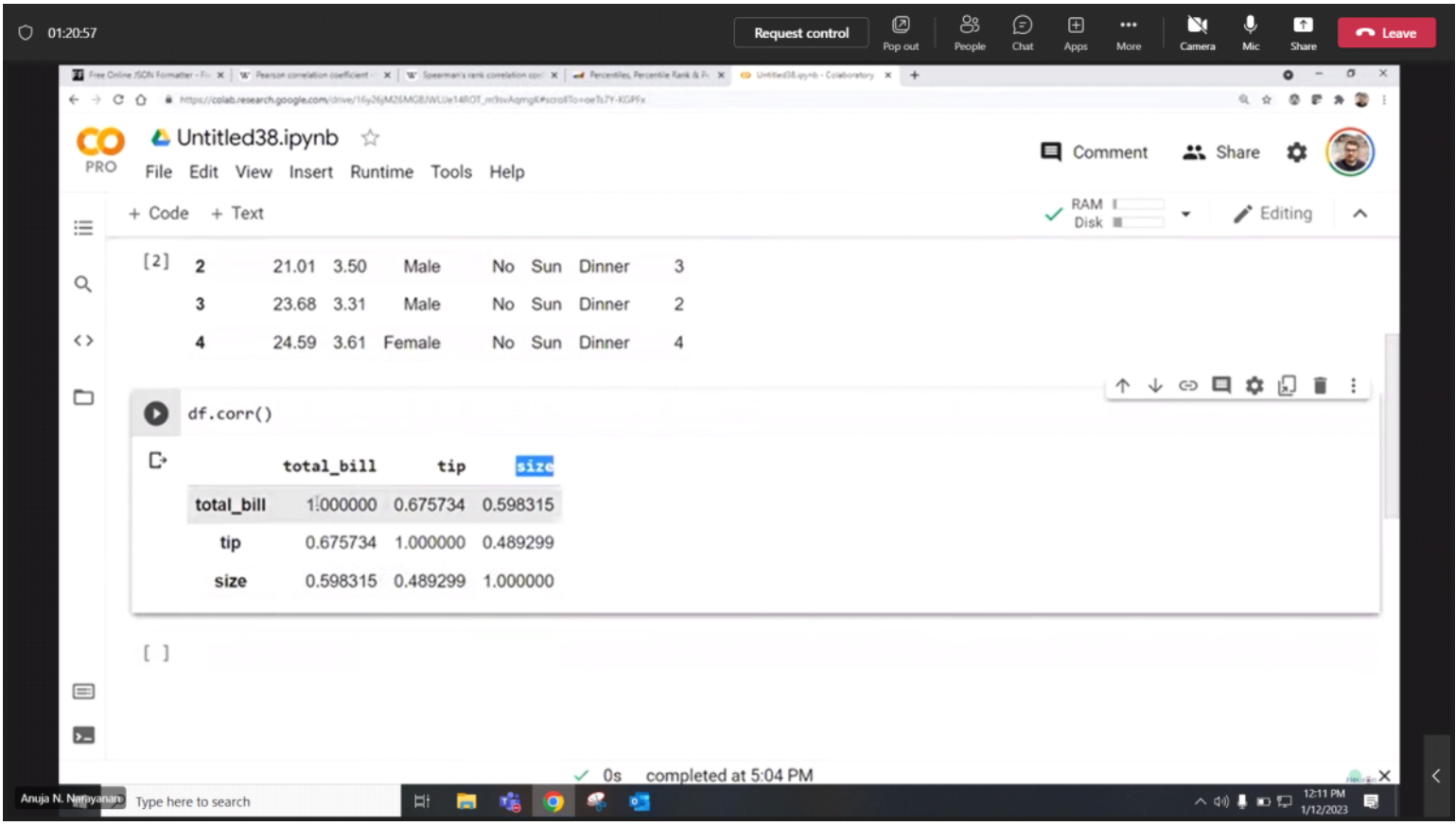
With d_i^2 found, add them to find $\sum d_i^2 = 194$. The value of n is 10. These values can now be substituted back into the equation

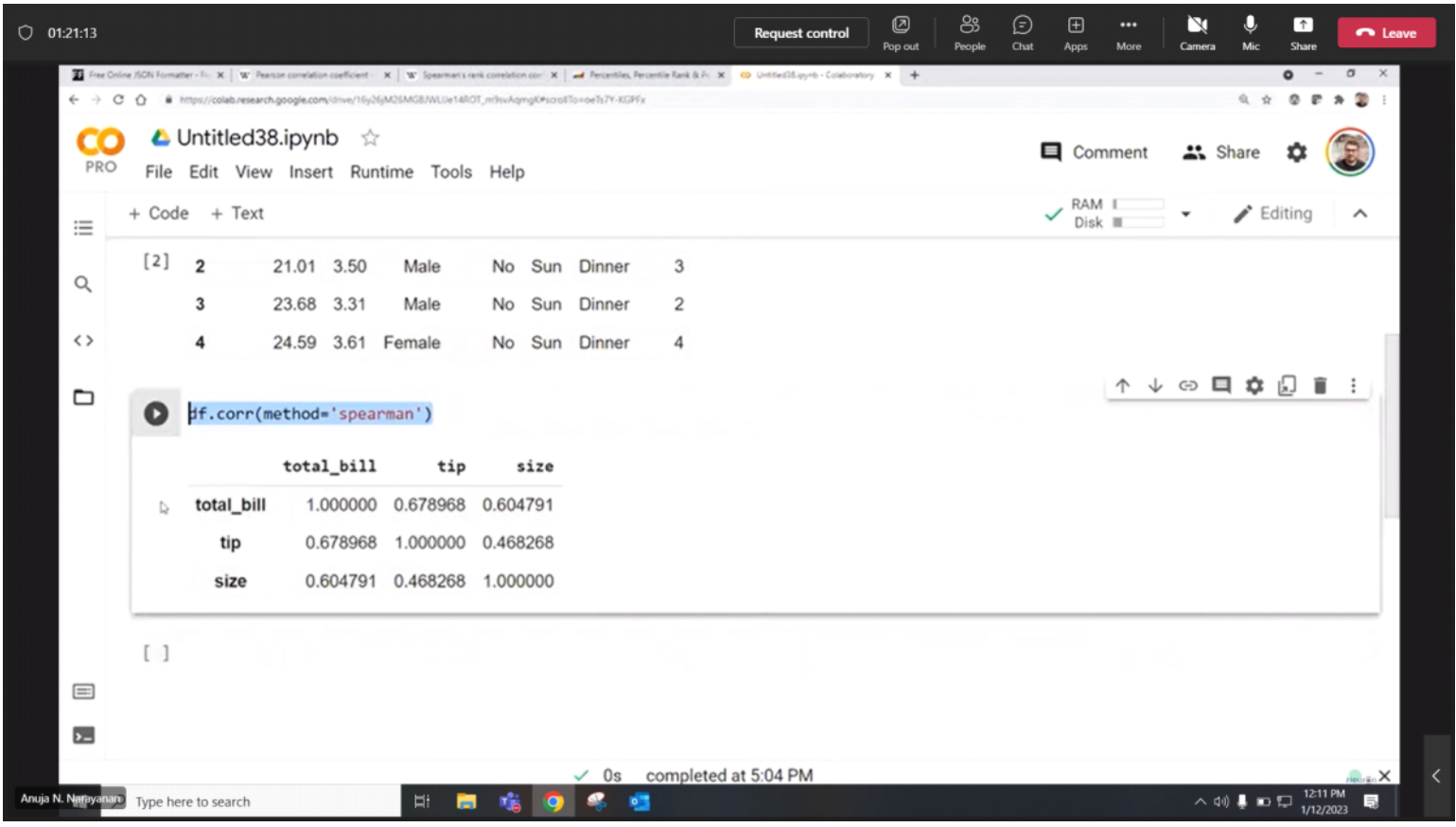
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

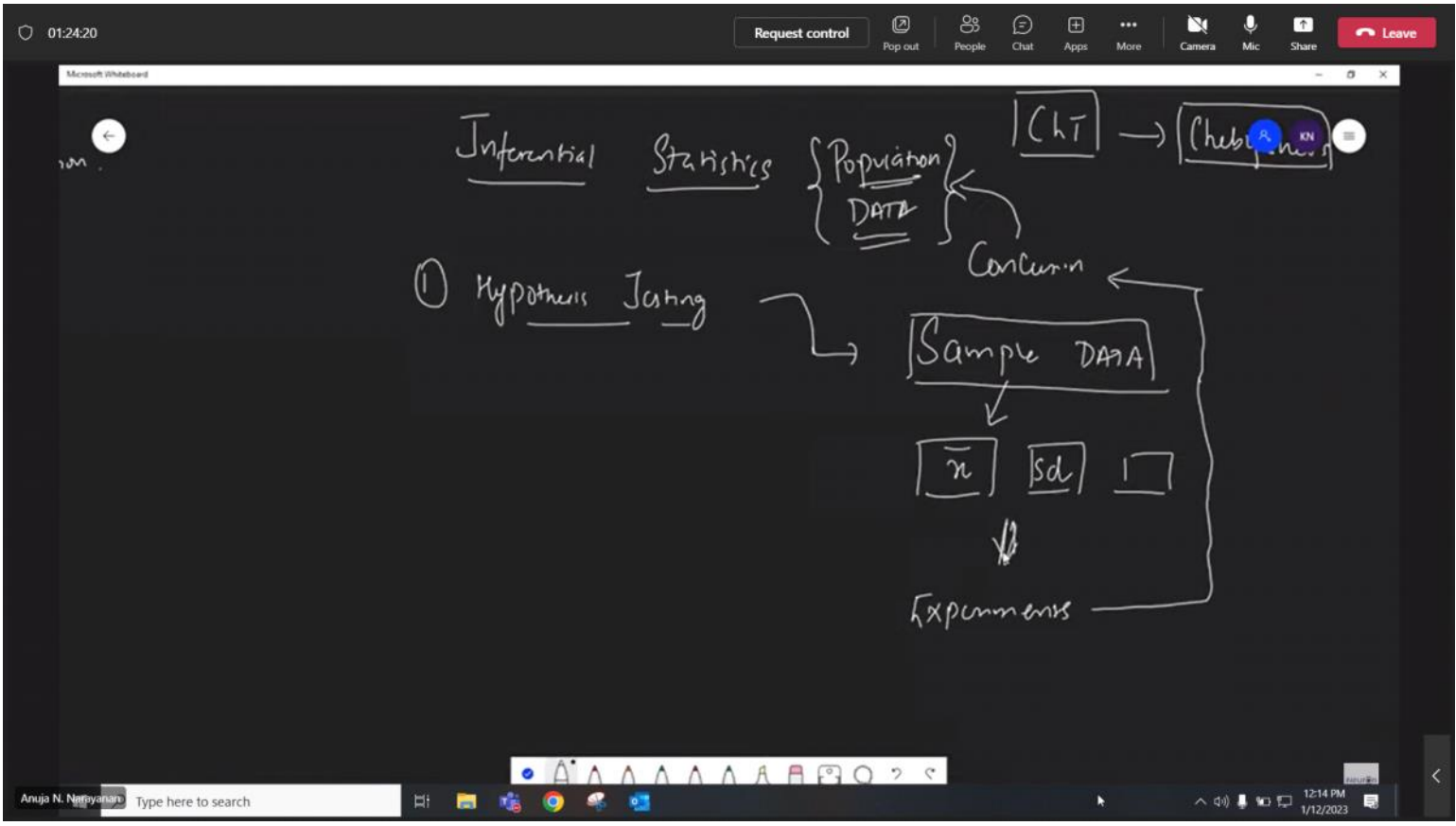
to give

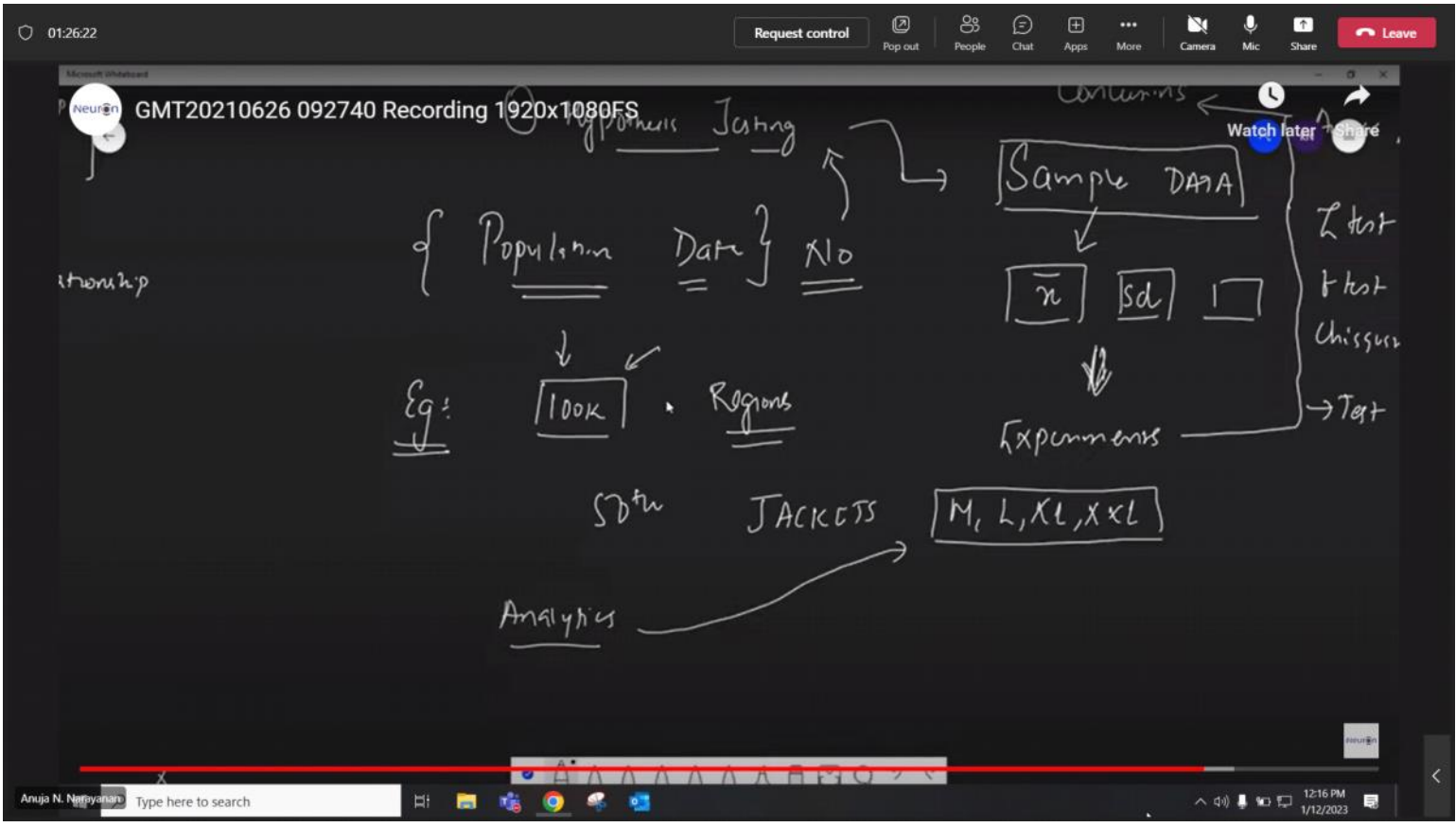
$$\rho = 1 - \frac{6 \times 194}{}$$

Anuja N. Narayanan | Type here to search | 12:09 PM 1/12/2023


01:20:57
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Untitled38.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
RAM
Disk
Editing
[2]
2 21.01 3.50 Male No Sun Dinner 3
3 23.68 3.31 Male No Sun Dinner 2
4 24.59 3.61 Female No Sun Dinner 4
df.corr()
total_bill tip size
total_bill 1.000000 0.675734 0.598315
tip 0.675734 1.000000 0.489299
size 0.598315 0.489299 1.000000
[ ]
0s completed at 5:04 PM
Anuja N. Narayanan
Type here to search
12:11 PM
1/12/2023


01:21:13
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Untitled38.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
RAM
Disk
Editing
[2]
2 21.01 3.50 Male No Sun Dinner 3
3 23.68 3.31 Male No Sun Dinner 2
4 24.59 3.61 Female No Sun Dinner 4
df.corr(method='spearman')
total_bill tip size
total_bill 1.000000 0.678968 0.604791
tip 0.678968 1.000000 0.468268
size 0.604791 0.468268 1.000000
[ ]
0s completed at 5:04 PM
Anuja N. Narayanan
Type here to search
12:11 PM
1/12/2023


01:24:20
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Inferential Statistics
Population
DATA
CHT
Hypothesis Testing
Sample DATA
Conclusion
Experiments
Anuja N. Narayanan
Type here to search
12:14 PM
1/12/2023


01:26:22
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
Hypothesis Testing
Sample DATA
Population
Data
No
Eg:
100K
Regions
Experiments
Conclusion
Z test
t test
Chi square
Test
50th
JACKETS
M, L, XL, XXL
Analysis
Anuja N. Narayanan
Type here to search
12:16 PM
1/12/2023

01:30:04

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

Central limit theorem

$n \geq 30$

$X \sim G.D(\mu, \sigma)$

$n \geq 30$

Central limit theorem

$S_1 \rightarrow x_1, x_2 \ldots x_n = \bar{x}_1$

$S_2 \rightarrow \ldots\ldots x_n = \bar{x}_2$

$S_3 \rightarrow$

$\{\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots$

pdf → Gauss

Anuja N. Narayanan

Type here to search

27°C Rain | ENG IN | 26-06-2021

01:32:45

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

$n \geq 30$

Central limit theorem

$\ldots x_n = \bar{x}_1$

$\{\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_m\}$

$\ldots x_n = \bar{x}_2$

$\left(\mu, \frac{\sigma^2}{n}\right)$

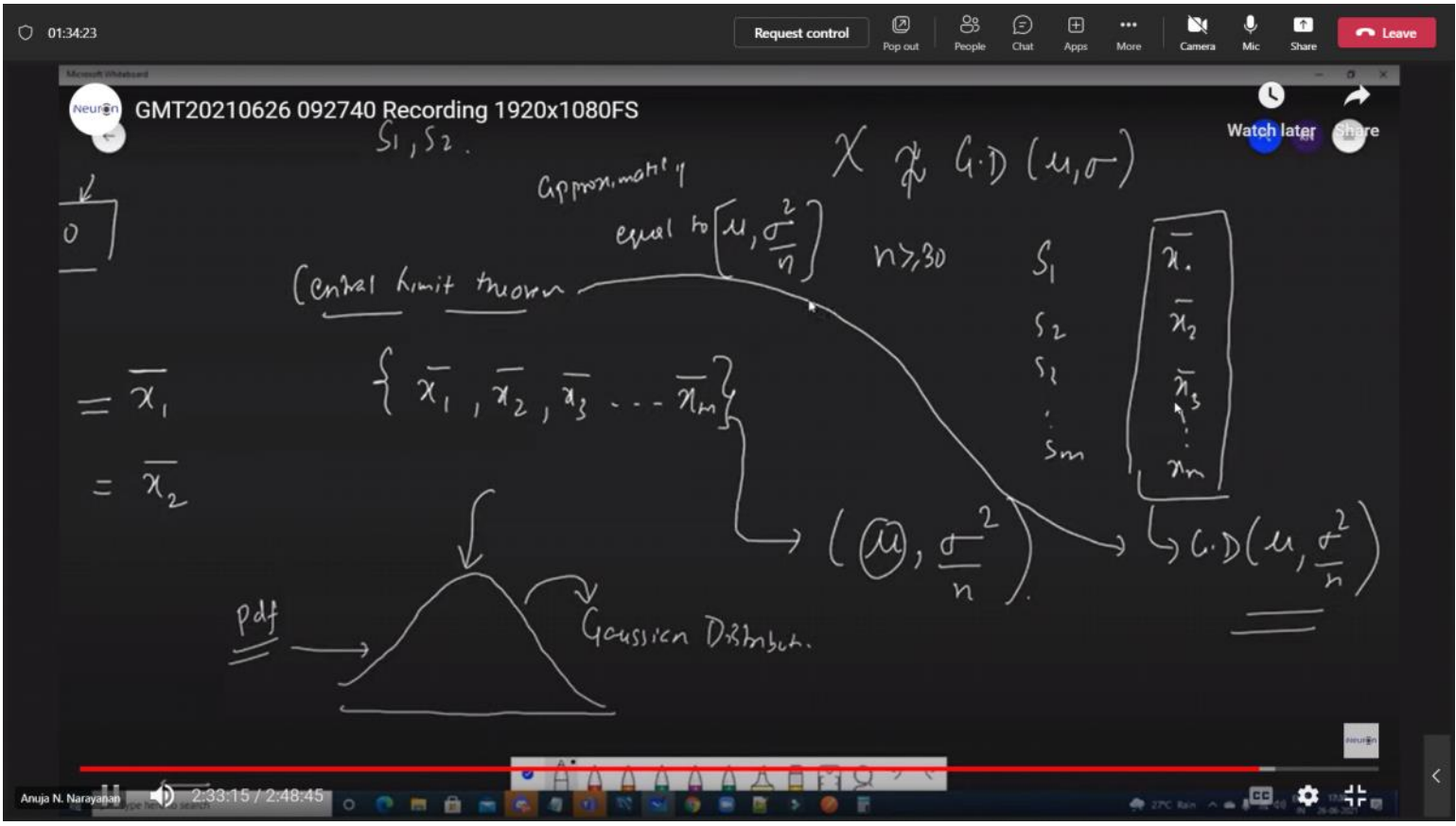
pdf → Gaussian Distribution

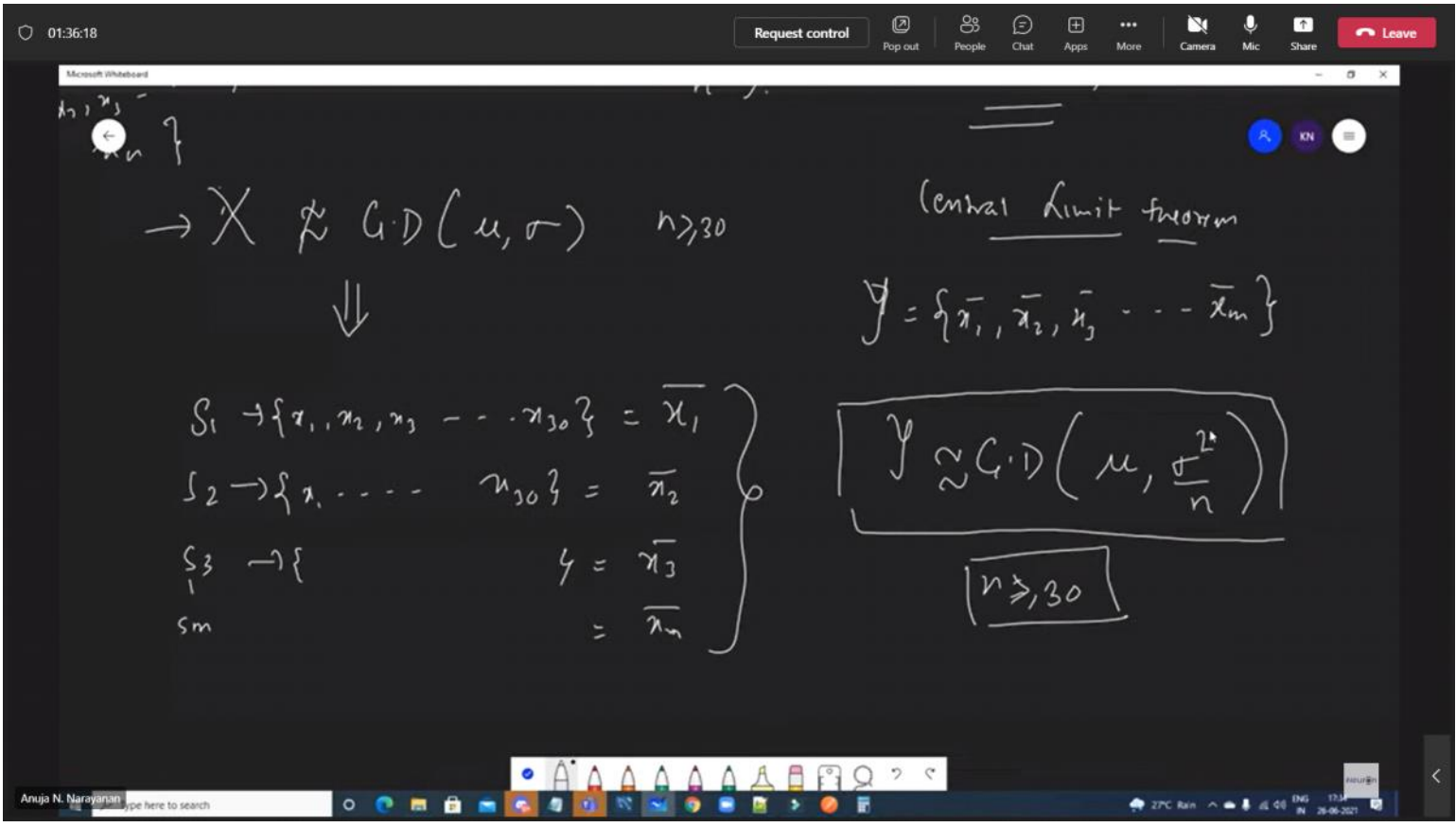
$\ldots x_n = \bar{x}_m$

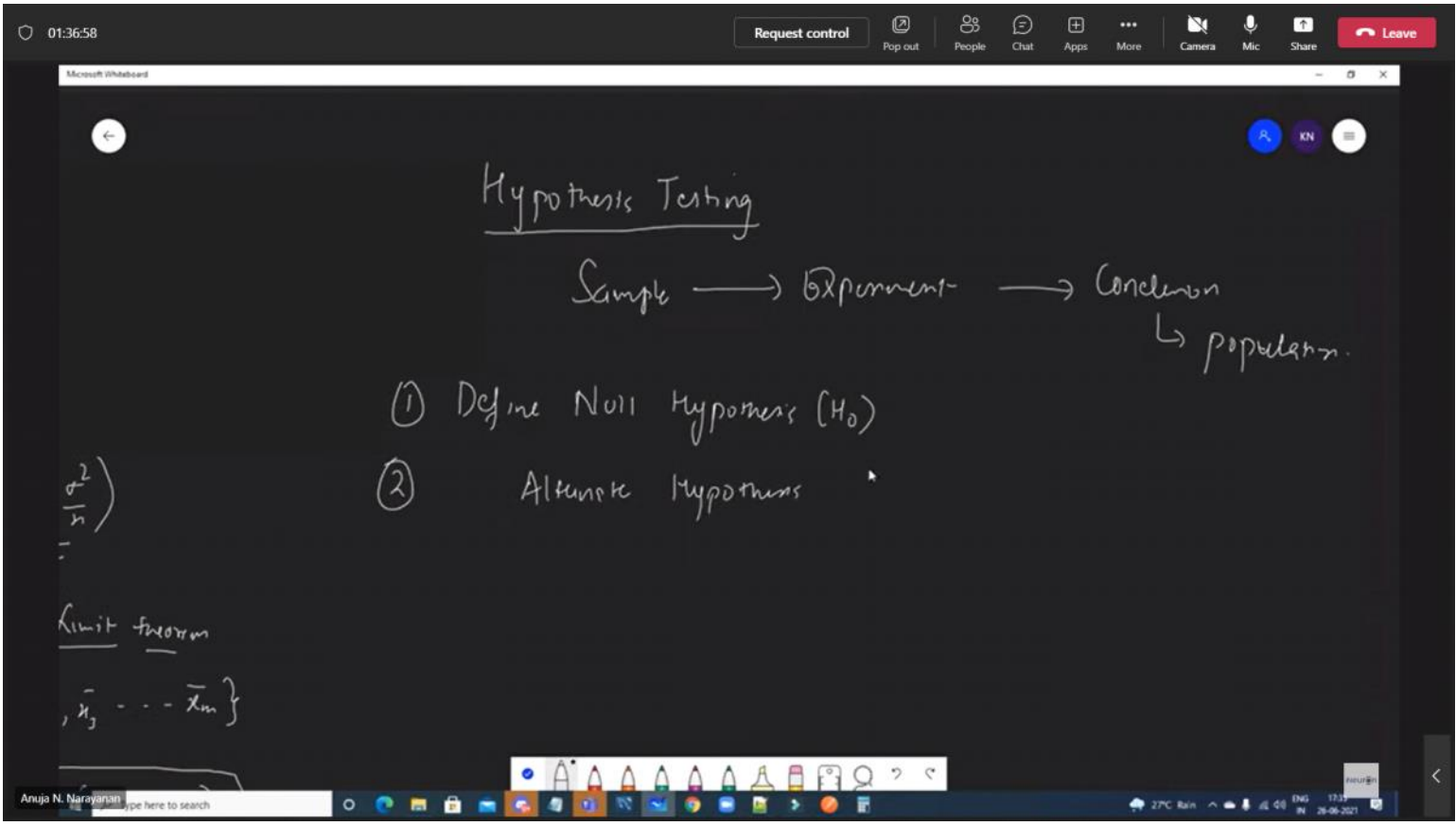
Anuja N. Narayanan

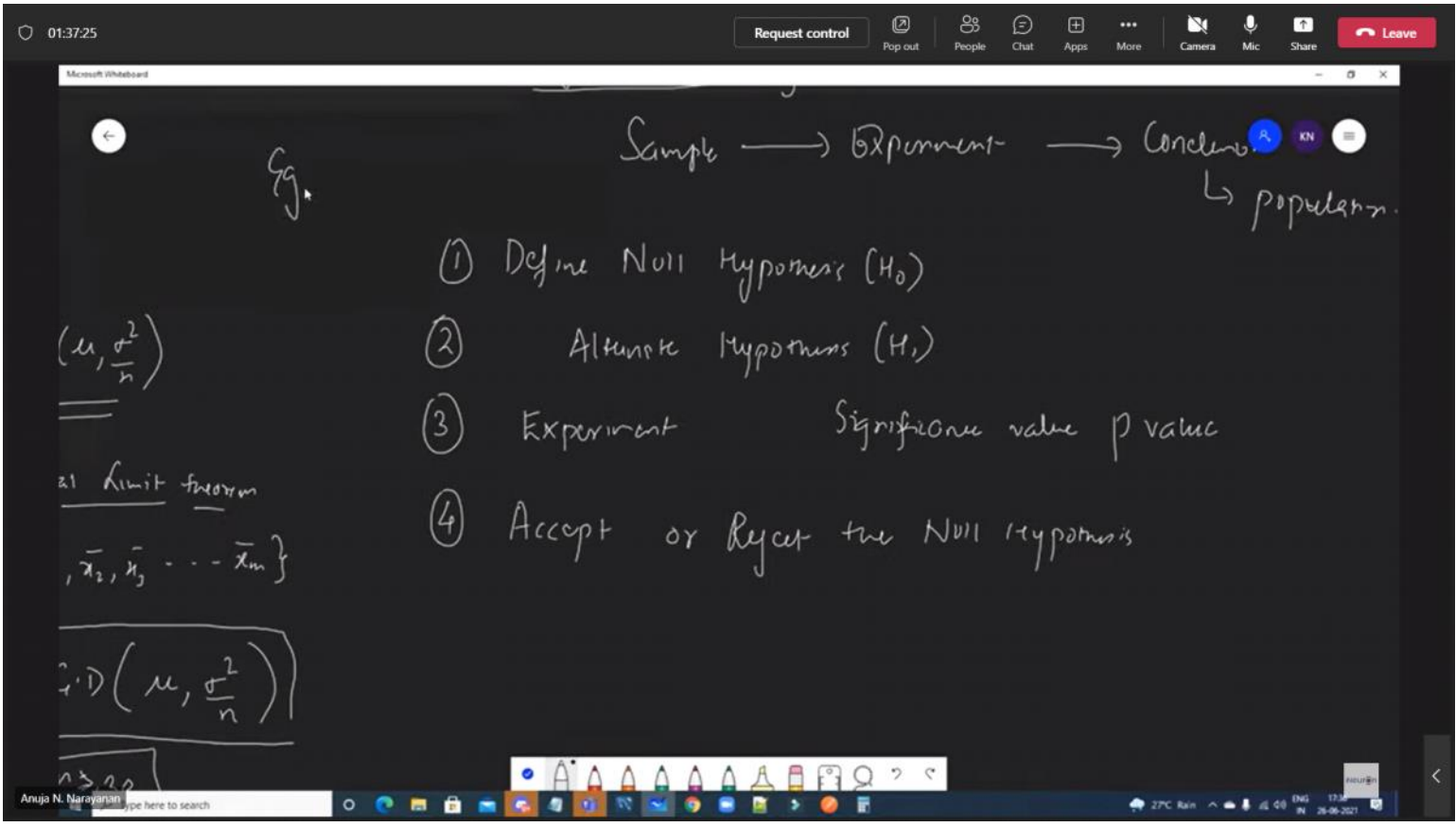
Type here to search

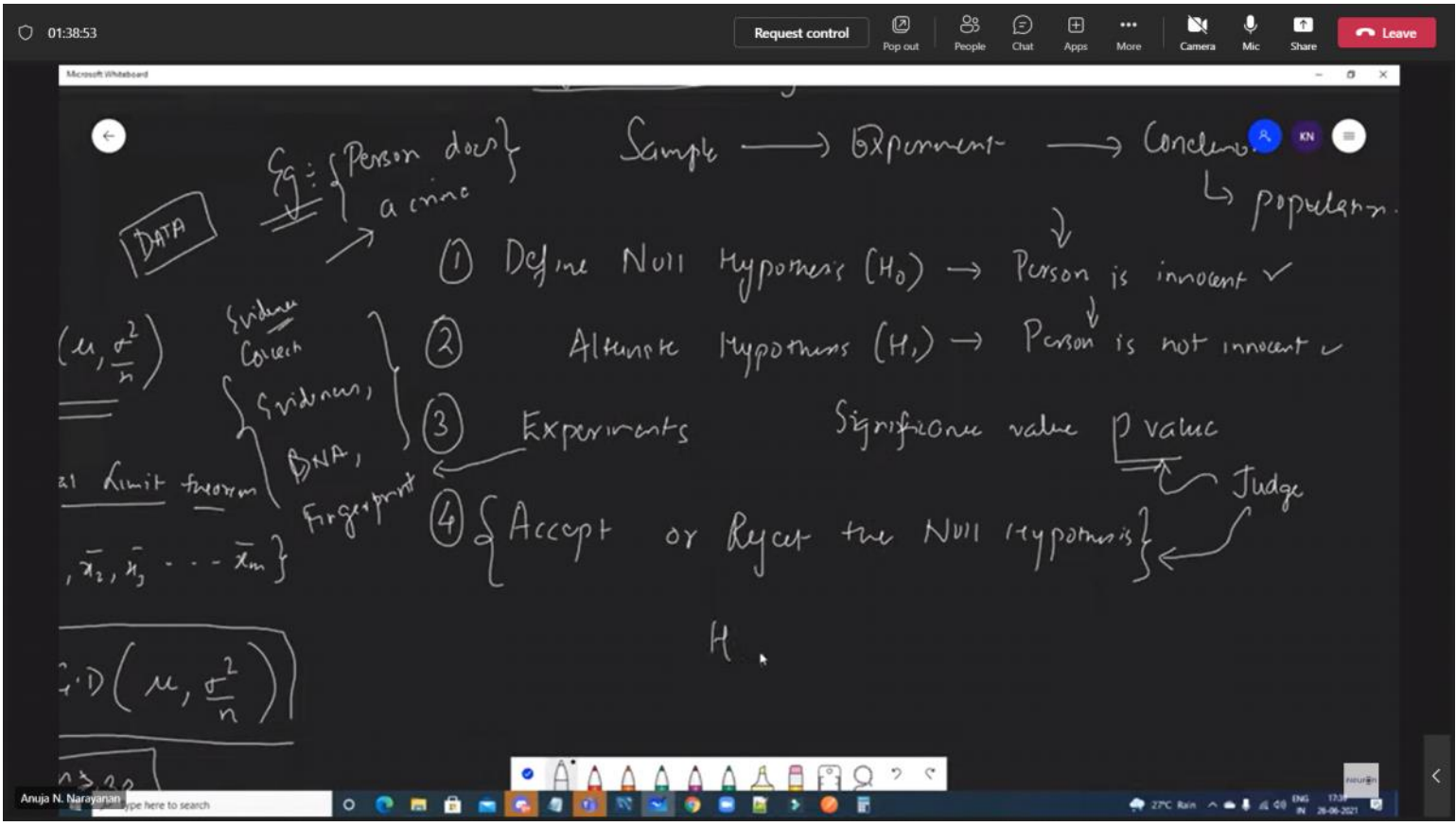
27°C Rain | ENG IN | 26-06-2021


01:34:23
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210626 092740 Recording 1920x1080FS
Watch later
Share
S_1, S_2.
$X \approx G.D(\mu, \sigma)$
Approximately
equal to $\left[\mu, \frac{\sigma^2}{n}\right]$
$n \geq 30$
(central limit theorem
S_1
S_2
S_3
S_m
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
$\bar{x}_m$
$= \bar{x}_1$
$= \bar{x}_2$
$\{\bar{x}_1, \bar{x}_2, \bar{x}_3 \cdots \bar{x}_m\}$
$\left(\mu, \frac{\sigma^2}{n}\right)$
$G.D\left(\mu, \frac{\sigma^2}{n}\right)$
pdf
Gaussian Distribution
Anuja N. Narayanan
2:33:15 / 2:48:45


01:36:18
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
$\rightarrow X \approx G.D(\mu, \sigma) \quad n \geq 30$
(central limit theorem
$\bar{y} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3 \cdots \bar{x}_m\}$
$S_1 \rightarrow \{x_1, x_2, x_3 \cdots x_{30}\} = \bar{x}_1$
$S_2 \rightarrow \{x_1 \cdots x_{30}\} = \bar{x}_2$
$S_3 \rightarrow \{ \quad = \bar{x}_3$
$S_m \quad = \bar{x}_m$
$\bar{y} \approx G.D\left(\mu, \frac{\sigma^2}{n}\right)$
$n \geq 30$
Anuja N. Narayanan


01:36:58
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Hypothesis Testing
Sample → Experiment → Conclusion
↳ population.
(1) Define Null Hypothesis (H_0)
(2) Alternate Hypothesis
Limit theorem
Anuja N. Narayanan


01:37:25
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Sample → Experiment → Conclusion
↳ population.
Eg.
(1) Define Null Hypothesis (H_0)
(2) Alternate Hypothesis (H_1)
(3) Experiment Significance value p value
(4) Accept or Reject the Null Hypothesis
Anuja N. Narayanan


01:38:53
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
DATA
Eg: {Person does a crime}
Sample → Experiment → Conclusion
↳ population
(1) Define Null Hypothesis (H_0) → Person is innocent
(2) Alternate Hypothesis (H_1) → Person is not innocent
(3) Experiments
Evidence Collection
Evidences, DNA, Fingerprint
Significance value
p value
Judge
(4) {Accept or Reject the Null Hypothesis}
Anuja N. Narayanan

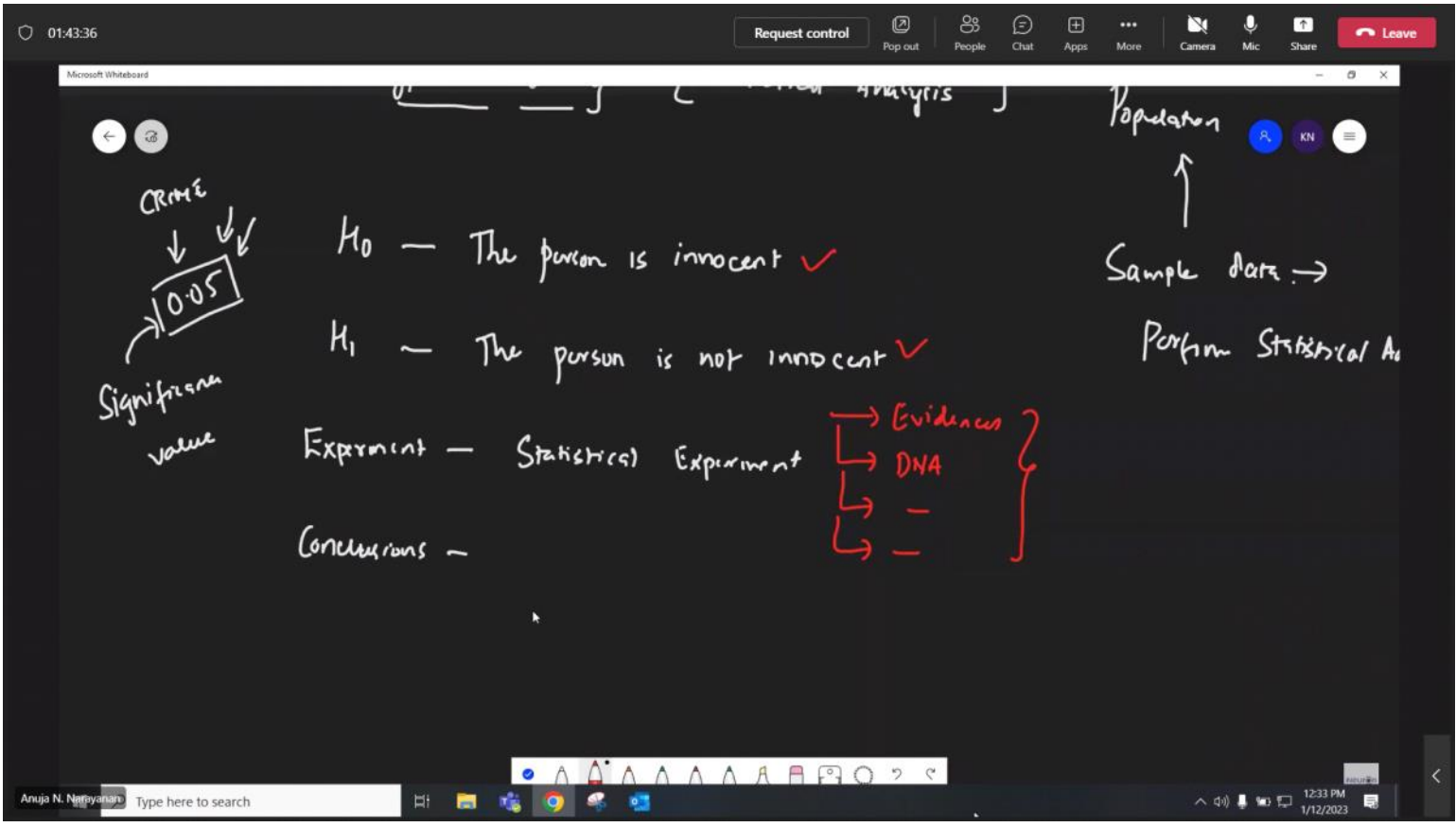
TO DO:

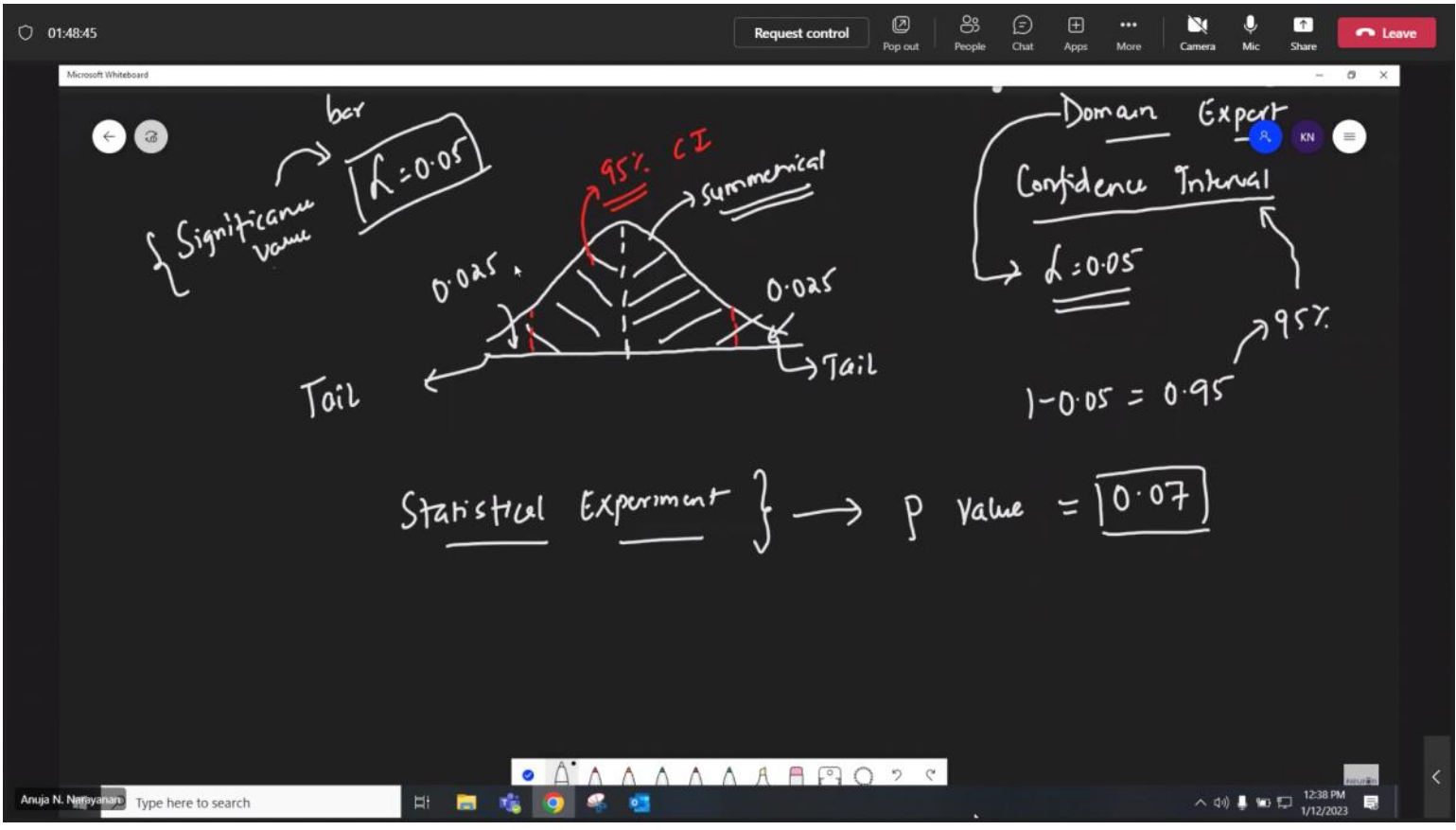
Thursday, January 12, 2023 12:06 PM

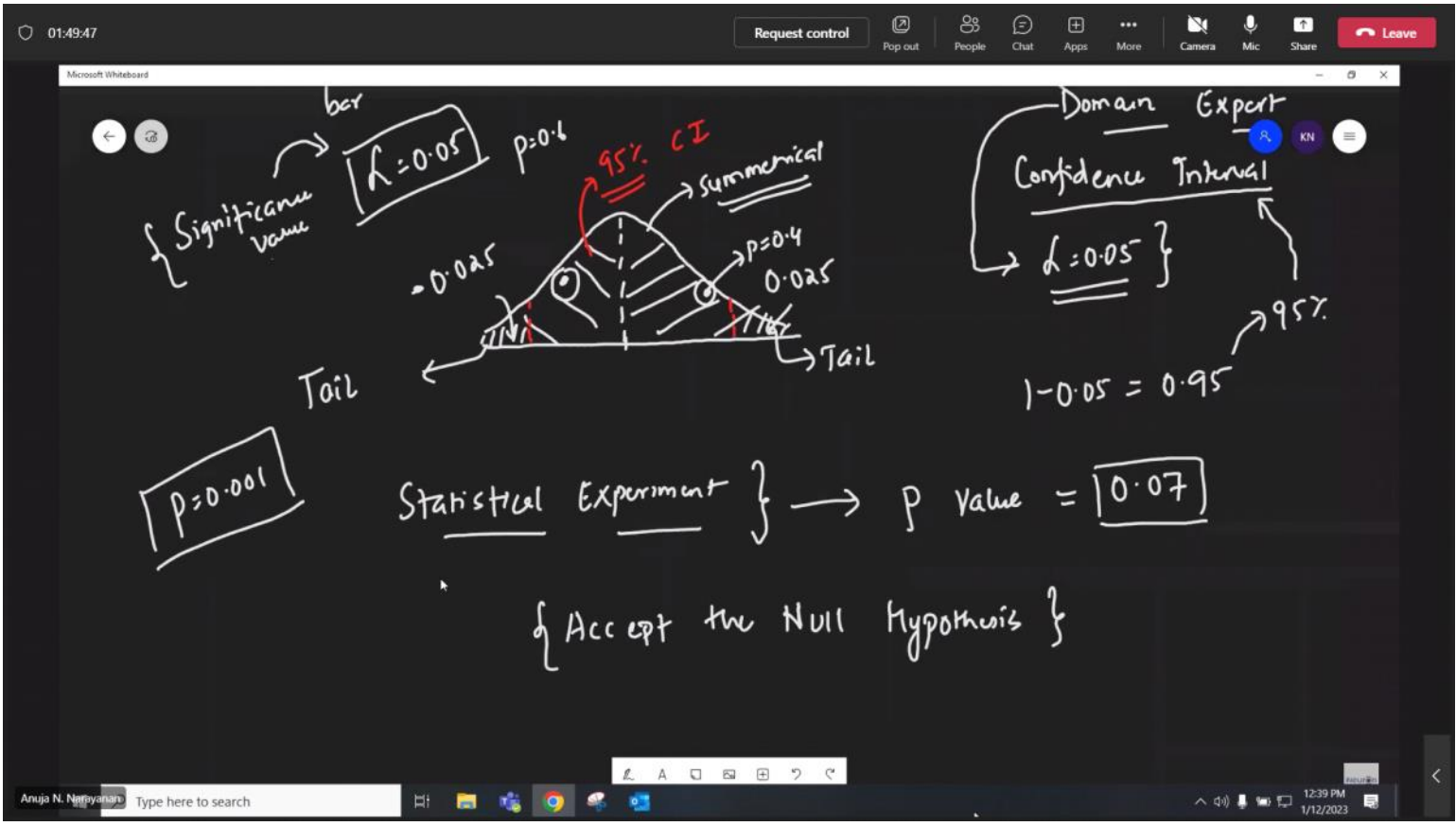
1. GO THROUGH INTERVIEW QUENTION FROM MONDAY
2. SEARCH FOR STATISTICS INTERVIEW QUESTION
3. SOLVED LINKEDIN STATISTICS QUESTION
4. SOLVE KN ASSIGNMENTS

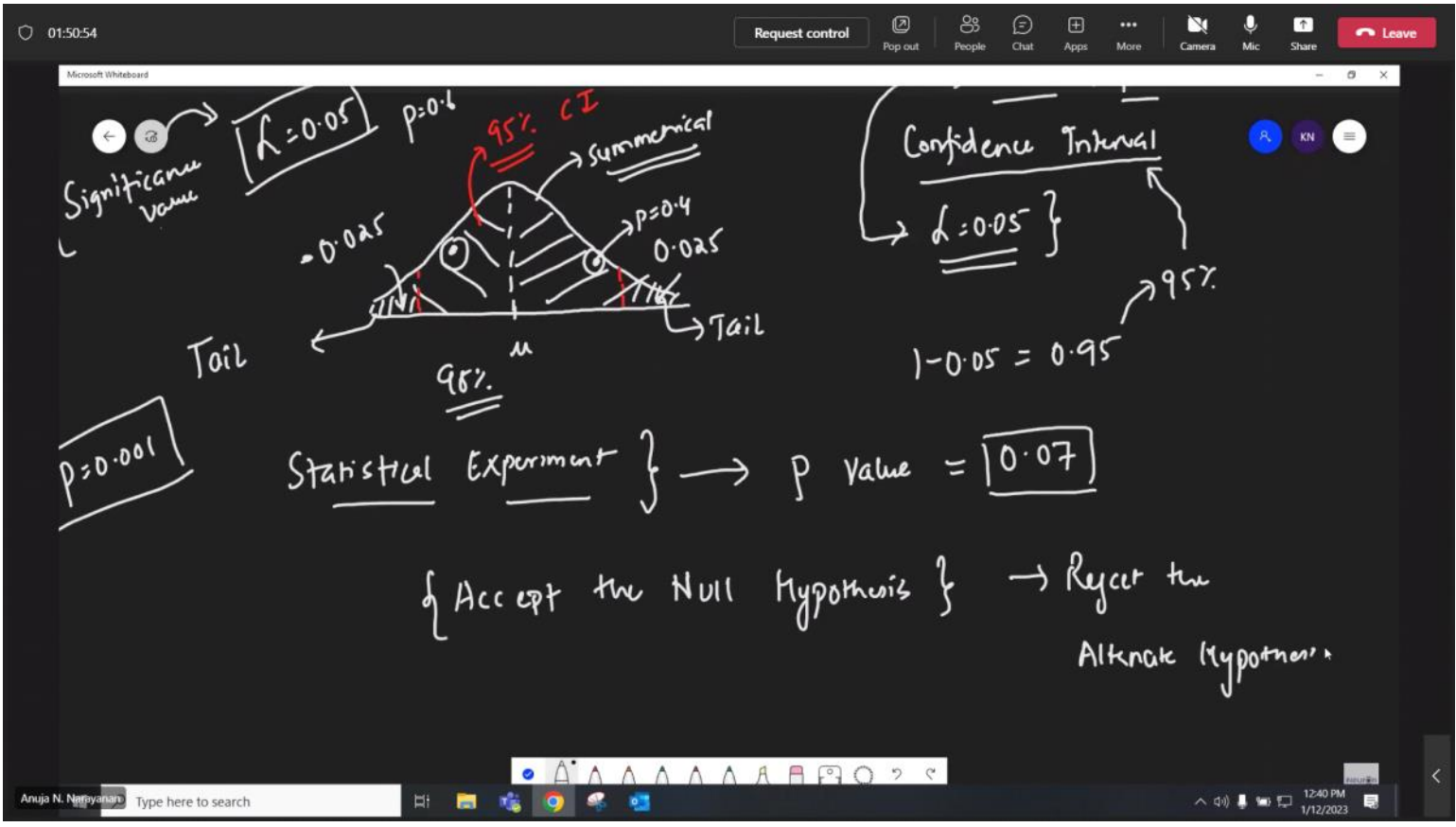
KN(lec-4)

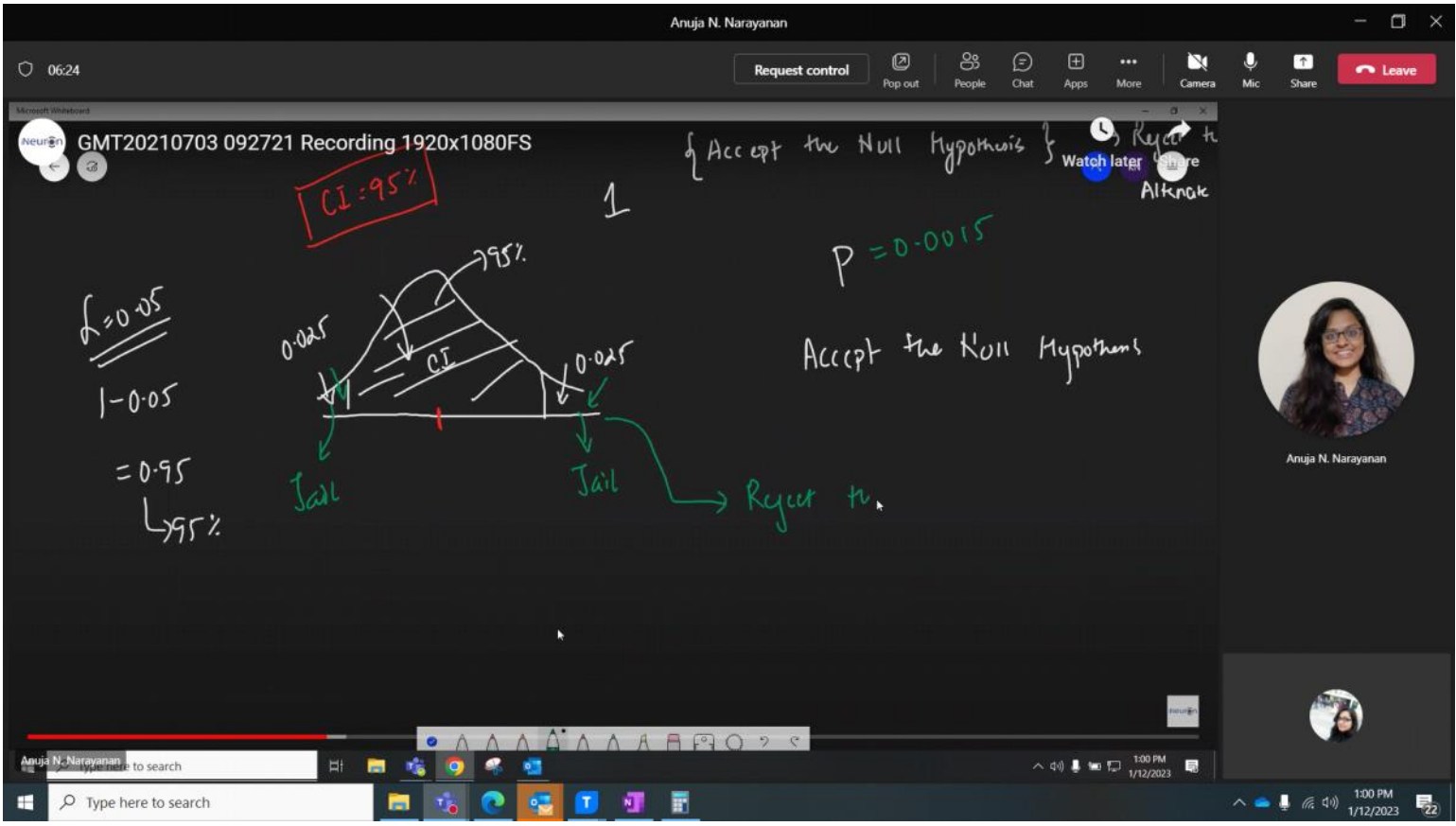
Thursday, January 12, 2023 12:33 PM

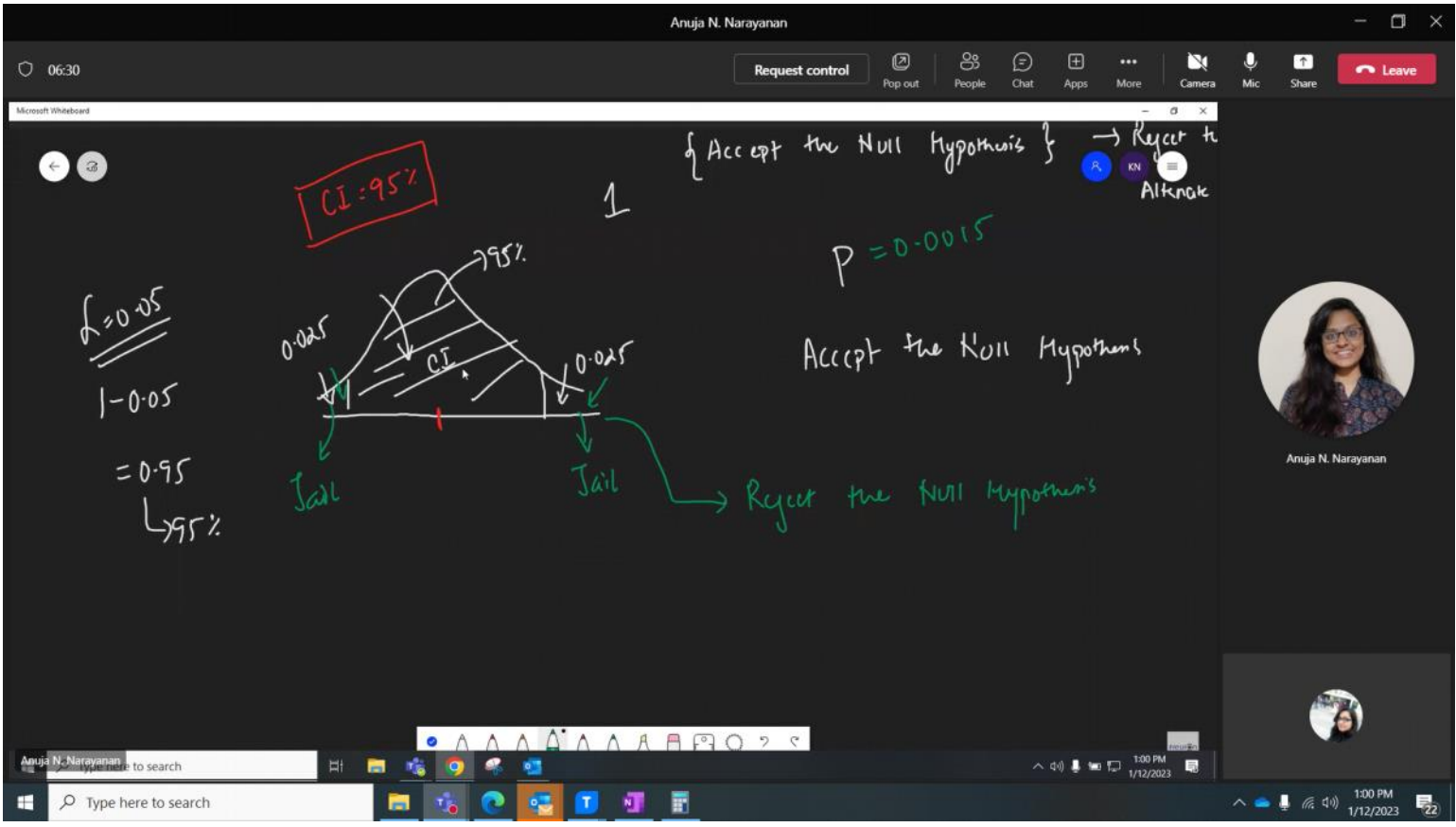


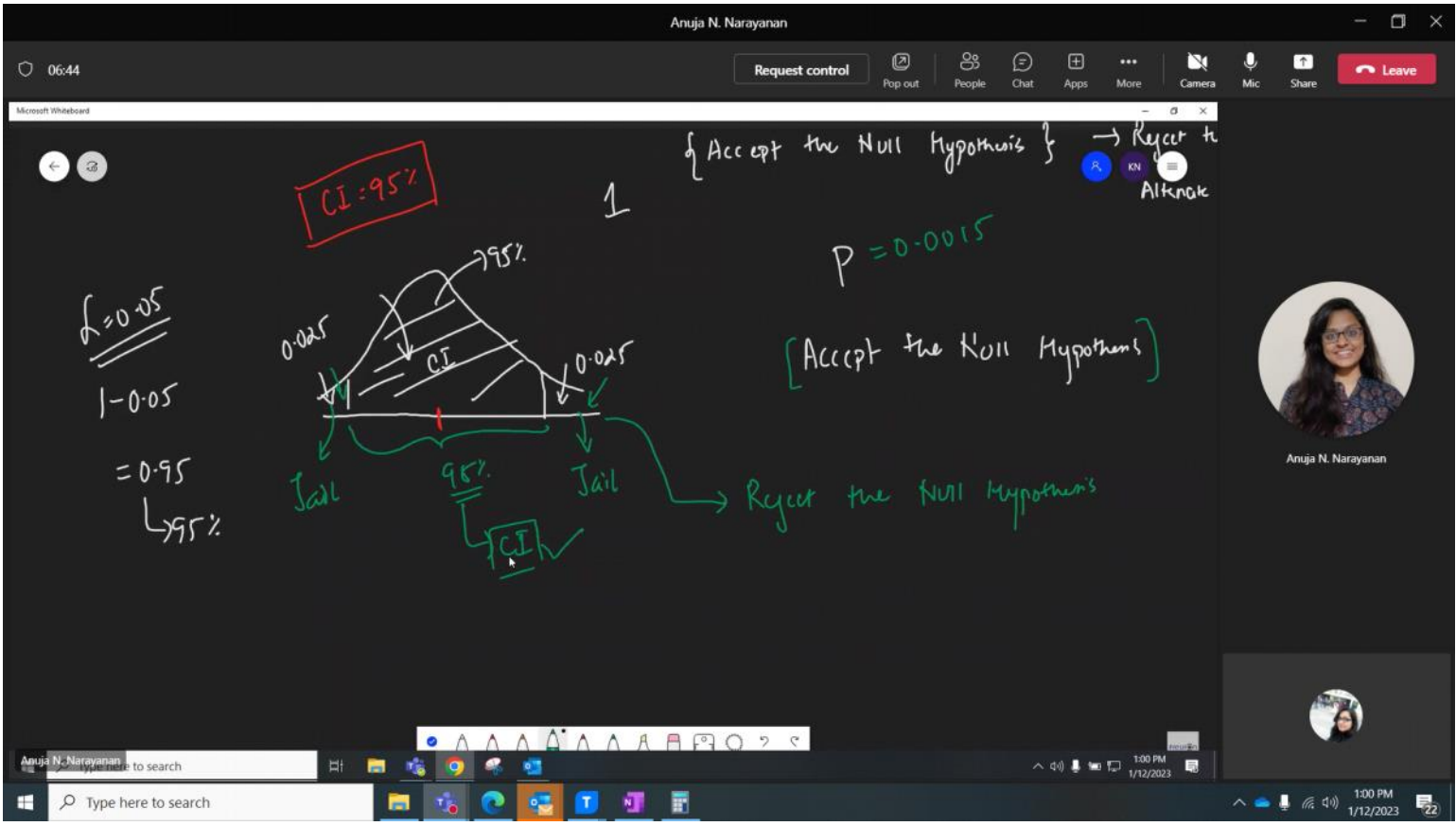


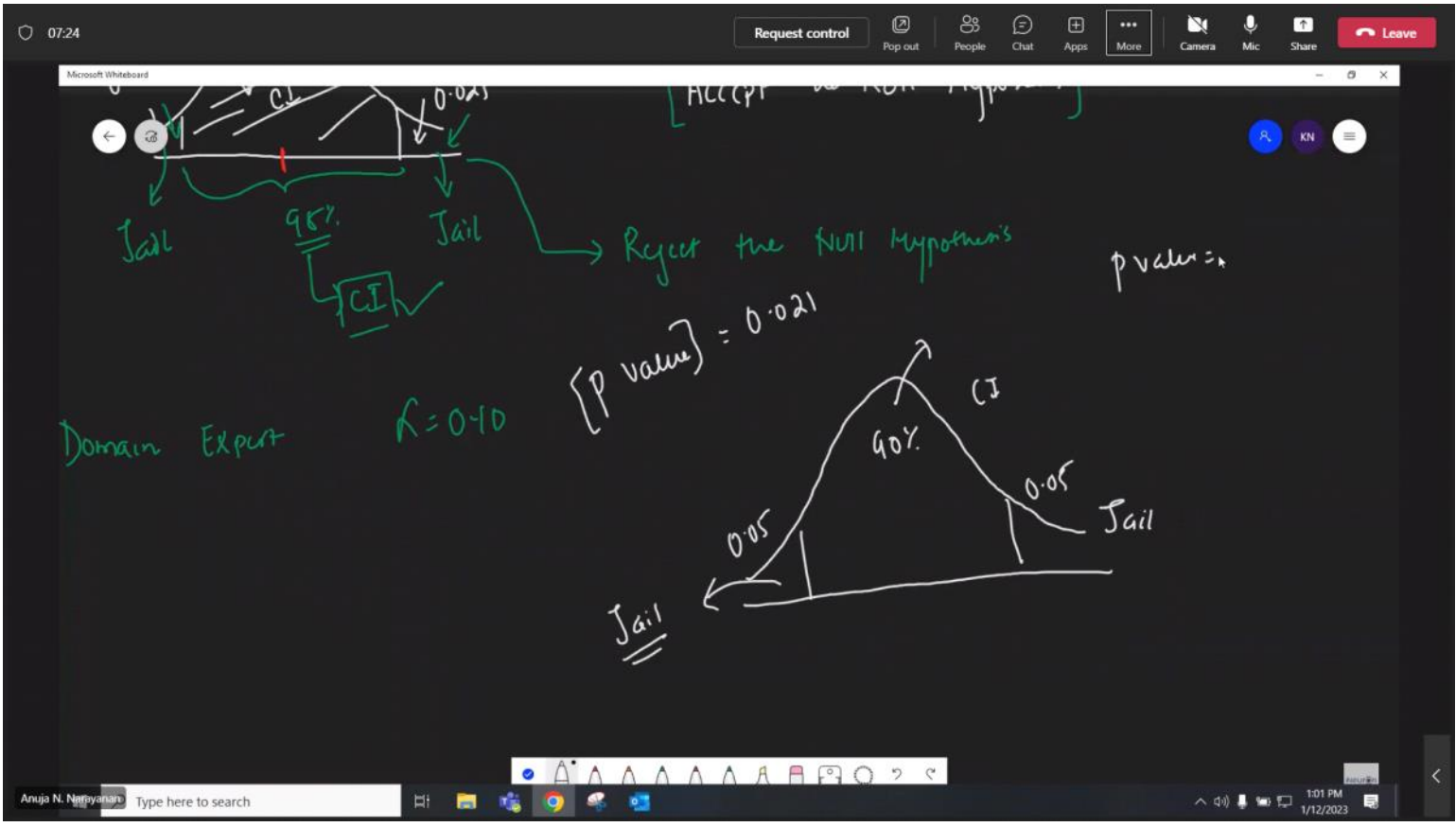

01:49:47
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
bar
Significance Value
α = 0.05
p=0.6
95% CI
Summerical
p=0.4
0.025
0.025
Tail
Tail
Domain Expert
Confidence Interval
α = 0.05
95%
1-0.05 = 0.95
p=0.001
Statistical Experiment → p Value = 0.07
Accept the Null Hypothesis
Anuja N. Narayanan
Type here to search
12:39 PM
1/12/2023

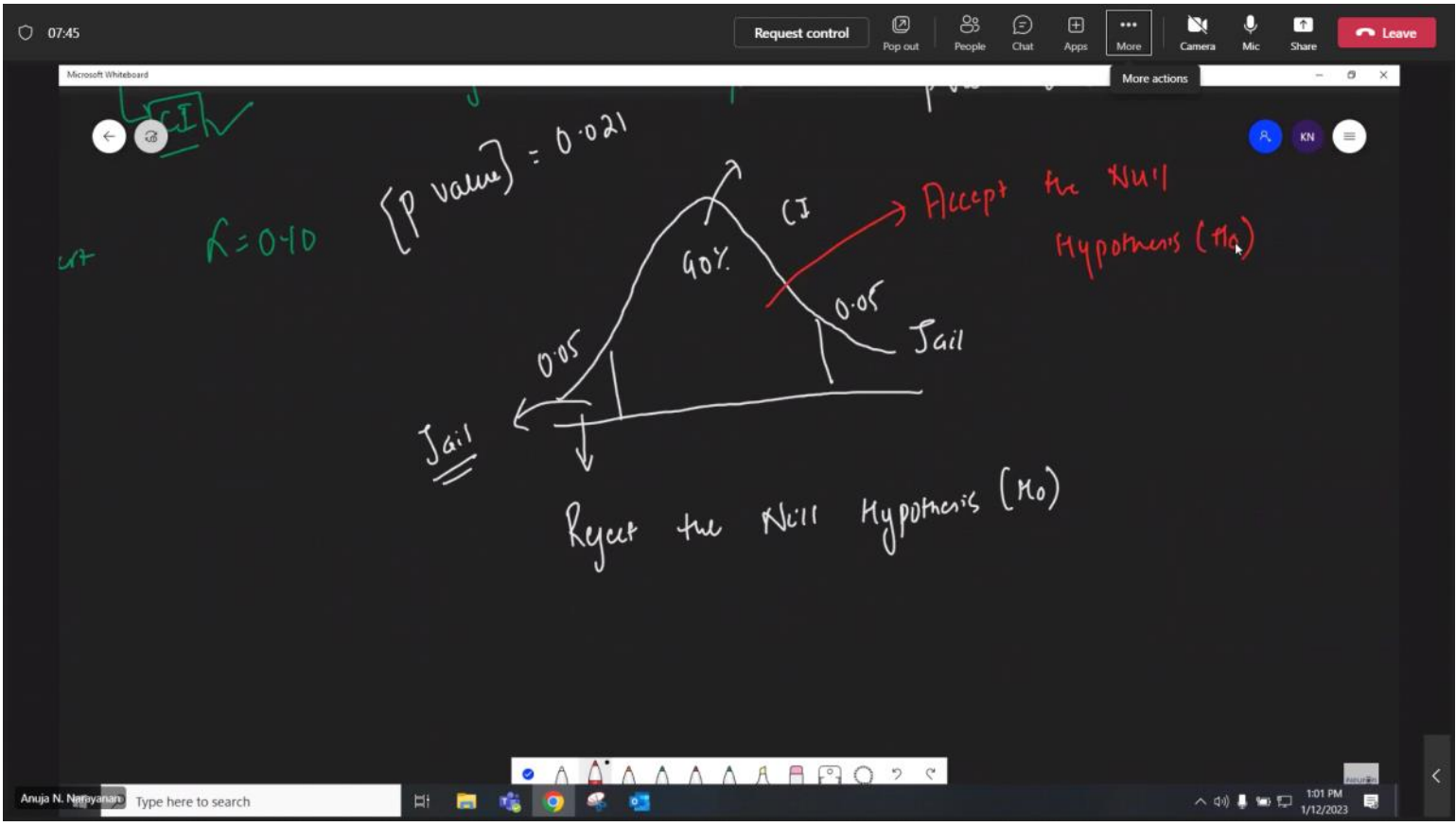

01:50:54
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Significance Value
α = 0.05
p=0.6
95% CI
Summerical
p=0.4
0.025
0.025
Tail
Tail
μ
95%
Confidence Interval
α = 0.05
95%
1-0.05 = 0.95
p=0.001
Statistical Experiment → p Value = 0.07
Accept the Null Hypothesis → Reject the
Alternate Hypothesis
Anuja N. Narayanan
Type here to search
12:40 PM
1/12/2023

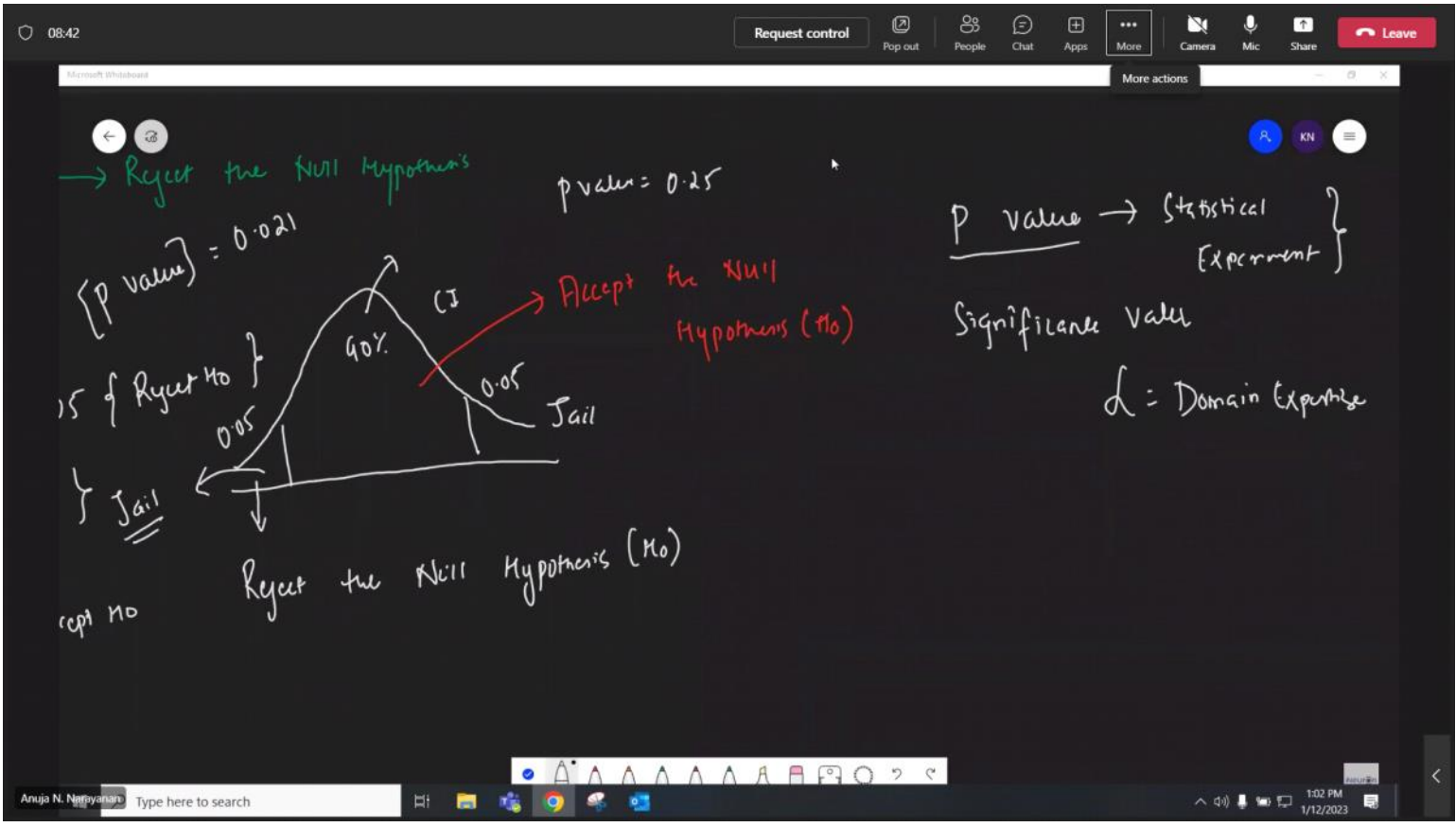

Anuja N. Narayanan
06:24
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
GMT20210703 092721 Recording 1920x1080FS
Watch later
CI : 95%
1
{ Accept the Null Hypothesis }
Alternate
P = 0.0015
Accept the Null Hypothesis
95%
0.025
CI
0.025
Jail
Jail
Reject the
Anuja N. Narayanan
Type here to search
1:00 PM
1/12/2023

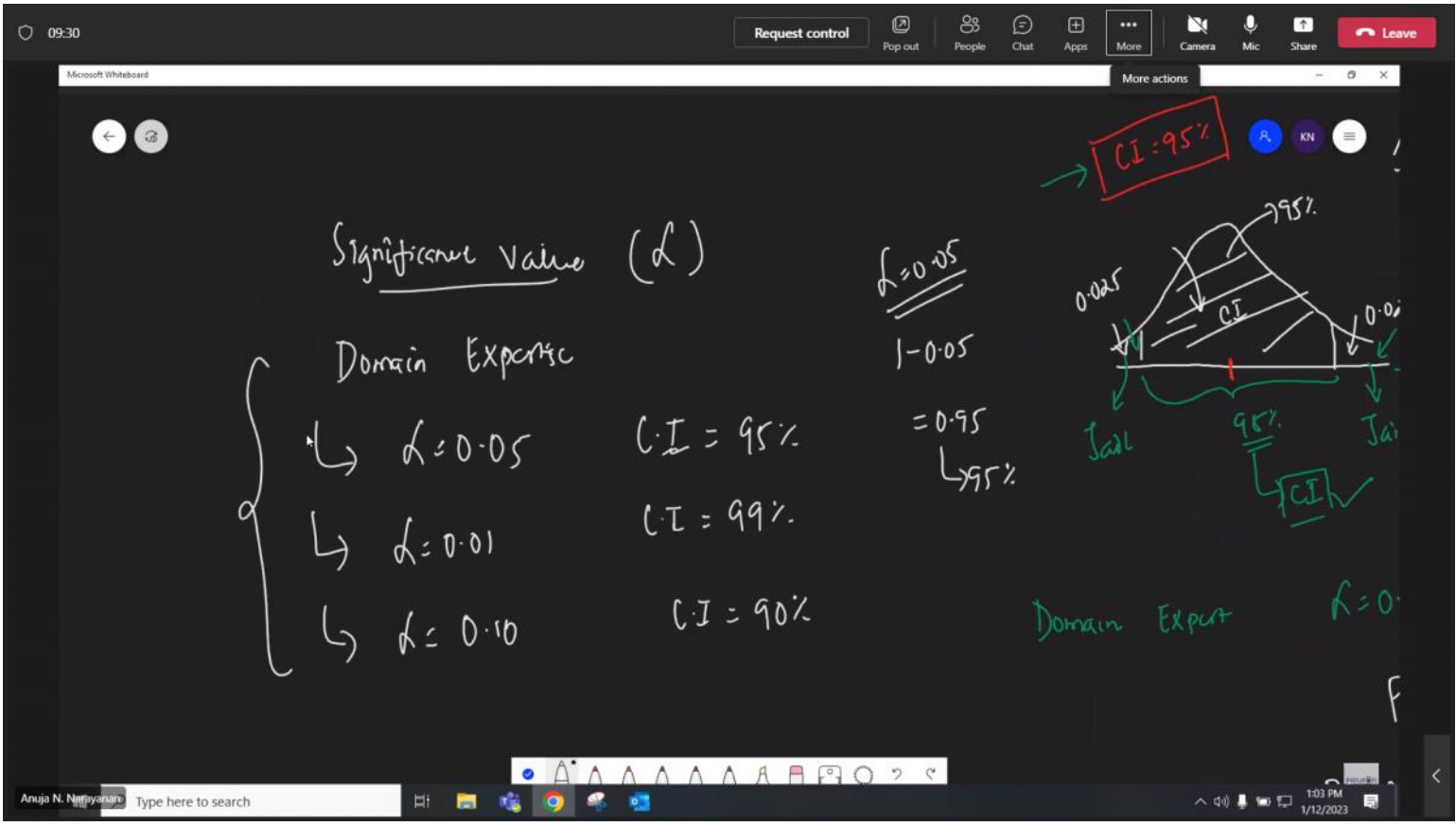

Anuja N. Narayanan
06:30
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
CI : 95%
1
{ Accept the Null Hypothesis }
Alternate
P = 0.0015
Accept the Null Hypothesis
95%
0.025
CI
0.025
Jail
Jail
Reject the Null Hypothesis
Anuja N. Narayanan
Type here to search
1:00 PM
1/12/2023

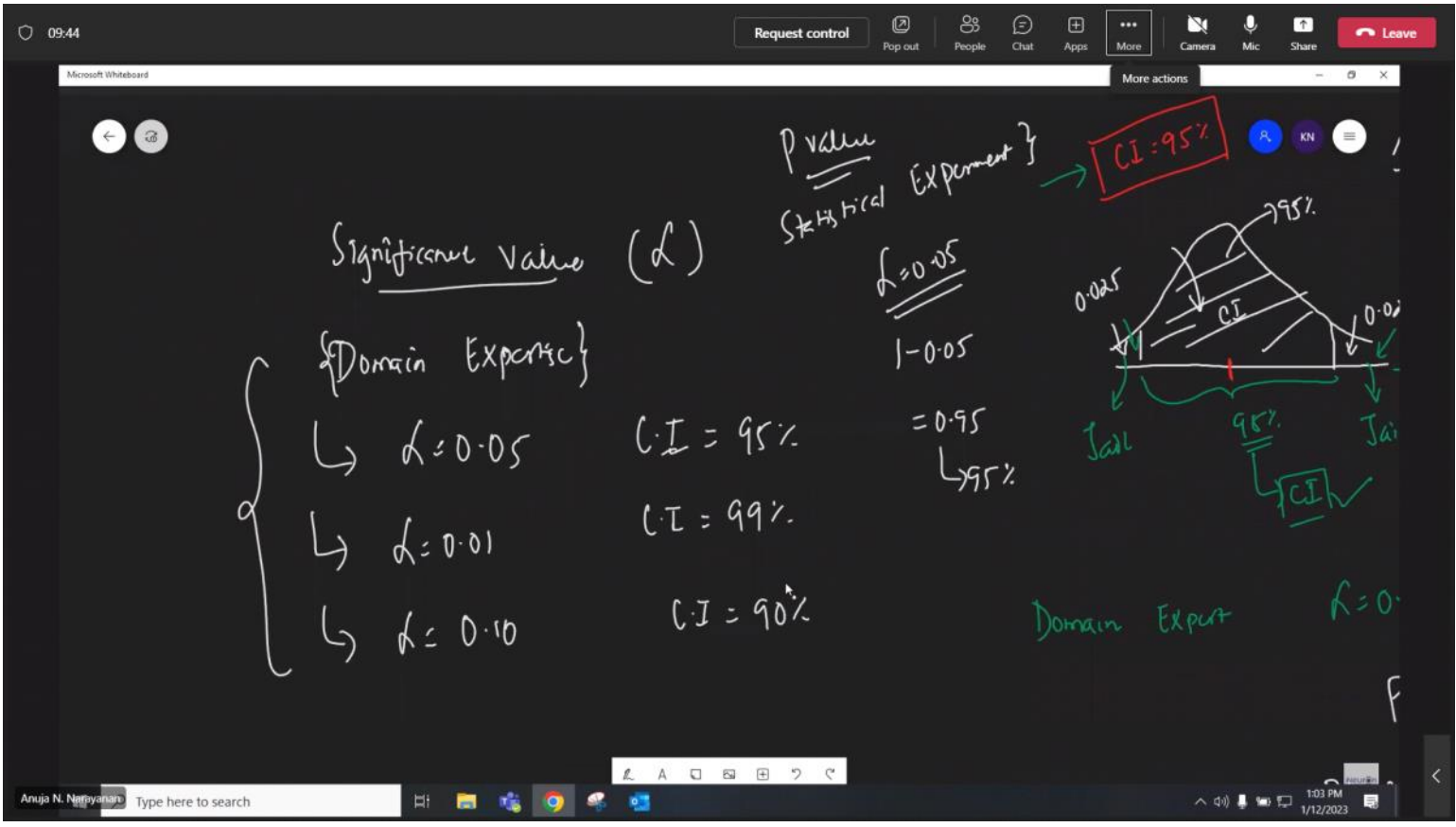

Anuja N. Narayanan
06:44
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
CI = 95%
1
{ Accept the Null Hypothesis } → Reject th
Alternate
P = 0.0015
[Accept the Null Hypothesis]
α = 0.05
1 − 0.05
= 0.95
↳ 95%
95%
0.025
CI
0.025
Jail
95%
Jail
CI
Reject the Null Hypothesis
Anuja N. Narayanan
Type here to search
1:00 PM
1/12/2023


07:24
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Jail
95%
Jail
CI
Reject the Null Hypothesis
p value =
[p value] = 0.021
Domain Expert
K = 0.10
CI
90%
0.05
0.05
Jail
Jail
Anuja N. Narayanan
Type here to search
1:01 PM
1/12/2023


07:45
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
KN
[P value] = 0.021
α = 0.10
CI
40%
0.05
0.05
Jail
Jail
Accept the Null Hypothesis (Ho)
Reject the Null Hypothesis (Ho)
Anuja N. Narayanan
Type here to search
1:01 PM
1/12/2023


08:42
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
KN
→ Reject the Null Hypothesis
pvalue = 0.25
[P value] = 0.021
P value → Statistical Experiment
Significance Value
α = Domain Expertise
CI
40%
0.05
0.05
Jail
Jail
Accept the Null Hypothesis (Ho)
Reject the Null Hypothesis (Ho)
Anuja N. Narayanan
Type here to search
1:02 PM
1/12/2023


09:30
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
KN
Significance Value (α)
Domain Expertise
α = 0.05
C.I = 95%
α = 0.01
C.I = 99%
α = 0.10
C.I = 90%
α = 0.05
1 - 0.05
= 0.95
95%
CI : 95%
0.025
CI
95%
Jail
Domain Expert
Anuja N. Narayanan
Type here to search
1:03 PM
1/12/2023


09:44
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
KN
P value
Statistical Experiment
Significance Value (α)
{Domain Expertise}
α = 0.05
C.I = 95%
α = 0.01
C.I = 99%
α = 0.10
C.I = 90%
α = 0.05
1 - 0.05
= 0.95
95%
CI : 95%
0.025
CI
95%
Jail
Domain Expert
Anuja N. Narayanan
Type here to search
1:03 PM
1/12/2023

15:14

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

More actions

in Experiment → Domain

$\alpha = 0.05$ C.I = 95% = 0.95 = 95%

$\alpha = 0.01$ C.I = 99%

$\alpha = 0.10$ C.I = 90%

1 - 0.05

Jail | 95% | Jail → Reject the Null Hypothesis

CI

pvalue = 0.2

Domain Expert $\alpha = 0.10$ {P value} = 0.021

$P \leq 0.05$ { Reject Ho }

$P > 0.05$ { } Jail

Accept Ho

90% | CI | 0.05 | 0.05 | Jail | Accept the H

Reject the Null Hypothesis (H_0)

Statistical Exp

{ loding → P value
Threrch → Region value }

Anuja N. Narayanan | Type here to search | 1:09 PM 1/12/2023

17:10

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

Statistical Test

σ / $\neq \sigma$

1. Z test { Comparison of Mean }
2. t test { Comparison of Mean }
3. F test { Comparison of Variance }
4. ANOVA { Analysis of Variance }
5. Chi Square { Comparison between 2 Categorical Variable }

Anuja N. Narayanan | Type here to search | 1:11 PM 1/12/2023

19:26 Request control Pop out People Chat Apps More Camera Mic Share Leave

Microsoft Whiteboard

{ Confidence Interval }

of Mean }

of Mean }

of Variance }

Variance }

between 2 Categorical Variable }

Point Estimate { The value of any statistic that estimates the value of a parameter is called a Point Estimate }

$\bar{x}$ —Estimate→ μ

2.95 3.00

Anuja N. Narayanan Type here to search 1:13 PM 1/12/2023

20:21 Request control Pop out People Chat Apps More Camera Mic Share Leave

Microsoft Whiteboard

Point Estimate { The value of any statistic that estimates the value of a parameter is called a Point Estimate }

$\bar{x}$ —Estimate→ μ

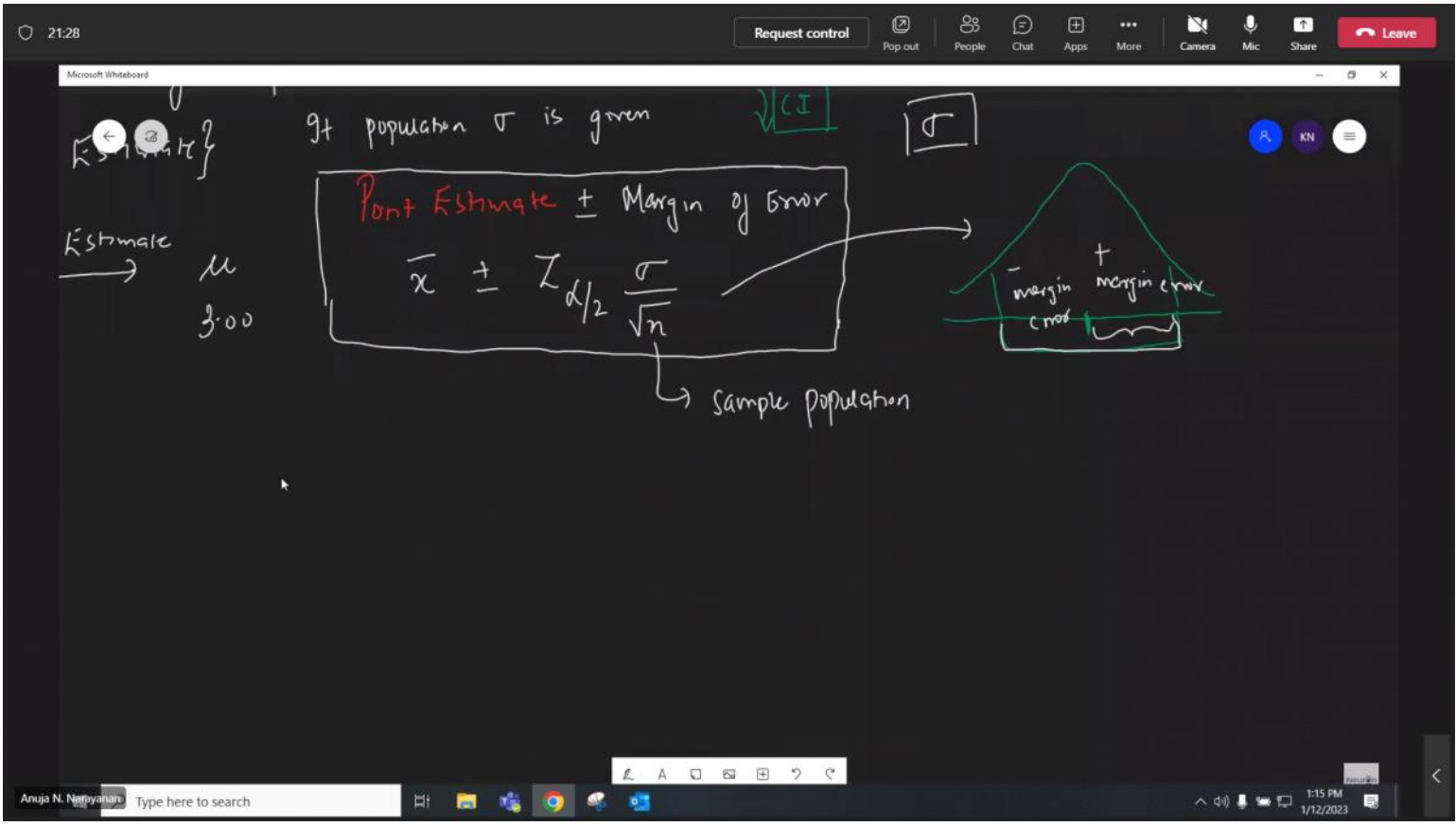
2.95 3.00

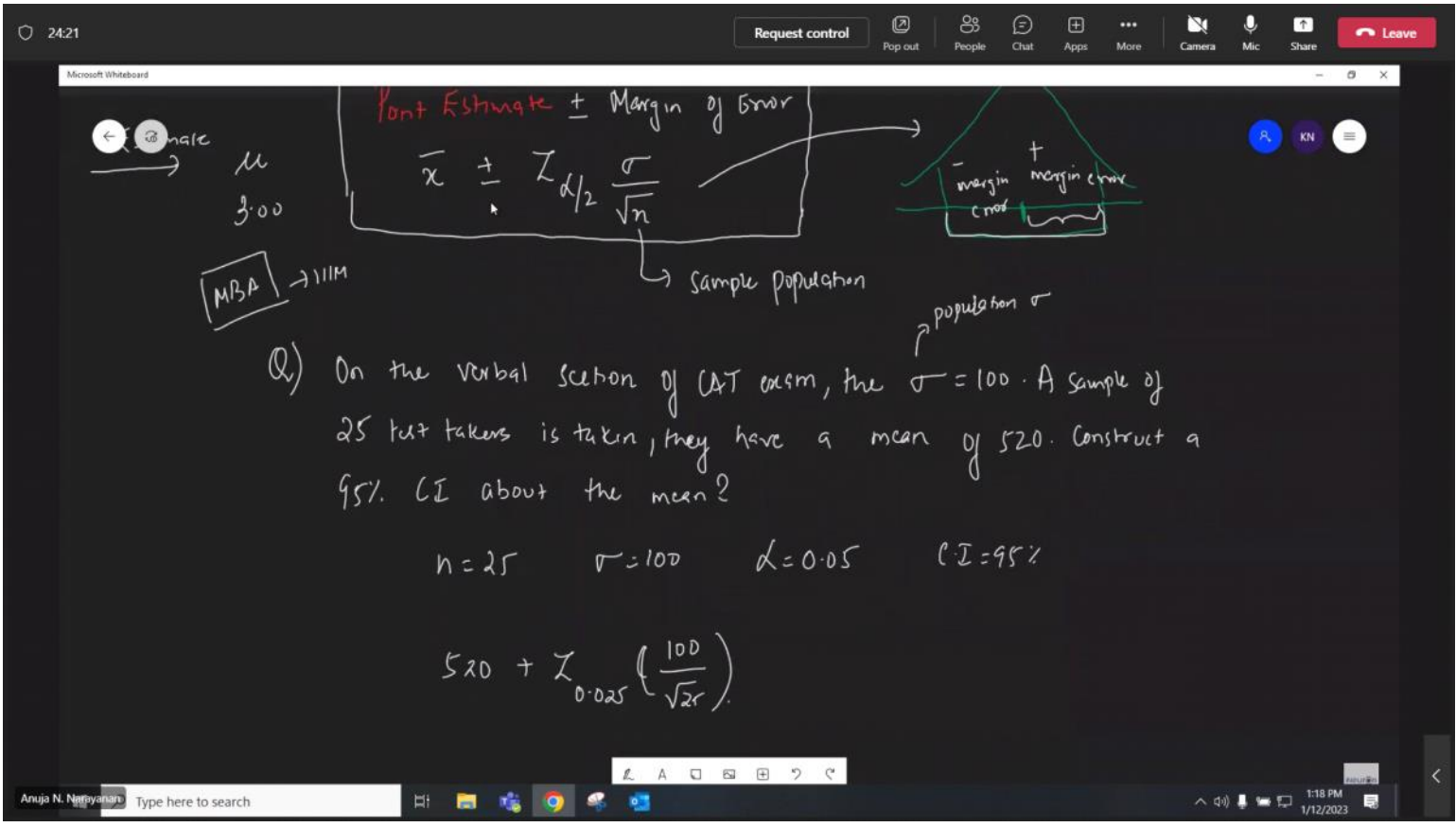
If population σ is given

Point Estimate ± Margin of Error

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Anuja N. Narayanan Type here to search 1:14 PM 1/12/2023


21:28
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
If population σ is given
σ
Point Estimate ± Margin of Error
$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Estimate
μ
3.00
Sample population
margin error
margin error
Anuja N. Narayanan
Type here to search
1:15 PM
1/12/2023


24:21
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Point Estimate ± Margin of Error
$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
μ
3.00
Sample population
MBA → IIIM
Q) On the verbal section of CAT exam, the σ = 100. A sample of 25 test takers is taken, they have a mean of 520. Construct a 95% CI about the mean?
population σ
n = 25
σ = 100
α = 0.05
C.I = 95%
$520 + Z_{0.025}\left(\frac{100}{\sqrt{25}}\right)$
Anuja N. Narayanan
Type here to search
1:18 PM
1/12/2023

26:08

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

Point Estimate ± Margin of Error

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

→ Sample Population

− margin error + margin error

MBA → IIIM

Q) On the verbal section of CAT exam, the $\sigma = 100$. A sample of 25 test takers is taken, they have a mean of 520. Construct a 95% CI about the mean?

↑ population σ

-1.96 $+1.96$

0.025 0.025

−ve +ve

$n = 25 \quad \sigma = 100 \quad \alpha = 0.05 \quad C.I = 95\%$

$$\Rightarrow 520 + Z_{0.025}\left(\frac{100}{\sqrt{25}}\right) =$$

Anuja N. Narayanan | Type here to search | 1:20 PM 1/12/2023

25:55

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

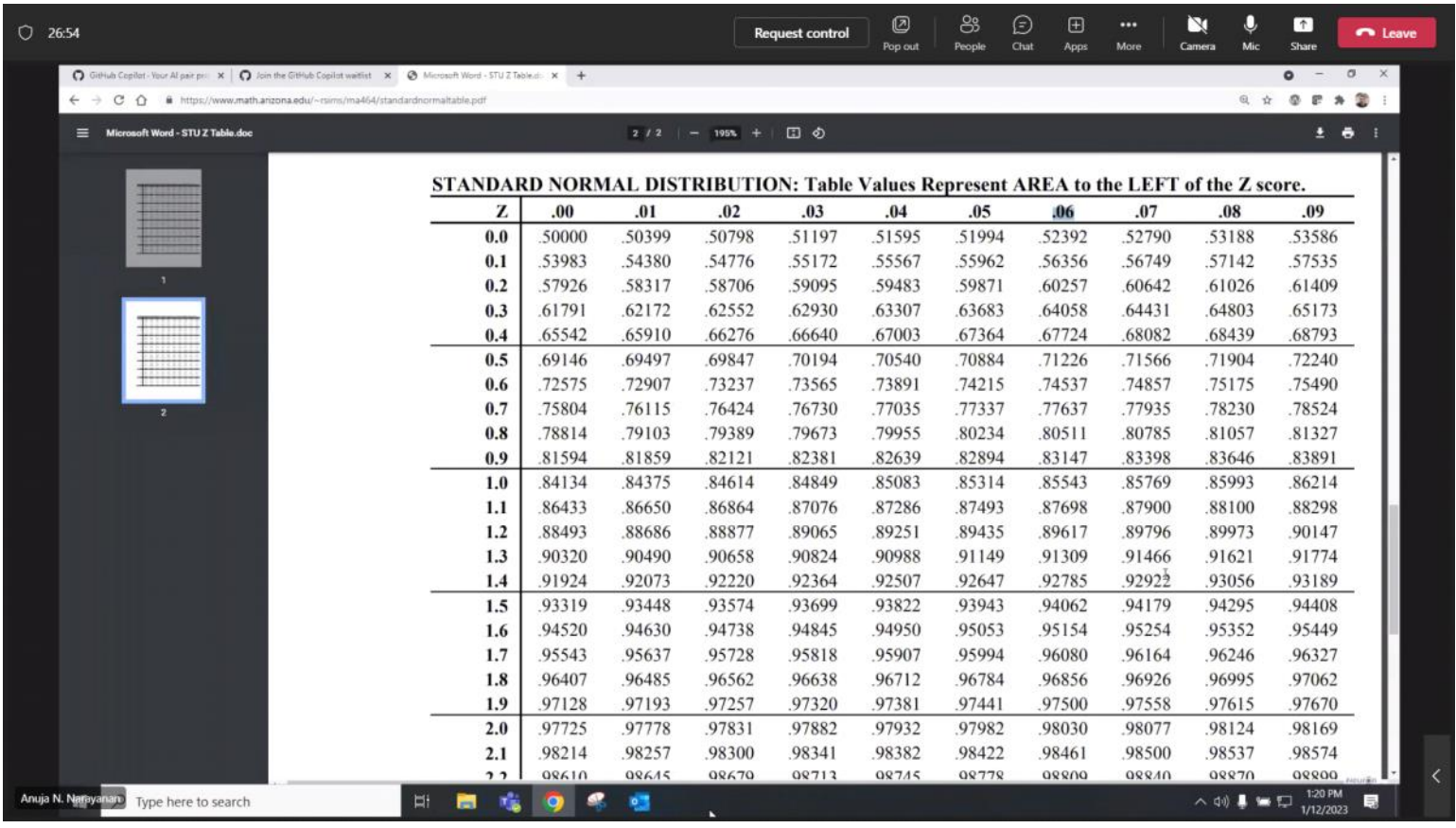
https://www.math.arizona.edu/~rsims/ma464/standardnormaltable.pdf

Microsoft Word - STU Z Table.doc

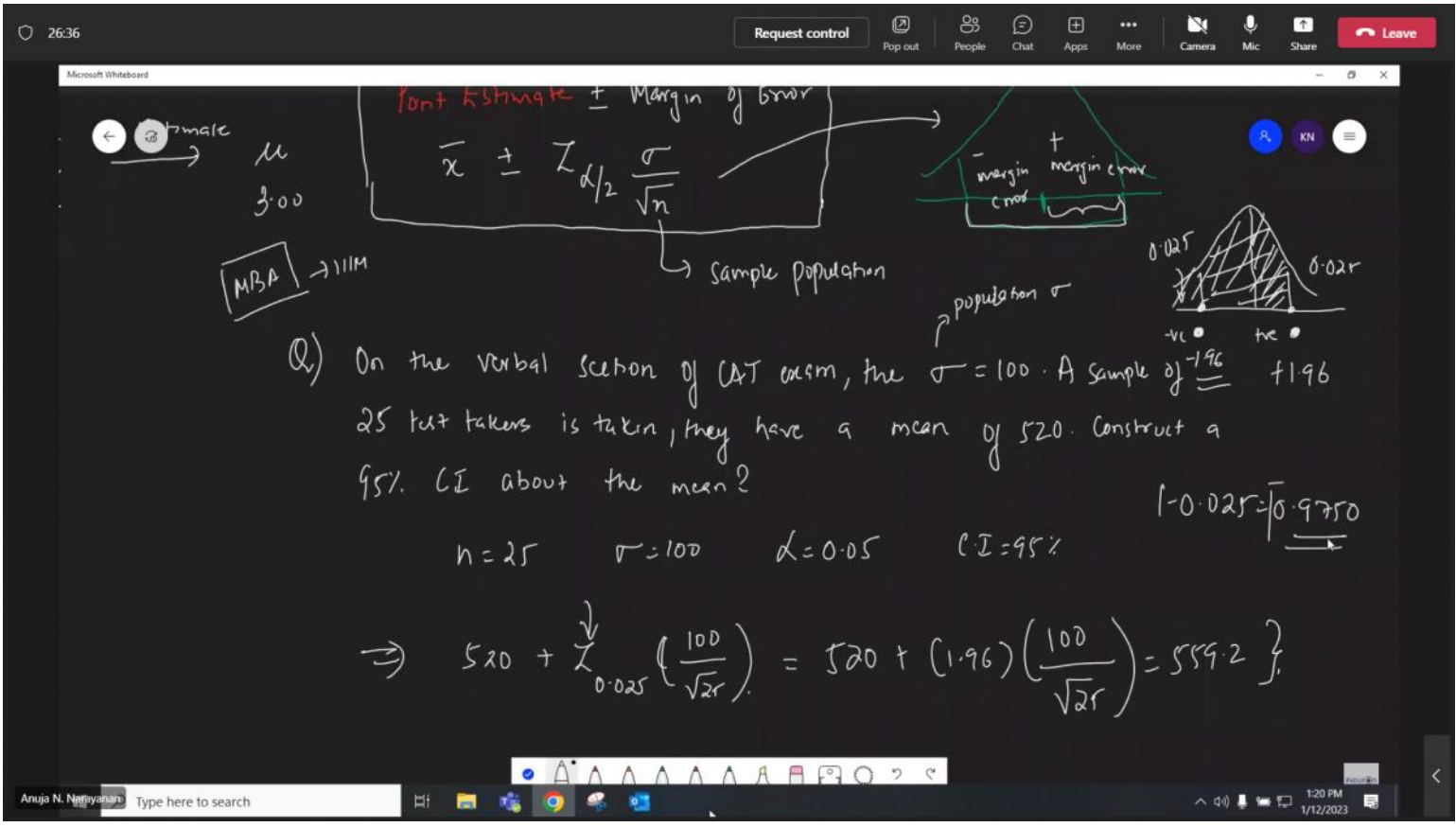
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
					.99683	.99693	.99702	.99711	.99720	.99728	.99736
					.99767	.99774	.99781	.99788	.99795	.99801	.99807
					.99831	.99836	.99841	.99846	.99851	.99856	.99861
					.99878	.99882	.99886	.99889	.99893	.99896	.99900

if p is less than .05 - Google Sea... | Post Attendee - Zoom - Google... | Microsoft Word - STU Z Table.d...

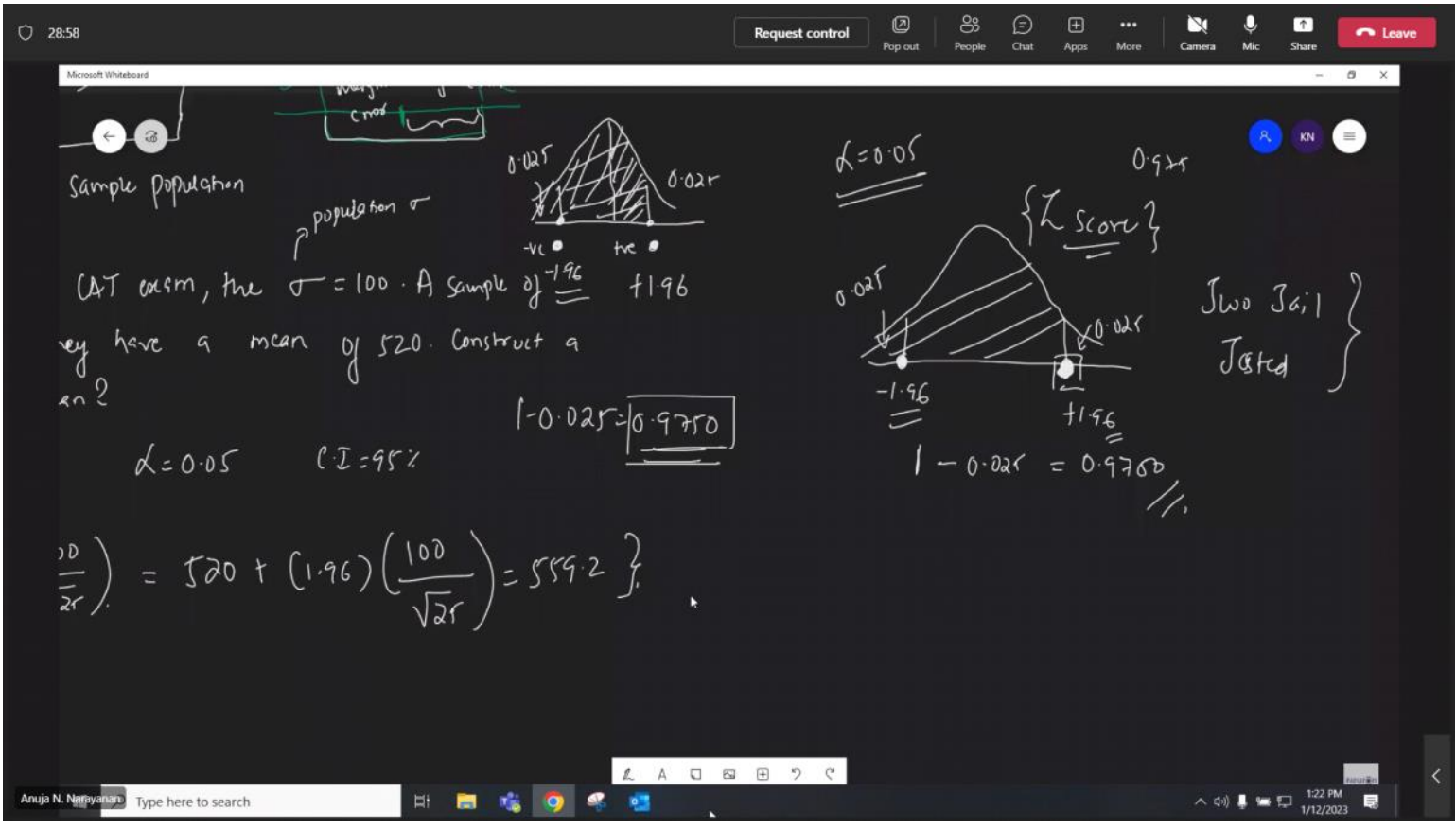
Anuja N. Narayanan | Type here to search | 1:19 PM 1/12/2023

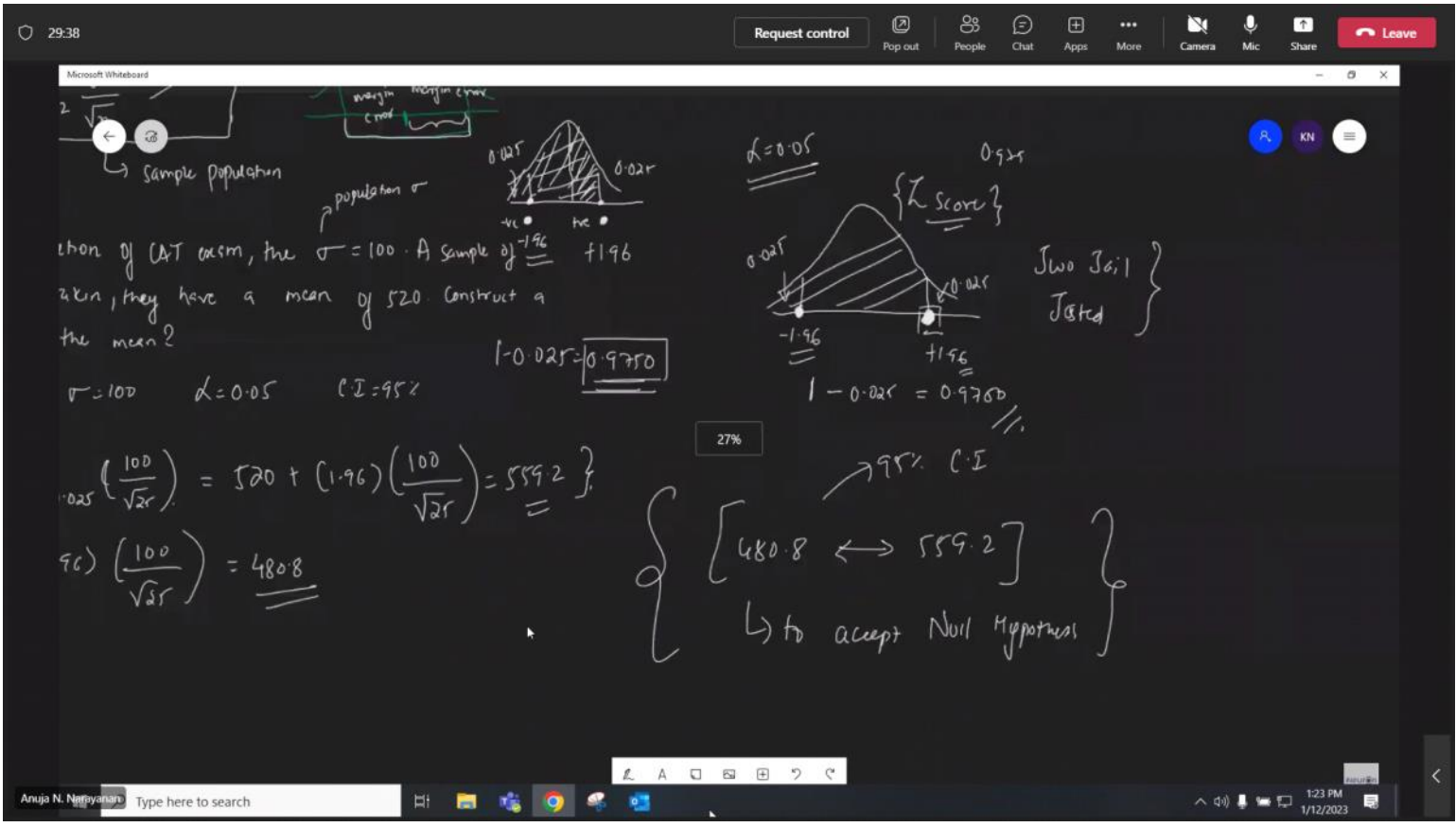


Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899



$$\Rightarrow 520 + Z_{0.025}\left(\frac{100}{\sqrt{25}}\right) = 520 + (1.96)\left(\frac{100}{\sqrt{25}}\right) = 559.2$$


28:58
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Sample Population
population σ
CAT exam, the σ = 100. A sample of
they have a mean of 520. Construct a
an?
α = 0.05
C.I = 95%
1 - 0.025 = 0.9750
-1.96
+1.96
0.025
0.025
-ve
+ve
α = 0.05
0.95
{Z score}
0.025
0.025
-1.96
+1.96
1 - 0.025 = 0.9750
Two Tail Tested
= 520 + (1.96)(100/√25) = 559.2
Anuja N. Narayanan
Type here to search
1:22 PM
1/12/2023


29:38
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Sample Population
population σ
tion of CAT exam, the σ = 100. A sample of
ation, they have a mean of 520. Construct a
the mean?
σ = 100
α = 0.05
C.I = 95%
1 - 0.025 = 0.9750
-1.96
+1.96
0.025
0.025
-ve
+ve
α = 0.05
0.95
{Z score}
0.025
0.025
-1.96
+1.96
1 - 0.025 = 0.9750
Two Tail Tested
= 520 + (1.96)(100/√25) = 559.2
= 480.8
95% C.I
[480.8 ⟷ 559.2]
↳ to accept Null Hypothesis
27%
Anuja N. Narayanan
Type here to search
1:23 PM
1/12/2023

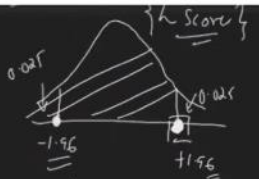
30:25 Request control Pop out People Chat Apps More Camera Mic Share Leave

Microsoft Whiteboard

Q) ... Verbal section of CAT exam, the $\sigma = 100$. A sample of 25 test takers is taken, they have a mean of 520. Construct a 95% CI about the mean?

$n = 25 \quad \sigma = 100 \quad \alpha = 0.05 \quad CI = 95\%$

$1 - 0.025 = 0.9750$



$1 - 0.025 = 0.9750$

$\Rightarrow 520 + Z_{0.025}\left(\frac{100}{\sqrt{25}}\right) = 520 + (1.96)\left(\frac{100}{\sqrt{25}}\right) = 559.2$

$\Rightarrow 520 - (1.96)\left(\frac{100}{\sqrt{25}}\right) = 480.8$

95% C.I

$[480.8 \longleftrightarrow 559.2]$

↳ to accept Null Hypothesis

$H_0 : \{\mu = 520\}$ ✓

$H_1 : \{\mu \neq 520\}$

Anuja N. Narayanan Type here to search 1:24 PM 1/12/2023

34:38 Request control Pop out People Chat Apps More Camera Mic Share Leave

Microsoft Whiteboard

$H_1 : \{\mu \neq 520\}$

Z test

↓

t-test

Population $\sigma \neq$ given

(2) On the verbal section of CAT, a sample of 25 test takers has a mean of 520. with standard deviation of 80. Construct 95% C.I about the mean?

Ans).

GMT20210703 092721 Recording 1920x1080FS

Anuja N. Narayanan Type here to search 1:28 PM 1/12/2023

35:42

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

; test

(2) On the verbal section of CAT, a sample of 25 test takers has a mean of 520. with standard deviation of 80. Construct 95% C.I about the mean?

Ans).

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$ ← Sample sd

$\alpha = 0.05$

$\bar{x} = 520$

$n = 25$

$s = 80$

$$= 520 + t_{0.025}\left(\frac{80}{\sqrt{25}}\right) =$$

GMT20210703 092721 Recording 1920x1080FS

Anuja N. Narayanan | Type here to search | 1:29 PM 1/12/2023

36:26

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

On the verbal section of CAT, a sample of 25 test takers has a mean of 520. with standard deviation of 80. Construct 95% C.I about the mean?

).

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$ ← Sample sd

$\alpha = 0.05$

$\bar{x} = 520$

$n = 25$

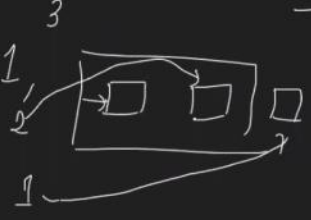
$s = 80$

Degree of freedom

$n - 1 = 25 - 1 = \{24\}$

$$= 520 + t_{0.025}\left(\frac{80}{\sqrt{25}}\right) =$$

3 people



$3 - 1 =$

Anuja N. Narayanan | Type here to search | 1:30 PM 1/12/2023

38:00

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

t-table.xls

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Anuja N. Narayanan | Type here to search | 1:32 PM 1/12/2023

41:30

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

GMT20210703 092721 Recording 1920x1080FS | Watch later | Share

a sample of 25 test takers has a mean of 520. with standard deviation of 80. Construct 95% C.I about the mean?

$\alpha = 0.05$

$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ → Sample sd

$\bar{x} = 520$

$n = 25$

$s = 80$

Degree of freedom

$n - 1 = 25 - 1 = 24$

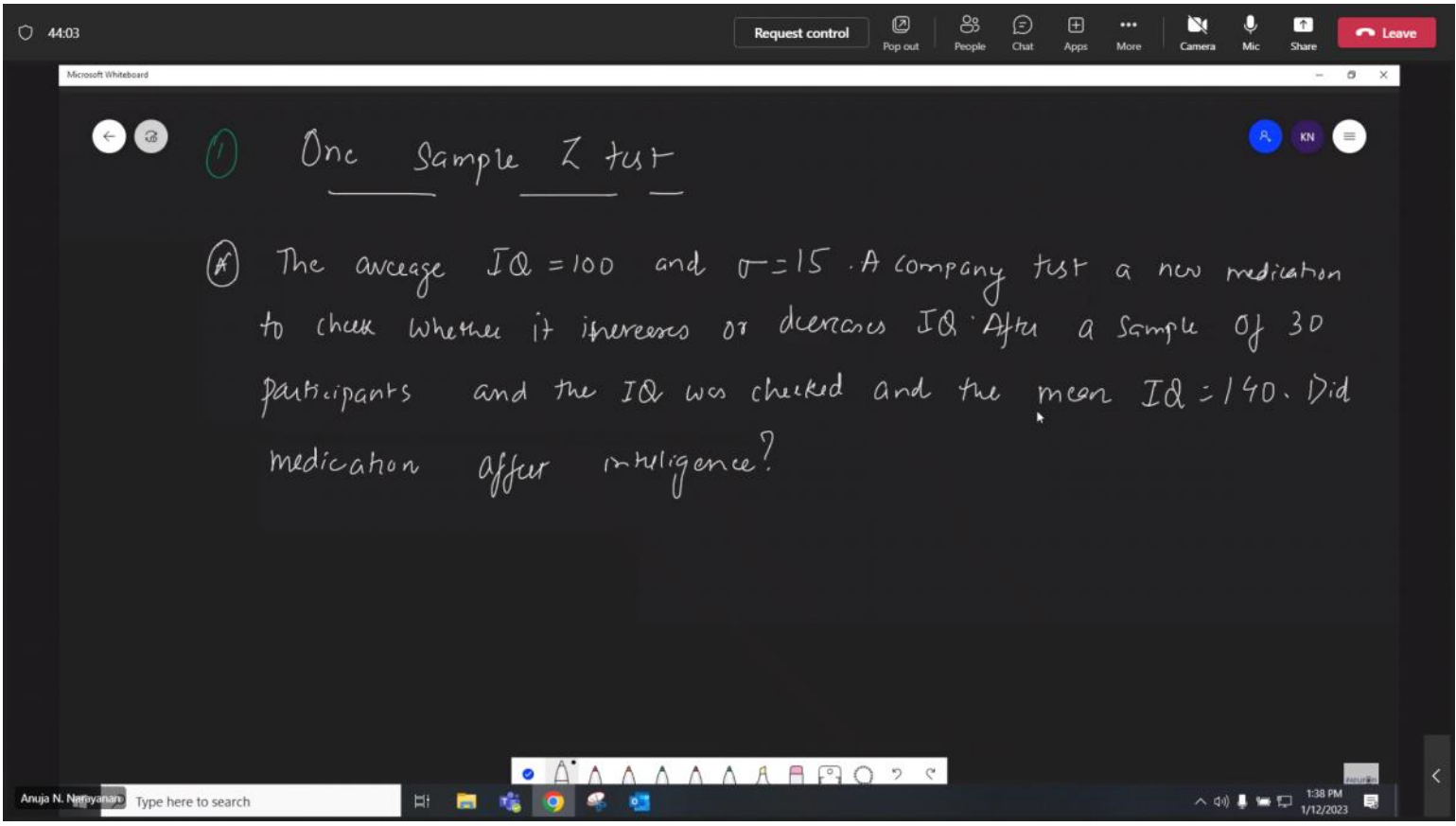
C.I

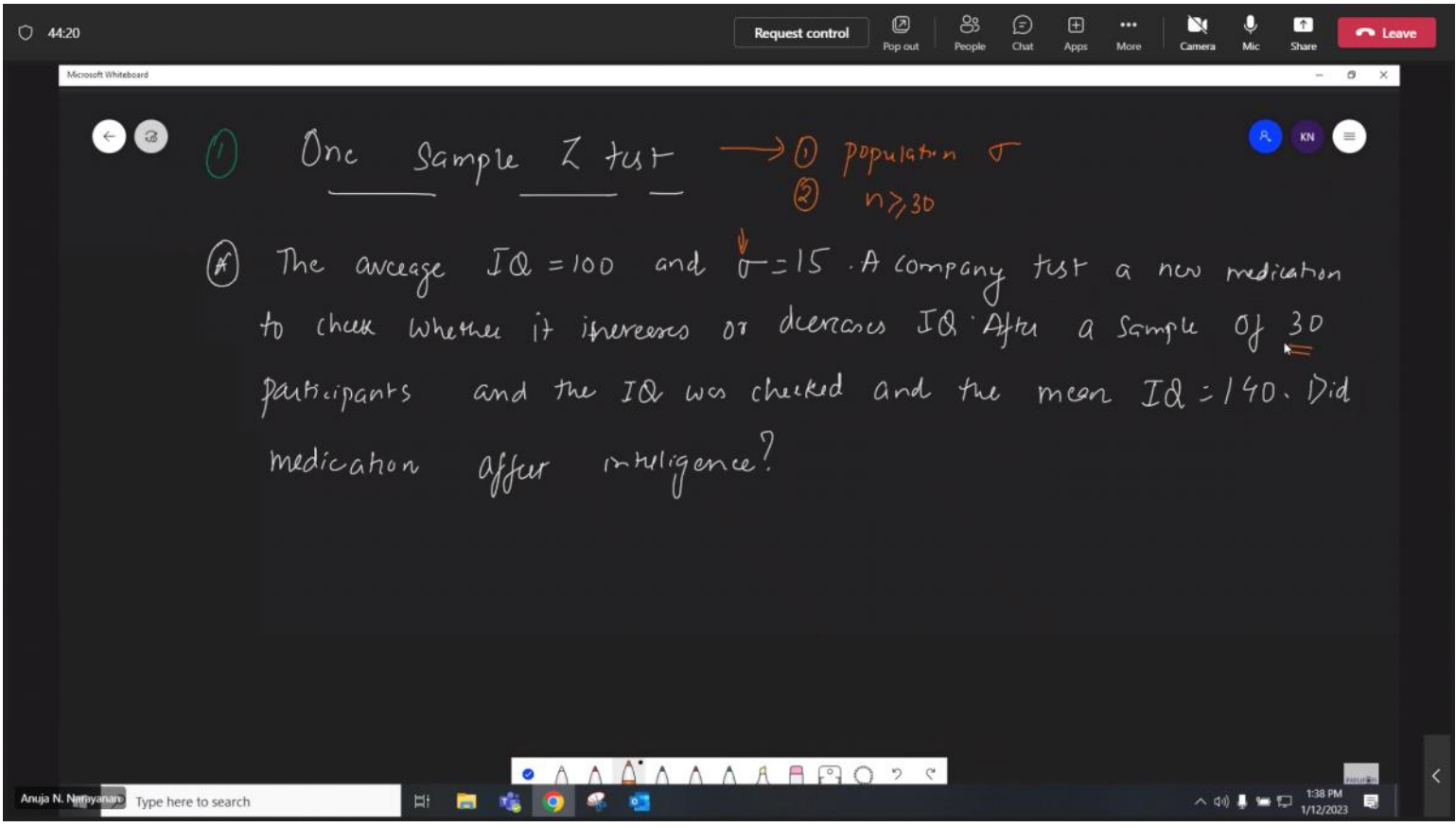
{ 520 → Accept Ho }

$= 520 + t_{0.025}\left(\frac{80}{\sqrt{25}}\right) = 520 + 2.064\left(\frac{80}{\sqrt{25}}\right) = 553.022$ } Upper

$= 520 - t_{0.025}\left(\frac{80}{\sqrt{25}}\right) = 520 - 2.064\left(\frac{80}{\sqrt{25}}\right) = 486.978$ } Lower

Anuja N. Narayanan | Type here to search | 1:35 PM 1/12/2023


44:03
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
KN
1 One Sample Z test
A The average IQ = 100 and σ = 15. A company test a new medication to check whether it increases or decreases IQ. After a sample of 30 participants and the IQ was checked and the mean IQ = 140. Did medication affect intelligence?
Anuja N. Narayanan
Type here to search
1:38 PM
1/12/2023


44:20
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
KN
1 One Sample Z test → ① population σ ② n ≥ 30
A The average IQ = 100 and σ = 15. A company test a new medication to check whether it increases or decreases IQ. After a sample of 30 participants and the IQ was checked and the mean IQ = 140. Did medication affect intelligence?
Anuja N. Narayanan
Type here to search
1:38 PM
1/12/2023

47:45

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

GMT20210703 092721 Recording 1920x1080FS

Watch later | Share

participants and the IQ was checked and the mean IQ = 140. Did medication after intelligence? The $\alpha = 0.05$

③ 2 Tail test

0.025 | 0.025

-1.96 | 1.96

① Null Hypothesis (H_0) : $\mu = 100$

② Alternate Hypothesis (H_1) : $\mu \neq 100$

④ Calculate Z Test Statistics

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$$

Anuja N. Narayanan

09:04

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Anuja N. Narayanan

medication after intelligence? The $\alpha = 0.05$

③ 2 Tail test

0.025 | 0.025

-1.96 | 1.96

① Null Hypothesis (H_0) : $\mu = 100$

② Alternate Hypothesis (H_1) : $\mu \neq 100$

④ Calculate Z Test Statistics

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$$

GMT20210703 092721 Recording 1920x1080FS

Anuja N. Narayanan

Type here to search

3:11 PM 1/12/2023

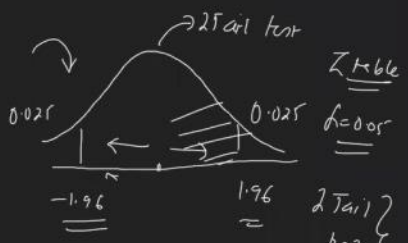
Anuja N. Narayanan

13:18 | Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

check whether it increases or decreases IQ. After a sample of 30 participants and the IQ was checked and the mean IQ = 140. Did medication (increase ✓ / decrease ✓) affect intelligence? The $\alpha = 0.05$ ③

1) Null Hypothesis (H_0): $\mu = 100$
2) Alternate Hypothesis (H_1): $\mu \neq 100$ { 1 Tail: $\mu > 100$, $\mu < 100$ }
3) Calculate Z Test Statistics }



$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$$

[-1.96 to 1.96]

$14.60 > 1.96$

Anuja N. Narayanan

Type here to search | 30°C AQI 77 | ENG IN | 03-07-2021 | 3:15 PM 1/12/2023

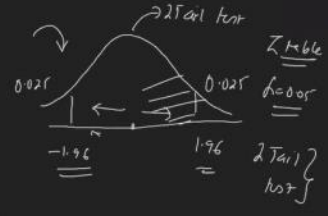
Anuja N. Narayanan

15:19 | Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

average IQ = 100 and σ = 15. A company test a new medication ... check whether it increases or decreases IQ. After a sample of 30 ... participants and the IQ was checked and the mean IQ = 140. Did ... medication (increase ✓ / decrease ✓) affect intelligence? The $\alpha = 0.05$ ③

Hypothesis (H_0): $\mu = 100$
... Hypothesis (H_1): $\mu \neq 100$ { 1 Tail: $\mu > 100$, $\mu < 100$ }
... Z Test Statistics }



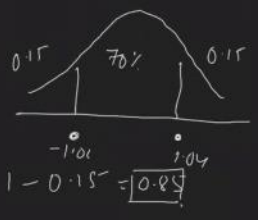
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$$

[-1.96 to 1.96]

$14.60 > 1.96$

↑ Significance value

$\alpha = 0.30$ } C.I } 70% ✓

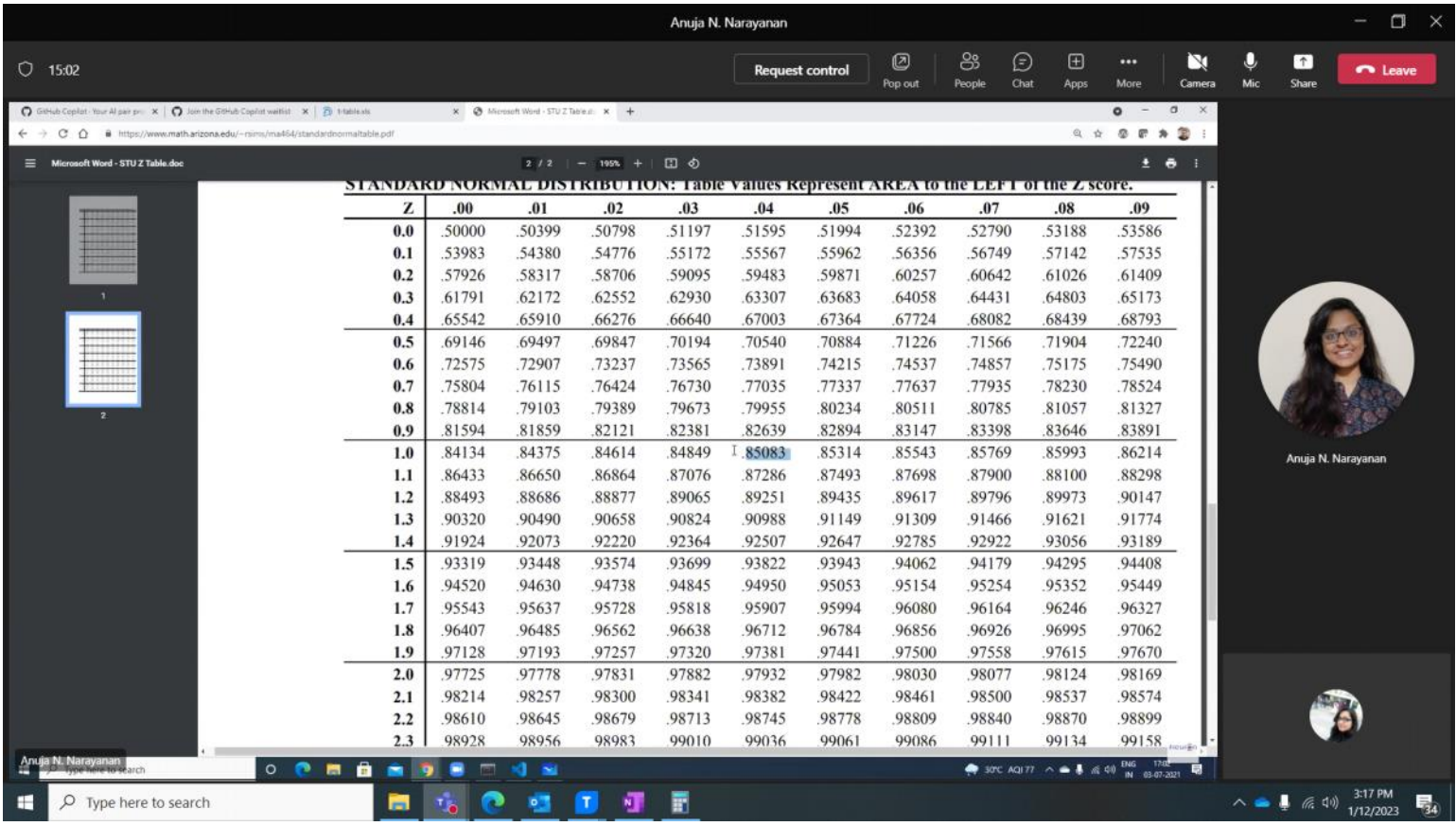


$1 - 0.15 = 0.85$

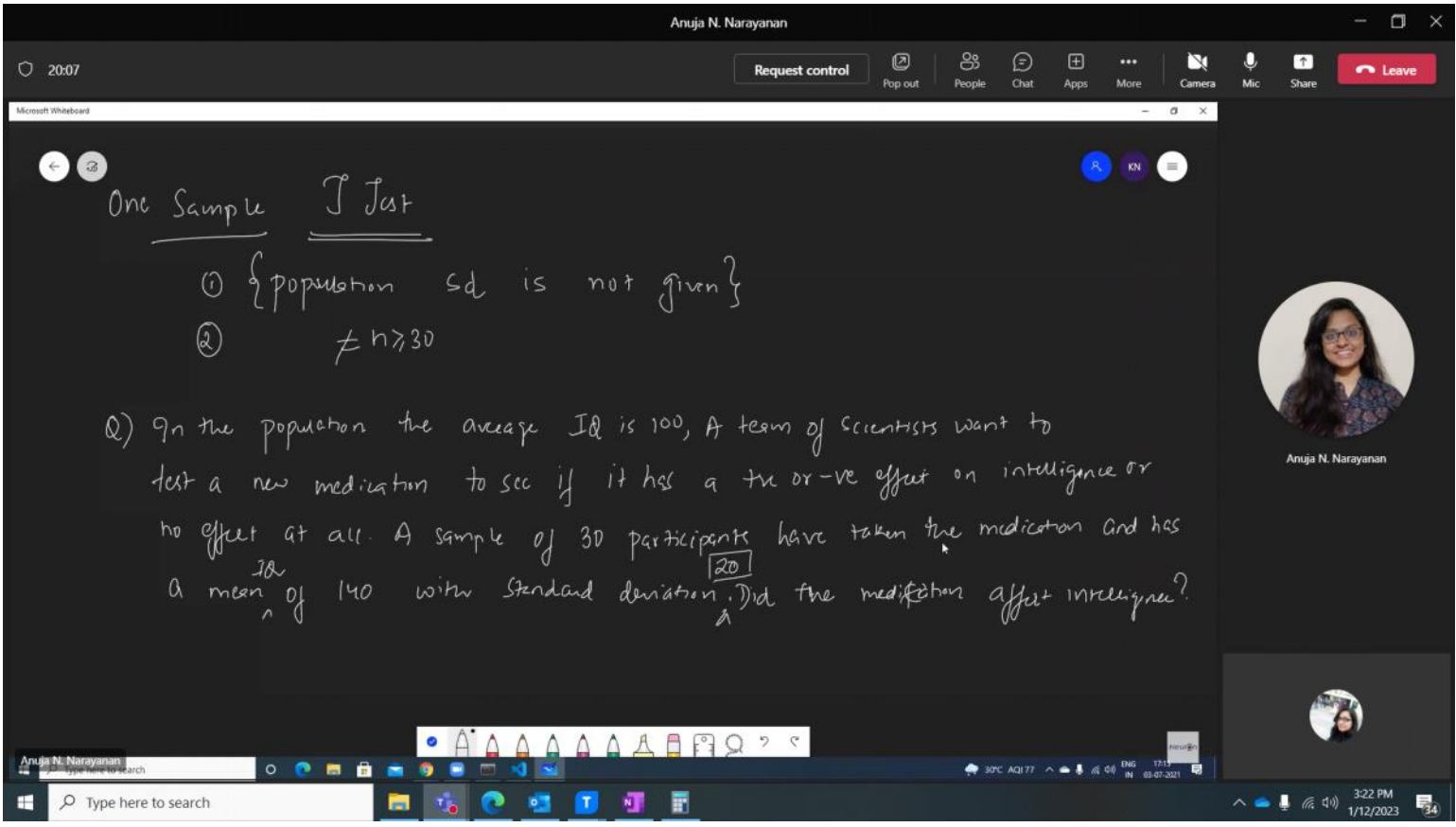
Conclusions : Z = 14.60 Reject Null Hypothesis & There is ... an effect of med...

Anuja N. Narayanan

Type here to search | 30°C AQI 77 | ENG IN | 03-07-2021 | 3:17 PM 1/12/2023



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158



Anuja N. Narayanan

21:58

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

data {mean, sd}

Q) In the population the average IQ is 100, A team of scientists want to test a new medication to see if it has a +ve or -ve effect on intelligence or no effect at all. A sample of 25 participants have taken the medication and has a mean IQ of 140 with standard deviation 20. Did the medication affect intelligence?

$\alpha = 0.05$

① $H_0 = \mu = 100$

$H_1 = \mu \neq 100$

② Degrees of freedom

$df = 24$

③

Anuja N. Narayanan

Type here to search

3:24 PM 1/12/2023

Anuja N. Narayanan

24:20

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

[-2.064 +2.064]

④ t test

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{25}} = \frac{40}{20/5} = \frac{40}{4} = 10$$

26%

$10 >$ Reject Null Hypothesis, Accept the alternate Hypothesis

Medication has increased the intelligence.

Anuja N. Narayanan

Type here to search

3:26 PM 1/12/2023

Anuja N. Narayanan

33:14 | Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

GMT20210703 092721 Recording 1920x1080FS

Watch later | Share

One Sample Z Test for {Proportion}

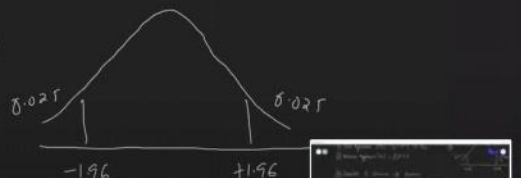
A survey claims that 9 out of 10 doctors recommend Aspirin for their patients with headache. To test this claim, a random sample of 100 doctors is obtained. Out of these 100 doctors, 82 indicate that they recommend Aspirin. Is this claim accurate? Using $\alpha = 0.05$.

Decision Rule

(1) Null Hypothesis $(H_0) = P_0 = 0.9 \Rightarrow 9/10$

(2) Alternate Hypothesis $(H_1) = P_0 \neq 0.9$

(3)



(4) Calculate Z statistics for proportion

$$\left\{ Z_0 = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \right\}$$

$$\hat{p} = \frac{82}{100} = 0.82$$

2:36:53

Anuja N. Narayanan 2:37:52 / 3:00:24

Anuja N. Narayanan

Type here to search

3:35 PM 1/12/2023

Anuja N. Narayanan

33:48 | Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

GMT20210703 092721 Recording 1920x1080FS

Watch later | Share

Decision Rule

(1) Null Hypothesis $(H_0) = P_0 = 0.9 \Rightarrow 9/10$

(2) Alternate Hypothesis $(H_1) = P_0 \neq 0.9$

(3)

(4) Calculate Z statistics for proportion

$$\left\{ Z_0 = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \right\} \{\text{formula}\}$$

$$\hat{p} = \frac{82}{100} = 0.82$$

$P_0 = 0.90$

$n = 100$

$$\left\{ = \frac{0.82 - 0.7}{\sqrt{\frac{0.7(0.3)}{100}}} \right\}$$

$Z = -2.667$

$-2.667 < -1.96$

Accept the Alternate Hypothesis $= 0.12$

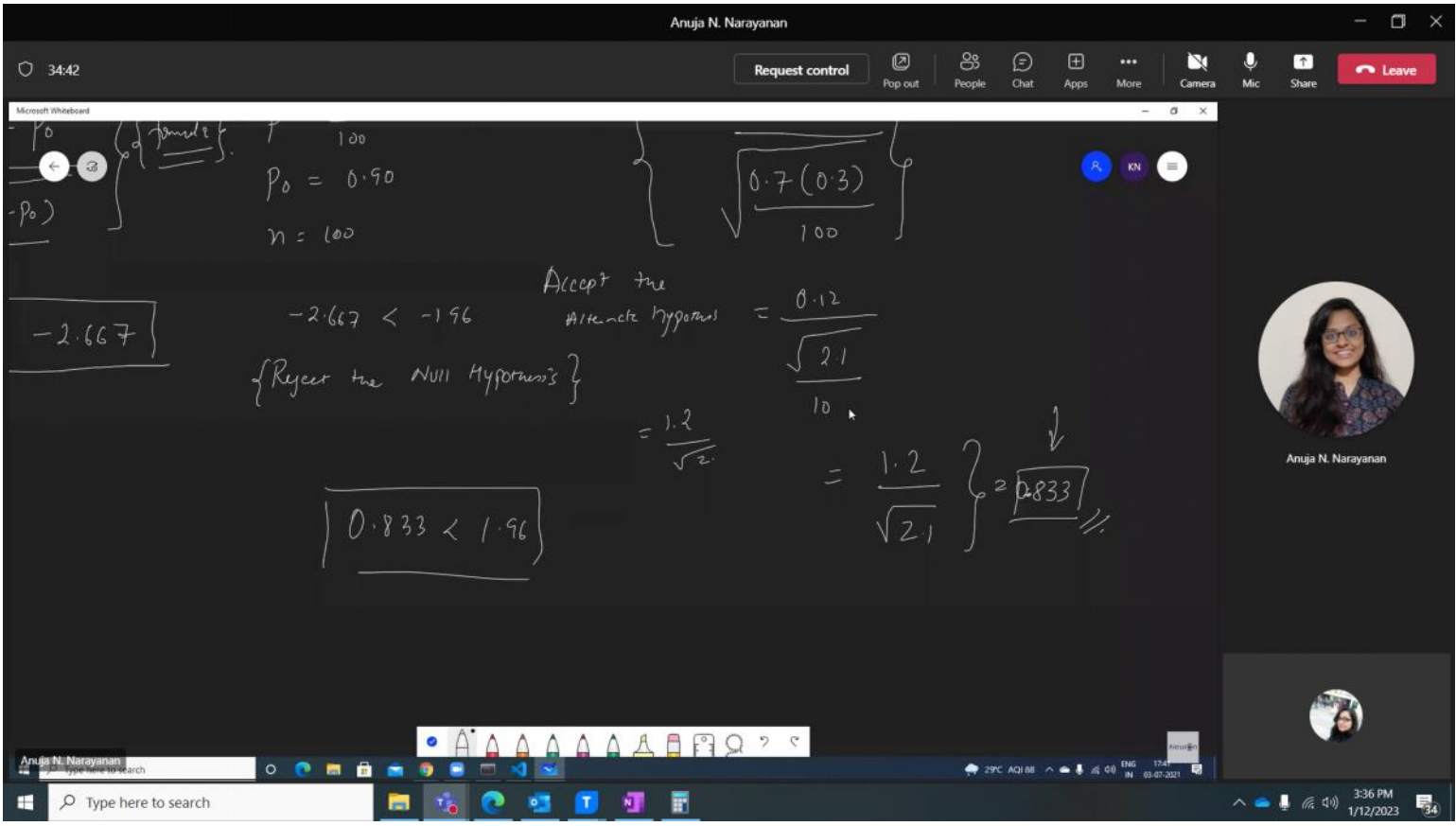
{Reject the Null Hypothesis}

Anuja N. Narayanan 2:41:26 / 3:00:24

Anuja N. Narayanan

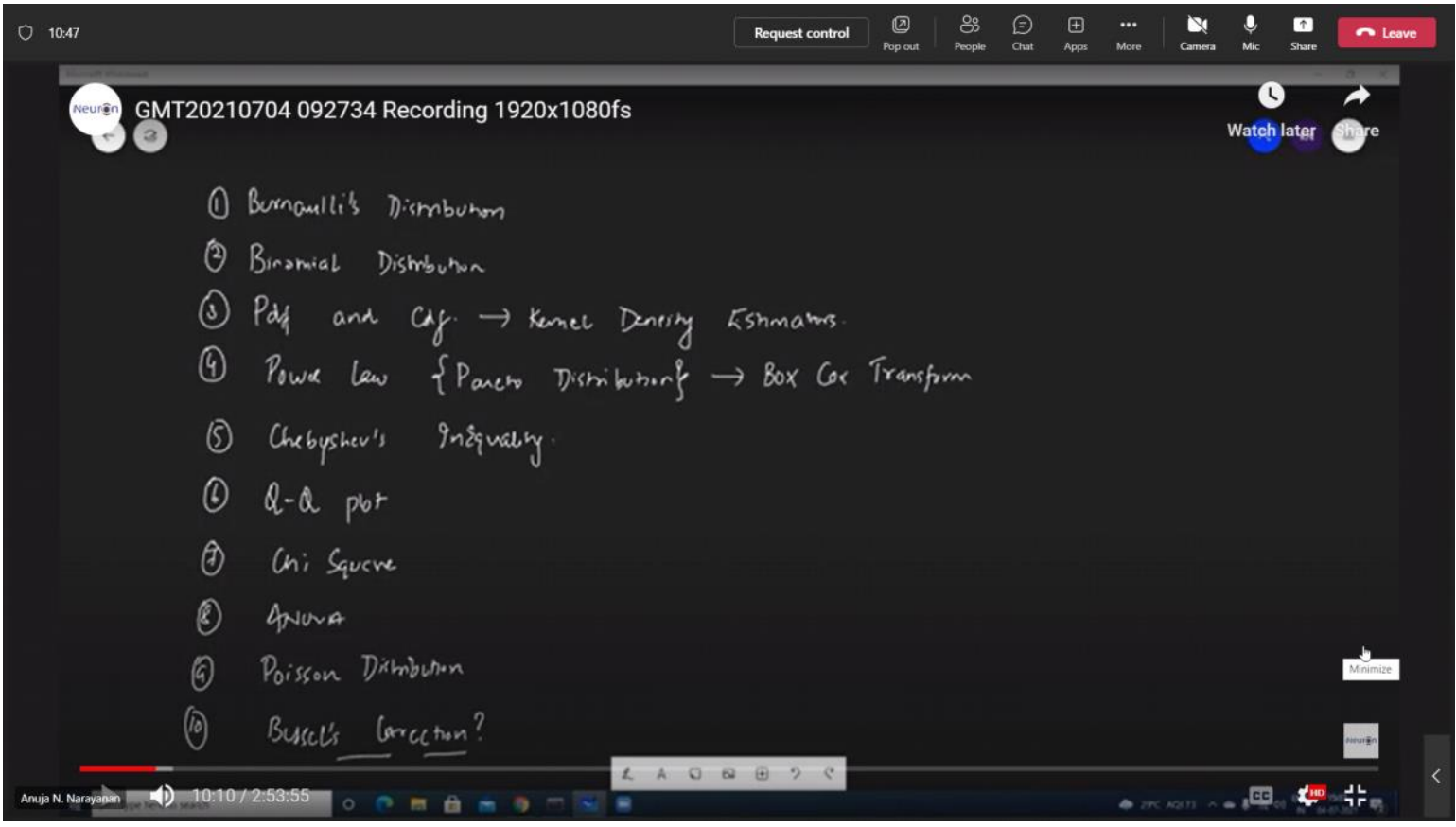
Type here to search

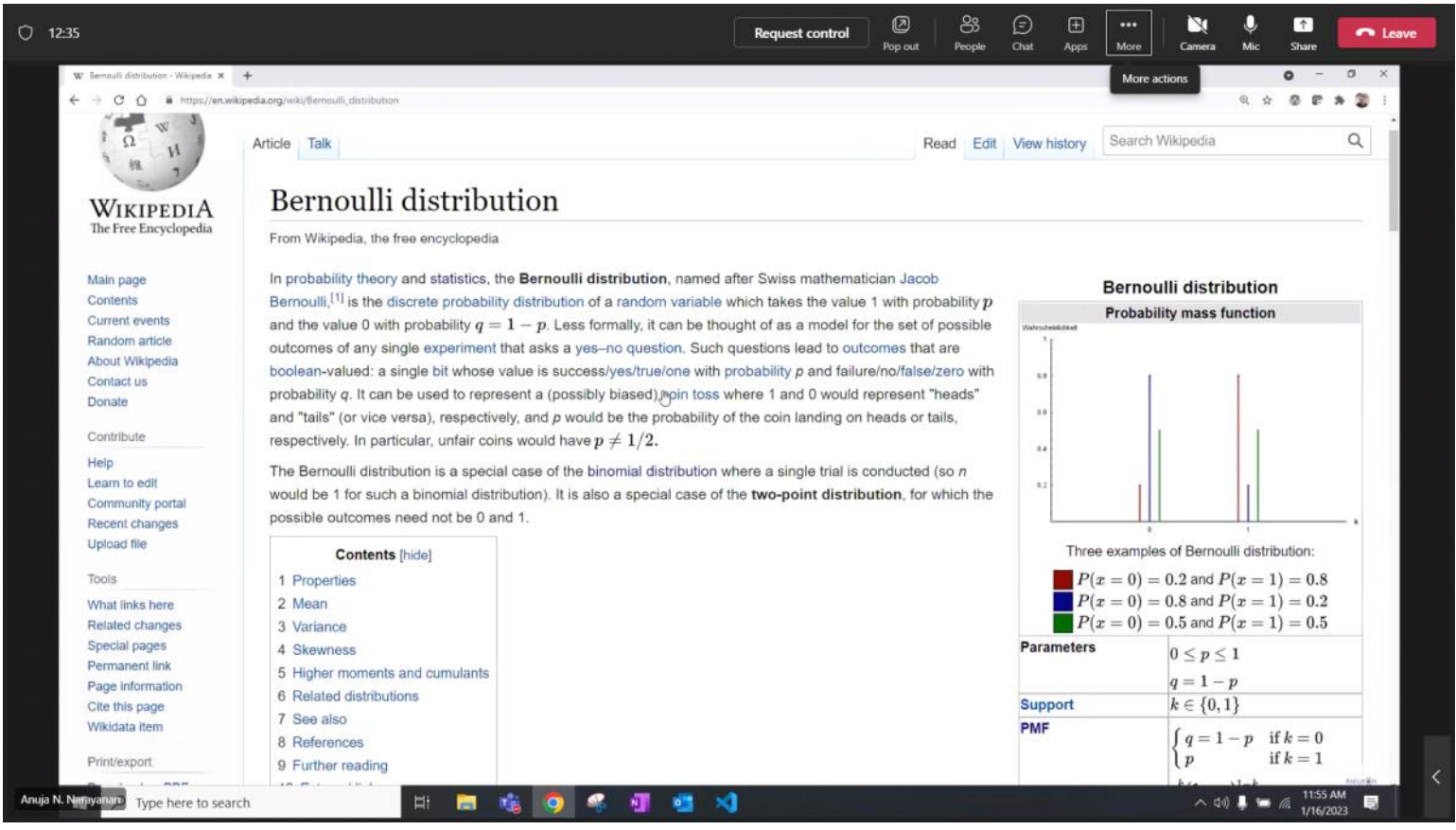
3:35 PM 1/12/2023


Anuja N. Narayanan
34:42
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
P_0
$P_0 = 0.90$
$n = 100$
$\sqrt{\frac{0.7(0.3)}{100}}$
-2.667
$-2.667 < -1.96$
Accept the Alternate Hypothesis
{Reject the NULL Hypothesis}
$= \frac{0.12}{\frac{\sqrt{2.1}}{10}}$
$= \frac{1.2}{\sqrt{2.1}}$
$= 0.833$
$0.833 < 1.96$
Anuja N. Narayanan
Type here to search
3:36 PM
1/12/2023

KN(lec-5)

Monday, January 16, 2023 11:52 AM





25:26

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

https://en.wikipedia.org/wiki/Bernoulli_distribution

WIKIPEDIA
The Free Encyclopedia

Bernoulli distribution

From Wikipedia, the free encyclopedia

Discrete ← pmf vs pdf → Continuous

In probability theory and statistics, the **Bernoulli distribution**, named after Swiss mathematician Jacob Bernoulli,[1] is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$. Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes–no question. Such questions lead to outcomes that are boolean-valued: a single bit whose value is success/yes/true/one with probability p and failure/no/false/zero with probability q. It can be used to represent a (possibly biased) coin toss where 1 and 0 would represent "heads" and "tails" (or vice versa), respectively, and p would be the probability of the coin landing on heads or tails, respectively. In particular, unfair coins would have $p \neq 1/2$.

The Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted (so n would be 1 for such a binomial distribution). It is also a special case of the **two-point distribution**, for which the possible outcomes need not be 0 and 1.

Contents [hide]

Bernoulli distribution

Probability mass function

Three examples of Bernoulli distribution:

- $P(x = 0) = 0.2$ and $P(x = 1) = 0.8$
- $P(x = 0) = 0.8$ and $P(x = 1) = 0.2$
- $P(x = 0) = 0.5$ and $P(x = 1) = 0.5$

Parameters	$0 \leq p \leq 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$ $p^k(1-p)^{1-k}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \end{cases}$

Anuja N. Narayanan | Type here to search | 12:08 PM 1/16/2023

26:57

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Unmute (Ctrl+Shift+M)

https://en.wikipedia.org/wiki/Bernoulli_distribution

WIKIPEDIA
The Free Encyclopedia

Bernoulli distribution

From Wikipedia, the free encyclopedia

Binary → 2 outcomes

$Pr(H=1) = 0.5 = p$ } { $p, 1-p = q$ }

In probability theory and statistics, the **Bernoulli distribution** named after Swiss mathematician Jacob Bernoulli,[1] is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$. Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes–no question. Such questions lead to outcomes that are boolean-valued: a single bit whose value is success/yes/true/one with probability p and failure/no/false/zero with probability q. It can be used to represent a (possibly biased) coin toss where 1 and 0 would represent "heads" and "tails" (or vice versa), respectively, and p would be the probability of the coin landing on heads or tails, respectively. In particular, unfair coins would have $p \neq 1/2$.

The Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted (so n would be 1 for such a binomial distribution). It is also a special case of the **two-point distribution**, for which the possible outcomes need not be 0 and 1.

$P(T=0) = 0.5 = 1-p$ }

Tossing a coin { H, T }

Unbiased coin

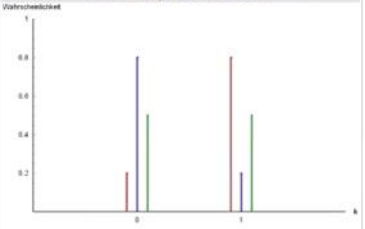
$P(H) = 1/2 \} = 0.5 = p$

$P(T) = 1/2 \} = 0.5 = 1-p$

Contents [hide]

Bernoulli distribution

Probability mass function



Three examples of Bernoulli distribution:

- $P(x = 0) = 0.2$ and $P(x = 1) = 0.8$
- $P(x = 0) = 0.8$ and $P(x = 1) = 0.2$
- $P(x = 0) = 0.5$ and $P(x = 1) = 0.5$

Parameters	$0 \leq p \leq 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$ $p^k(1-p)^{1-k}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \end{cases}$

Anuja N. Narayanan | Type here to search | 12:09 PM 1/16/2023

30:00 Request control Pop out People Chat Apps More Camera Mic Share Leave

Mute (Ctrl+Shift+M)

https://en.wikipedia.org/wiki/Bernoulli_distribution

Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
العربية
Deutsch
Español
Français
Bahasa Melayu
日本語
Português
Русский
中文
26 more
Edit links

possible outcomes need not be 0 and 1.

Contents [hide]

{K = 0, 1}
↳ Binary.

Properties [edit]

If X is a random variable with this distribution, then:

$$\Pr(X = 1) = p = 1 - \Pr(X = 0) = 1 - q.$$

The probability mass function f of this distribution, over possible outcomes k, is

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$ [2]

This can also be expressed as

$$f(k;p) = p^k(1-p)^{1-k} \quad \text{for } k \in \{0,1\}$$

or as

$$f(k;p) = pk + (1-p)(1-k) \quad \text{for } k \in \{0,1\}.$$

The Bernoulli distribution is a special case of the binomial distribution with $n = 1$.[3]

Three examples of Bernoulli distribution:
$P(x = 0) = 0.2$ and $P(x = 1) = 0.8$
$P(x = 0) = 0.8$ and $P(x = 1) = 0.2$
$P(x = 0) = 0.5$ and $P(x = 1) = 0.5$

Parameters	$0 \le p \le 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$ $p^k(1-p)^{1-k}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \le k < 1 \\ 1 & \text{if } k \ge 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Mode	$\begin{cases} 0 & \text{if } p < 1/2 \\ 0, 1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Variance	$p(1-p) = pq$
Skewness	$\frac{q-p}{\sqrt{pq}}$
Ex. kurtosis	$\frac{1-6pq}{pq}$
Entropy	$-q \ln q - p \ln p$

Anuja N. Narayanan Type here to search 12:12 PM 1/16/2023

33:04 Request control Pop out People Chat Apps More Camera Mic Share Leave

Unmute (Ctrl+Shift+M)

https://en.wikipedia.org/wiki/Binomial_distribution

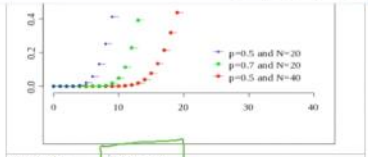
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
العربية
Español
★ Français
हिन्दी
Bahasa Indonesia
Русский
தமிழ்
اردو
中文
39 more
Edit links

B(n, p)
← n, p
Bernoullis
$\binom{n}{k} p^k q^{n-k}$



Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \ldots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0, 1, \ldots, n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	npq
Skewness	$\frac{q-p}{\sqrt{npq}}$
Ex. kurtosis	$\frac{1-6pq}{npq}$
Entropy	$\frac{1}{2}\log_2(2\pi enpq) + O\left(\frac{1}{n}\right)$ in shannons. For nats, use the natural log in the log.
MGF	$(q + pe^t)^n$

Anuja N. Narayanan Type here to search 12:15 PM 1/16/2023

33:44

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

KN

Estimators.

Box Cox Trar.

Binomial distribution

(i) An experiment is a binomial experiment if

(A) It is repeated for fixed number of times.

(K) The trials are independent

(A) Trails have 2 mutually exclusive outcomes, either Success or failure.

Anuja N. Narayanan | Type here to search | 12:16 PM 1/16/2023

35:39

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

Mute (Ctrl+Shift+M)

KN

failure.

{not biased} (A) The probability of success is same for all trials.

(K) In the recent survey, it was found that 85% of housholds in the United States have a High Speed Internet. If you take the sample of 18 households, what is the probability that exactly 15 will have High Speed Internet?

Anuja N. Narayanan | Type here to search | 12:18 PM 1/16/2023

38:29

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

Mute (Ctrl+Shift+M)

in the United States have a High Speed Internet. If you take the sample of 18 households, what is the probability that exactly 15 will have High Speed Internet?

Ans

$$P(x=15) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x}$$

$x = 15$

$n = 18$

$p = 85\% = 0.85$

$$= {}^{18}C_{15} (0.85)^{15} (0.15)^{18-15}$$

$$= {}^{18}C_{15} (0.85)^{15} (0.15)^{3}$$

Anuja N. Narayanan

Type here to search

12:21 PM
1/16/2023

Anuja N. Narayanan

08:25

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

interview → Chebyshev's Inequality

$X \sim G.D$

$$\mu - \sigma \le X \le \mu + \sigma = 68\%$$

$$\mu - 2\sigma \le X \le \mu + 2\sigma = 95\%$$

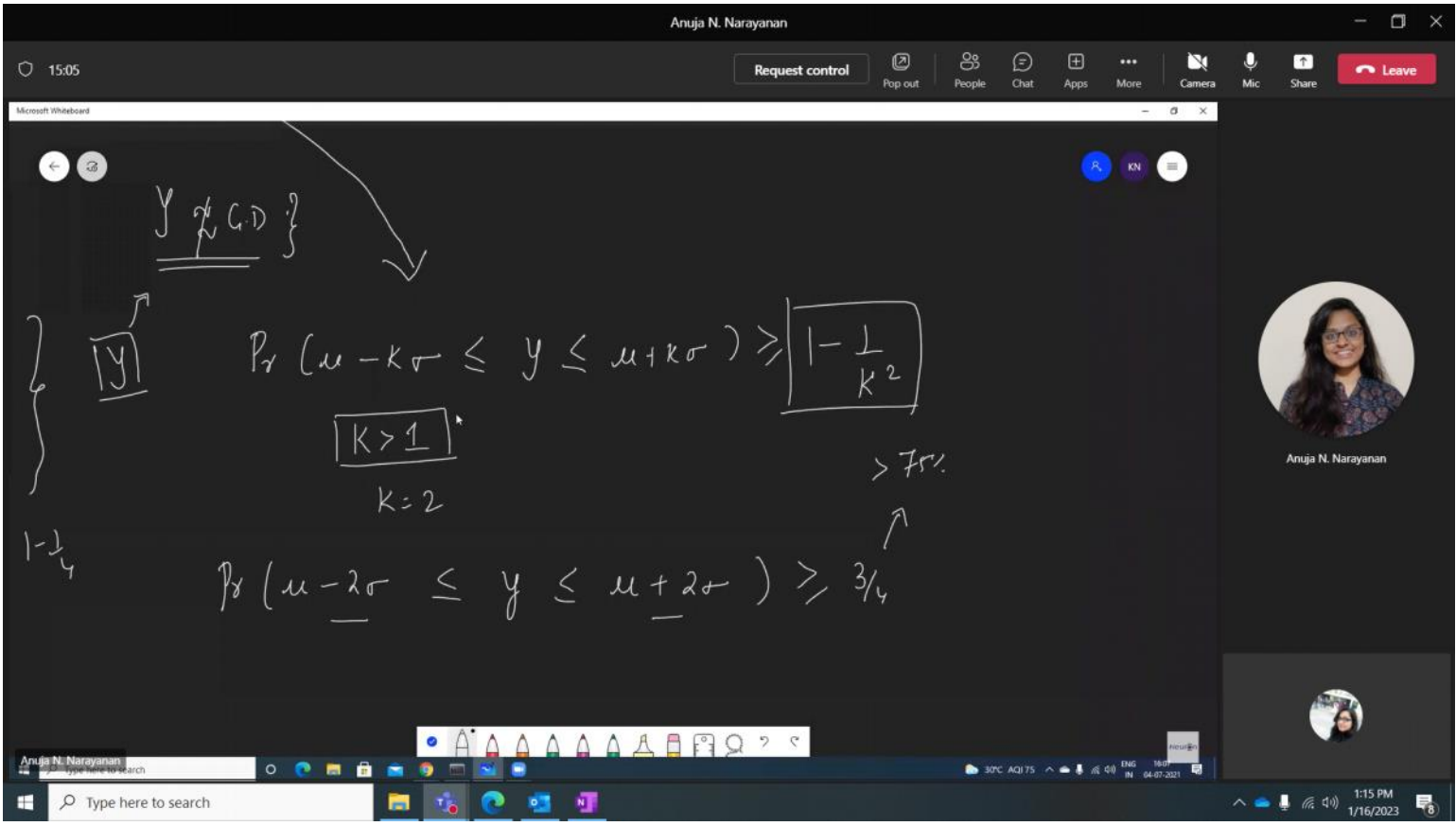
$$\mu - 3\sigma \le X \le \mu + 3\sigma = 99.7\%$$

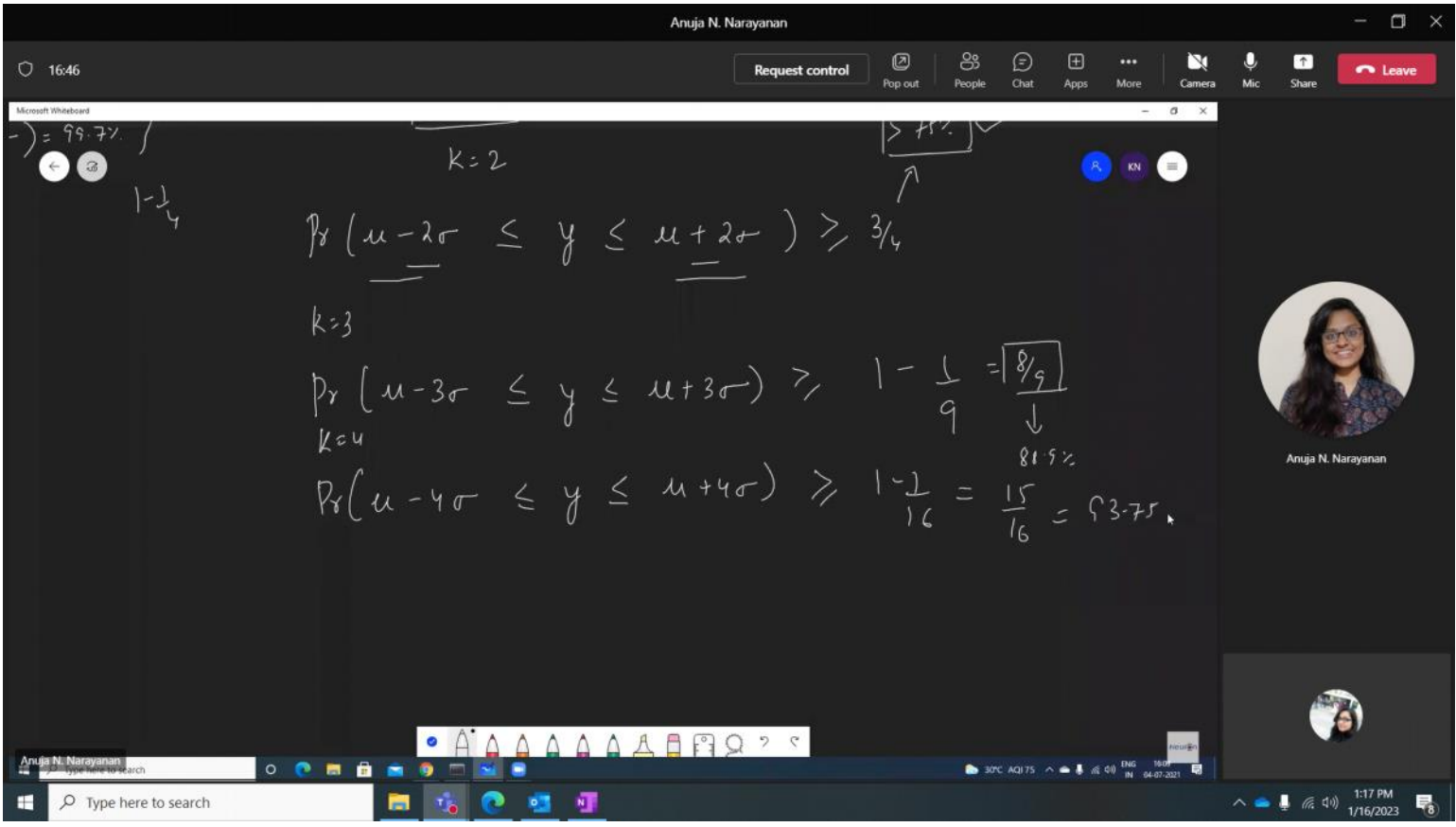
Anuja N. Narayanan

Anuja N. Narayanan

Type here to search

1:09 PM
1/16/2023





Power law distribution

Anuja N. Narayanan

24:09

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Power law

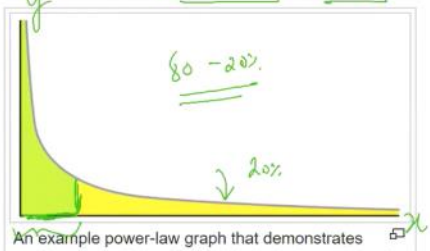
From Wikipedia, the free encyclopedia

IPL

Not to be confused with Force (law). For other uses, see Power (disambiguation)

Power Law relationship

In statistics, a **power law** is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another. For instance, considering the area of a square in terms of the length of its side, if the length is doubled, the area is multiplied by a factor of four.[1]



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

Contents [hide]

1 Empirical examples
2 Properties
  2.1 Scale invariance
  2.2 Lack of well-defined average value
  2.3 Universality
3 Power-law functions
  3.1 Examples

Anuja N. Narayanan

Type here to search

1:24 PM 1/16/2023

Anuja N. Narayanan

27:08

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

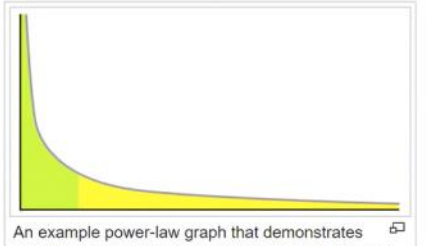
Power law

From Wikipedia, the free encyclopedia

Pareto distribution

Not to be confused with Force (law). For other uses, see Power (disambiguation)

In statistics, a **power law** is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another. For instance, considering the area of a square in terms of the length of its side, if the length is doubled, the area is multiplied by a factor of four.[1]



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

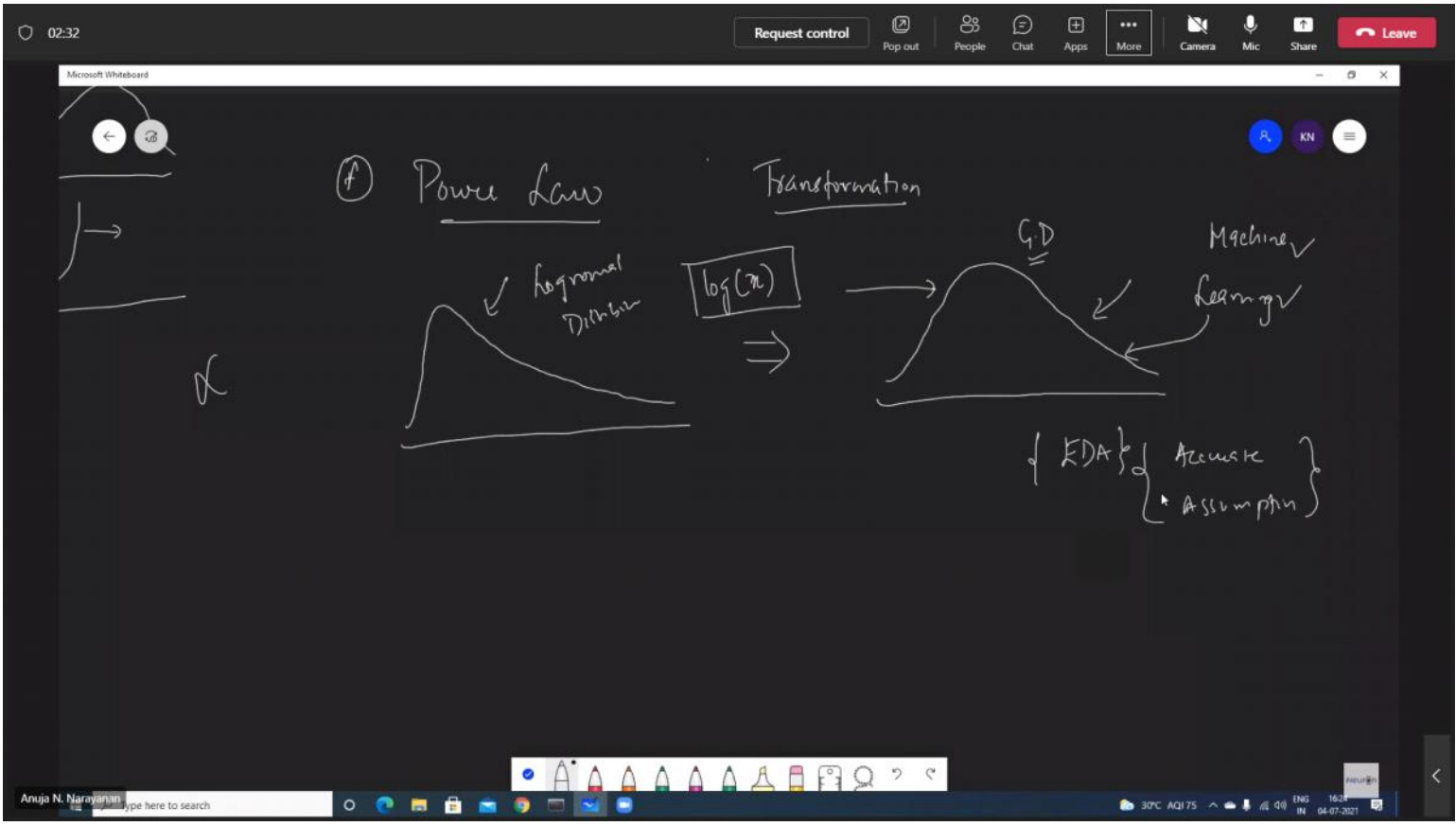
Contents [hide]

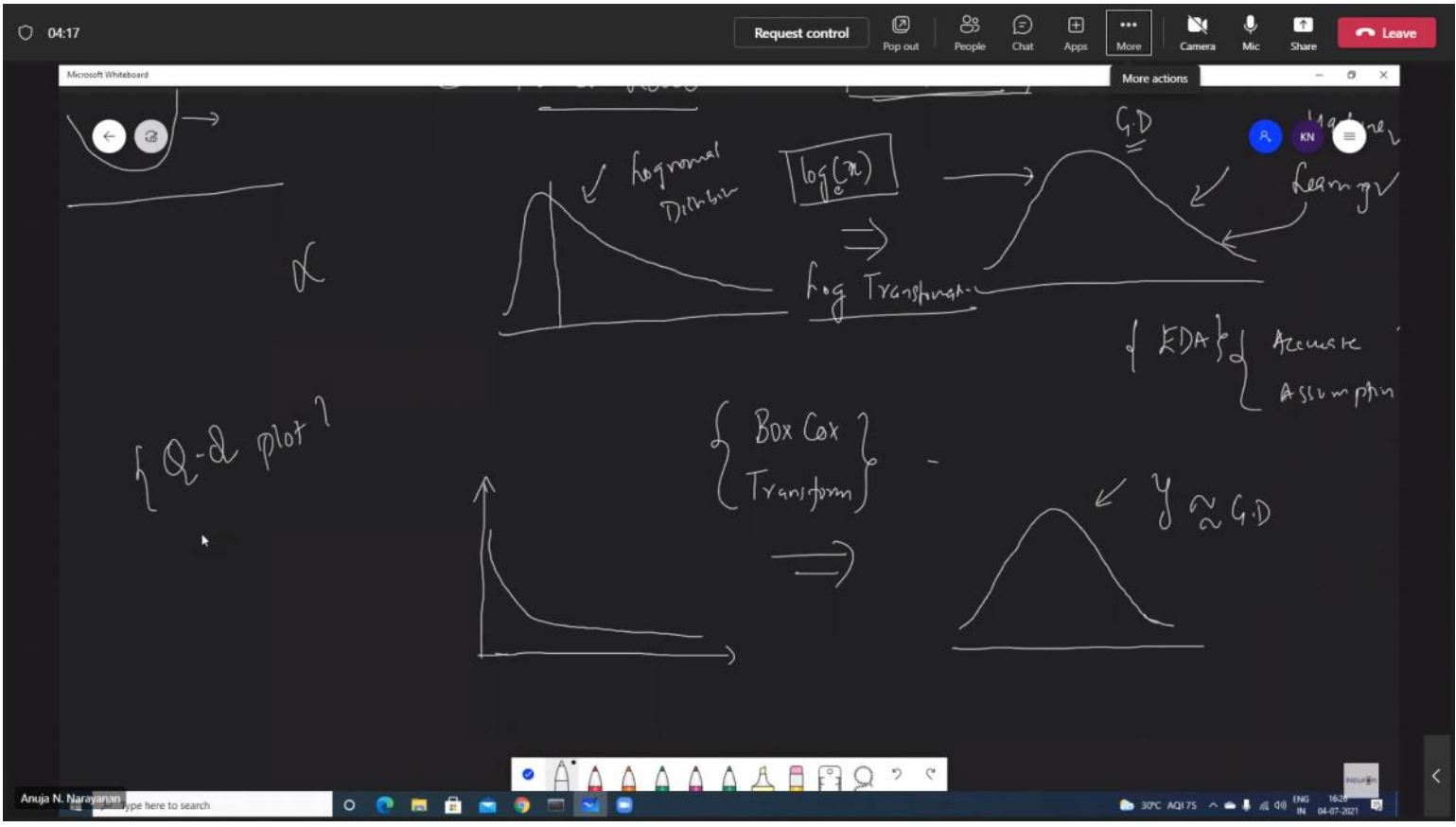
1 Empirical examples
2 Properties
  2.1 Scale invariance
  2.2 Lack of well-defined average value

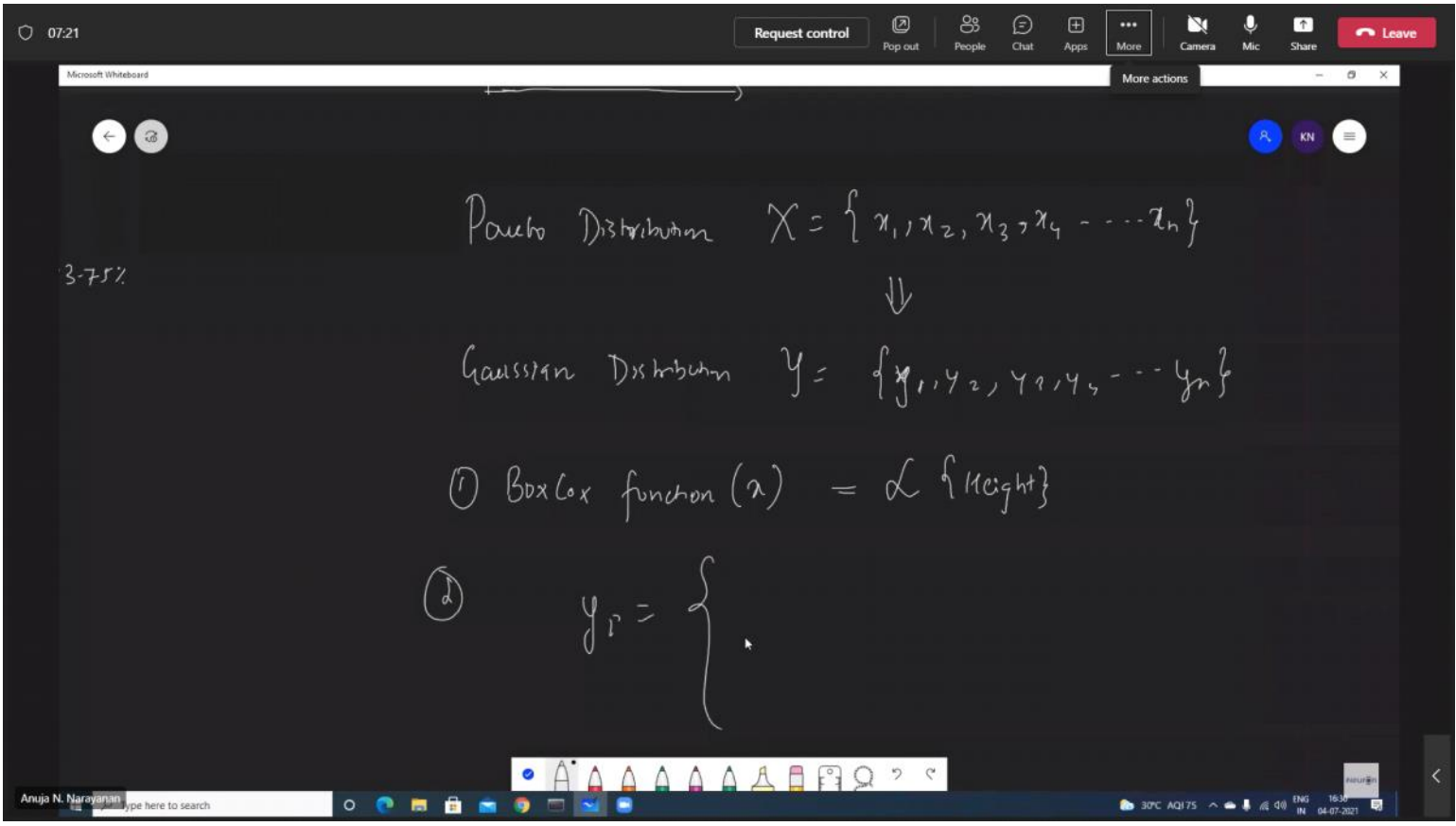
Anuja N. Narayanan

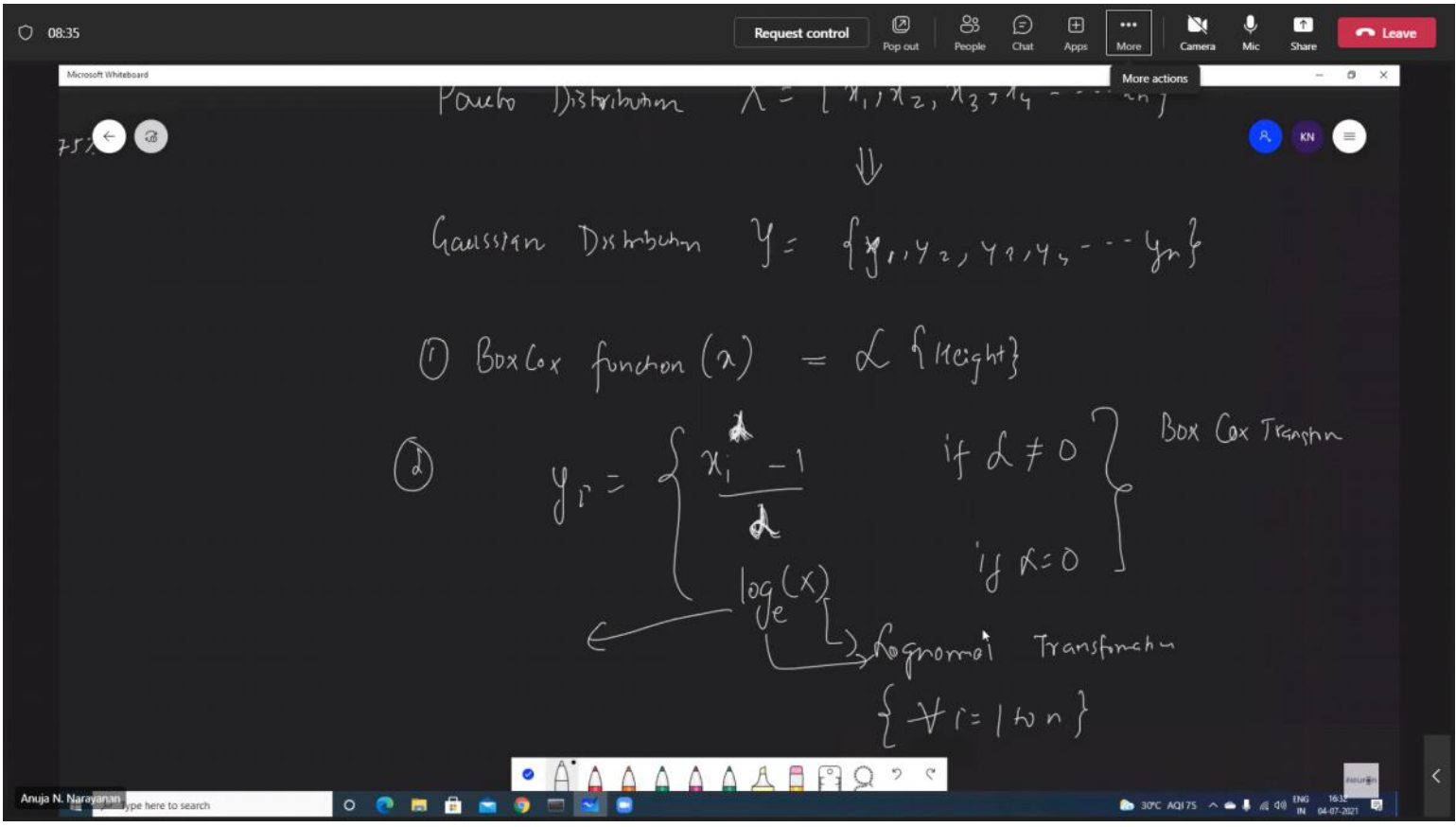
Type here to search

1:27 PM 1/16/2023


02:32
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
KN
Power Law Transformation
Lognormal Distribn
log(x)
G.D
Machine Learning
EDA
Accurate
Assumption
Anuja N. Narayanan


04:17
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
KN
G.D
Lognormal Distribn
log(x)
log Transformation
EDA
Accurate
Assumption
Q-Q plot
Box Cox Transform
y ≈ G.D
Anuja N. Narayanan


07:21
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
KN
Pareto Distribution $X = \{x_1, x_2, x_3, x_4 \cdots x_n\}$
3-75%
Gaussian Distribution $Y = \{y_1, y_2, y_3, y_4 \cdots y_n\}$
① Box Cox function (λ) = α {Weight}
②
Anuja N. Narayanan
Type here to search
30°C AQI 75
ENG IN
16:30
04-07-2021


08:35
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
KN
75%
Gaussian Distribution $Y = \{y_1, y_2, y_3, y_4 \cdots y_n\}$
① Box Cox function (λ) = α {Weight}
②
$y_i = \begin{cases} \frac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log_e(x) & \text{if } \lambda = 0 \end{cases}$
Box Cox Transform
Lognormal Transformation
$\{\forall i = 1 \text{ to } n\}$
Anuja N. Narayanan
Type here to search
30°C AQI 75
ENG IN
16:35
04-07-2021

25:38

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Untitled18 Last Checkpoint: 11 minutes ago (unsaved changes)

Logout

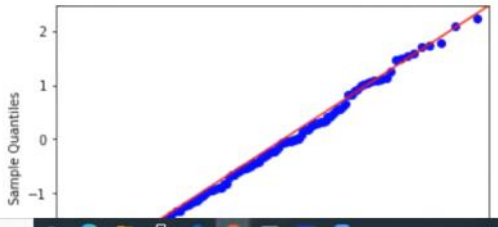
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

In [1]:

```
import numpy as np
import statsmodels.api as sm
import pylab as py

# np.random generates different random numbers
# whenever the code is executed
# Note: When you execute the same code
# the graph look different than shown below.

# Random data points generated
data_points = np.random.normal(0, 1, 100)

sm.qqplot(data_points, line ='45')
py.show()
```



Anuja N. Narayanan

26:50

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Untitled18 Last Checkpoint: 13 minutes ago (unsaved changes)

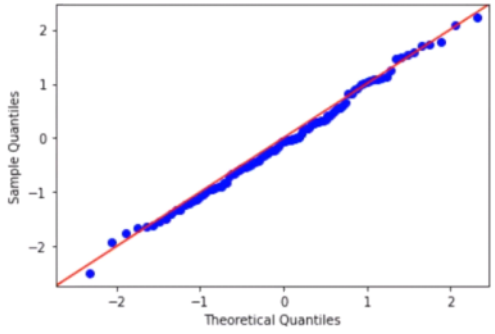
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

```
sm.qqplot(data_points, line ='45')
py.show()
```

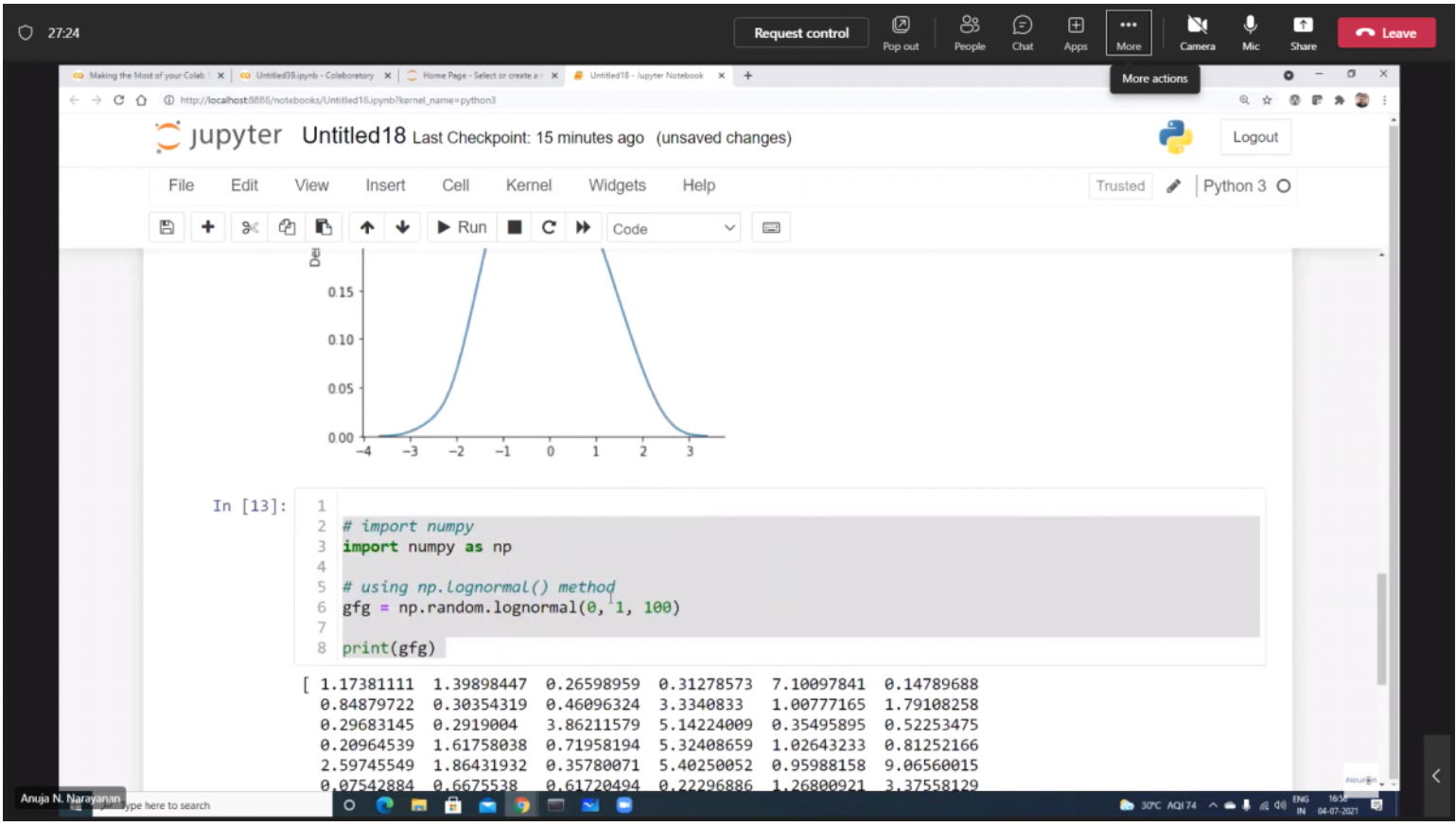


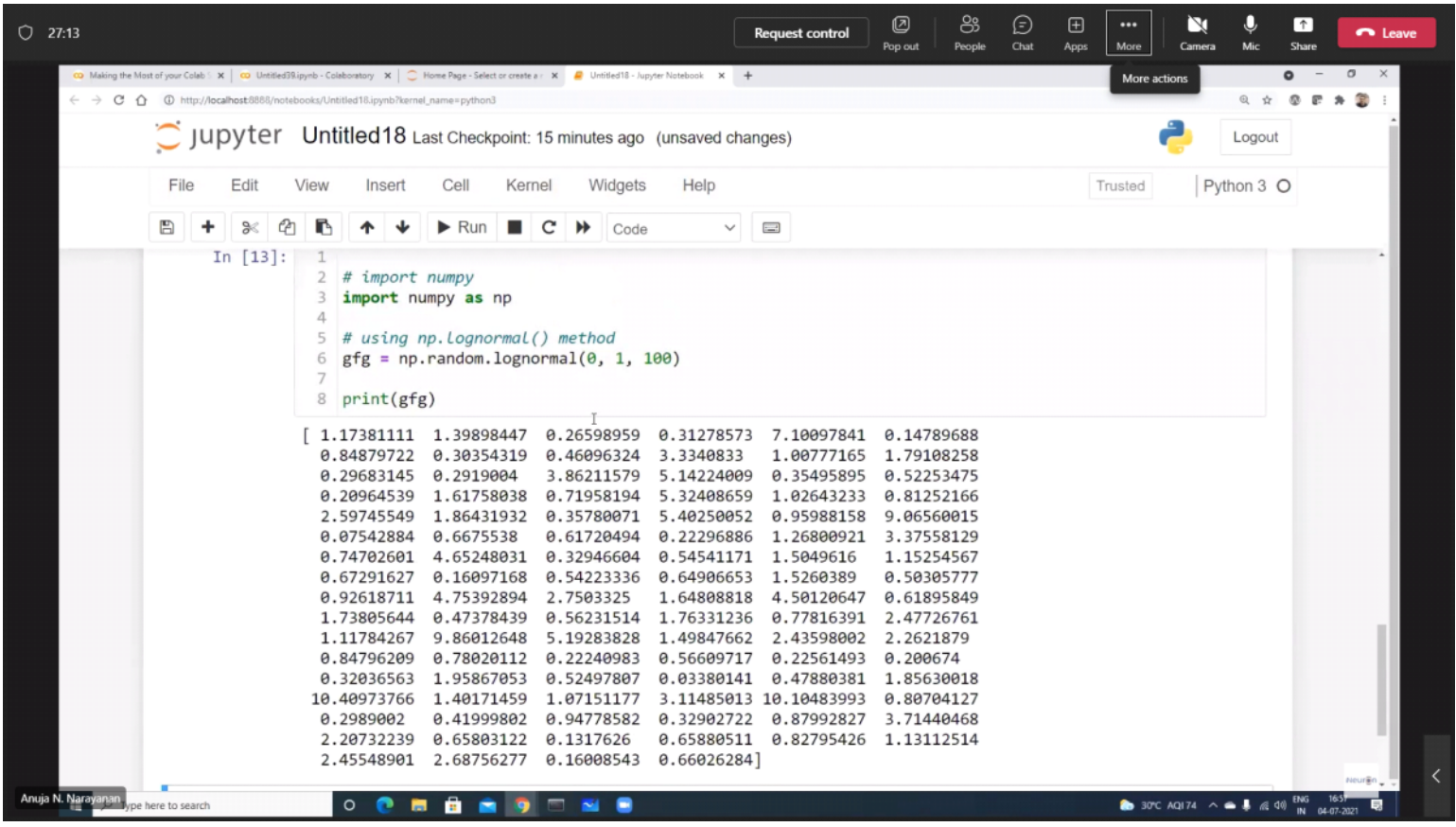
In [6]:

```
import seaborn as sns
sns.displot(data_points,kind="kde")
```

Out[6]: <seaborn.axisgrid.FacetGrid at 0x1f8dbca55c8>

Anuja N. Narayanan


27:24
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
jupyter Untitled18 Last Checkpoint: 15 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [13]:
# import numpy
import numpy as np
# using np.lognormal() method
gfg = np.random.lognormal(0, 1, 100)
print(gfg)
[1.17381111 1.39898447 0.26598959 0.31278573 7.10097841 0.14789688
0.84879722 0.30354319 0.46096324 3.3340833 1.00777165 1.79108258
0.29683145 0.2919004 3.86211579 5.14224009 0.35495895 0.52253475
0.20964539 1.61758038 0.71958194 5.32408659 1.02643233 0.81252166
2.59745549 1.86431932 0.35780071 5.40250052 0.95988158 9.06560015
0.07542884 0.6675538 0.61720494 0.22296886 1.26800921 3.37558129
Anuja N. Narayanan


27:13
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
jupyter Untitled18 Last Checkpoint: 15 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [13]:
# import numpy
import numpy as np
# using np.lognormal() method
gfg = np.random.lognormal(0, 1, 100)
print(gfg)
[1.17381111 1.39898447 0.26598959 0.31278573 7.10097841 0.14789688
0.84879722 0.30354319 0.46096324 3.3340833 1.00777165 1.79108258
0.29683145 0.2919004 3.86211579 5.14224009 0.35495895 0.52253475
0.20964539 1.61758038 0.71958194 5.32408659 1.02643233 0.81252166
2.59745549 1.86431932 0.35780071 5.40250052 0.95988158 9.06560015
0.07542884 0.6675538 0.61720494 0.22296886 1.26800921 3.37558129
0.74702601 4.65248031 0.32946604 0.54541171 1.5049616 1.15254567
0.67291627 0.16097168 0.54223336 0.64906653 1.5260389 0.50305777
0.92618711 4.75392894 2.7503325 1.64808818 4.50120647 0.61895849
1.73805644 0.47378439 0.56231514 1.76331236 0.77816391 2.47726761
1.11784267 9.86012648 5.19283828 1.49847662 2.43598002 2.2621879
0.84796209 0.78020112 0.22240983 0.56609717 0.22561493 0.200674
0.32036563 1.95867053 0.52497807 0.03380141 0.47880381 1.85630018
10.40973766 1.40171459 1.07151177 3.11485013 10.10483993 0.80704127
0.2989002 0.41999802 0.94778582 0.32902722 0.87992827 3.71440468
2.20732239 0.65803122 0.1317626 0.65880511 0.82795426 1.13112514
2.45548901 2.68756277 0.16008543 0.66026284]
Anuja N. Narayanan

29:57

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

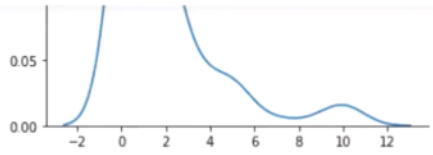
jupyter Untitled18 Last Checkpoint: 17 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code



In [13]:

```
# import numpy
import numpy as np

# using np.lognormal() method
gfg = np.random.lognormal(0, 1, 100)

print(gfg)
```

```
[ 1.17381111  1.39898447  0.26598959  0.31278573  7.10097841  0.14789688
  0.84879722  0.30354319  0.46096324  3.3340833   1.00777165  1.79108258
  0.29683145  0.2919004   3.86211579  5.14224009  0.35495895  0.52253475
  0.20964539  1.61758038  0.71958194  5.32408659  1.02643233  0.81252166
  2.59745549  1.86431932  0.35780071  5.40250052  0.95988158  9.06560015
  0.07542884  0.6675538   0.61720494  0.22296886  1.26800921  3.37558129
  0.74702601  4.65248031  0.32946604  0.54541171  1.5049616   1.15254567
  0.67291627  0.16097168  0.54223336  0.64906653  1.5260389   0.50305777
  0.92618711  4.75392894  2.7503325   1.64808818  4.50120647  0.61895849
```

Anuja N. Narayanan

28:07

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Untitled18 Last Checkpoint: 16 minutes ago (unsaved changes)

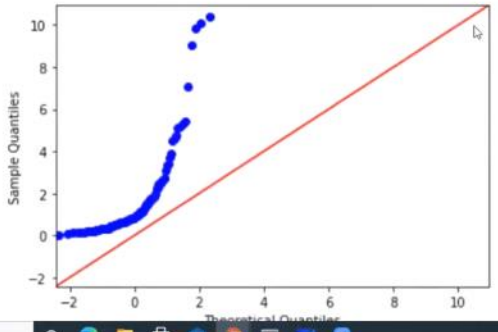
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

```
import pylab as py

# np.random generates different random numbers
# whenever the code is executed
# Note: When you execute the same code
# the graph look different than shown below.

# Random data points generated

sm.qqplot(gfg, line ='45')
py.show()
```



Anuja N. Narayanan

30:28

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Untitled18 Last Checkpoint: 17 minutes ago (unsaved changes)

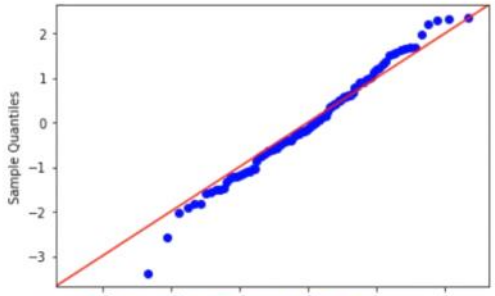
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

```
import statsmodels.api as sm
import pylab as py

# np.random generates different random numbers
# whenever the code is executed
# Note: When you execute the same code
# the graph look different than shown below.

# Random data points generated

sm.qqplot(np.log(gfg), line ='45')
py.show()
```



Anuja N. Narayanan

33:23

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Transformation Last Checkpoint: 28 minutes ago (autosaved)

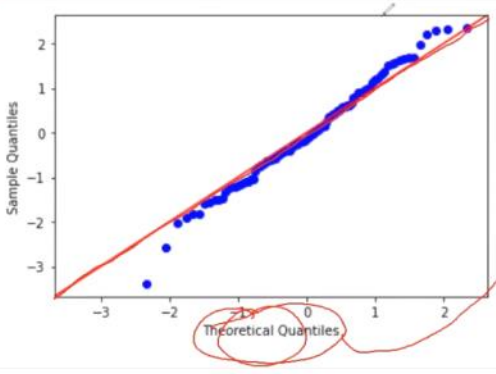
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

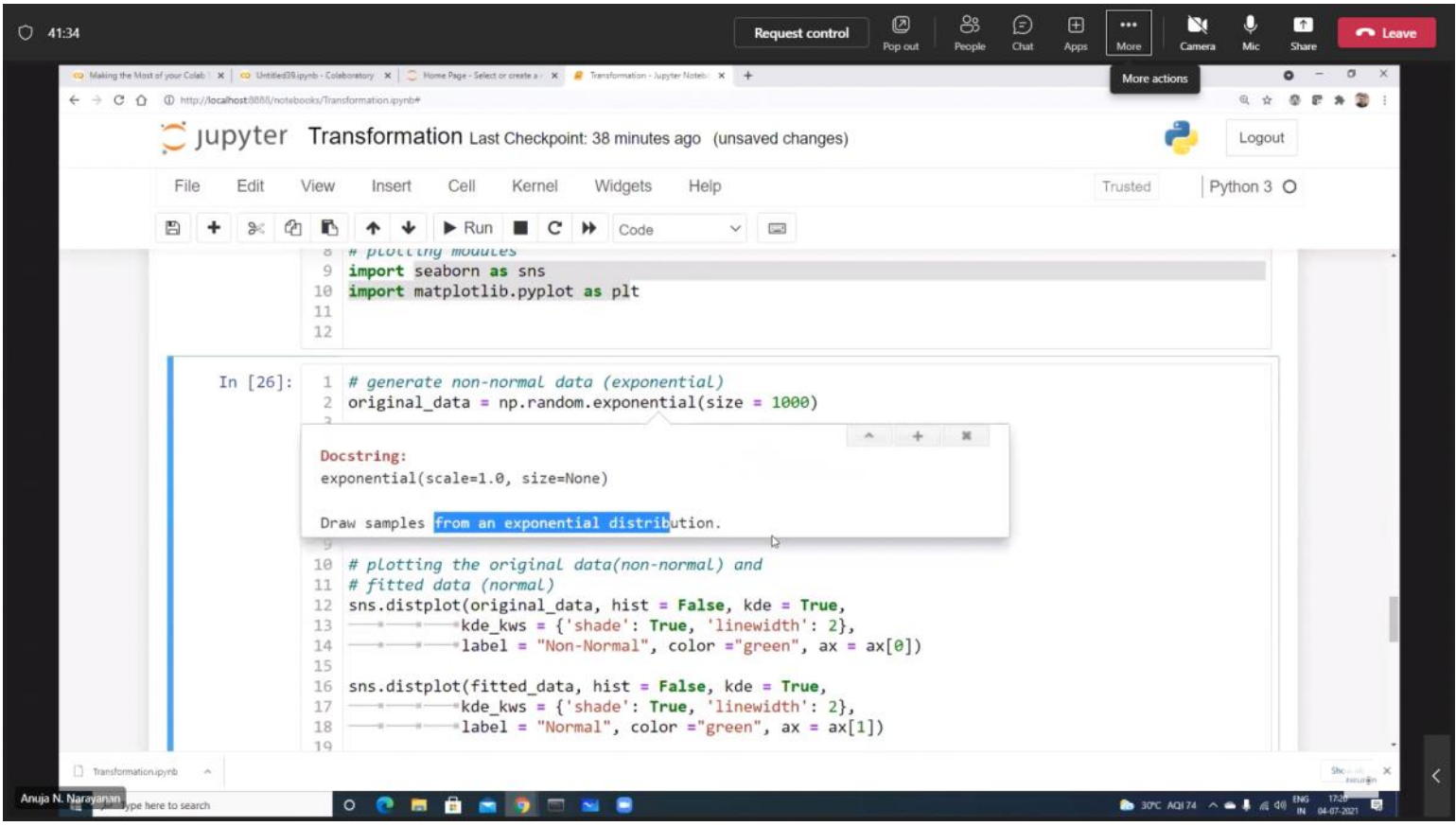
Run | Code

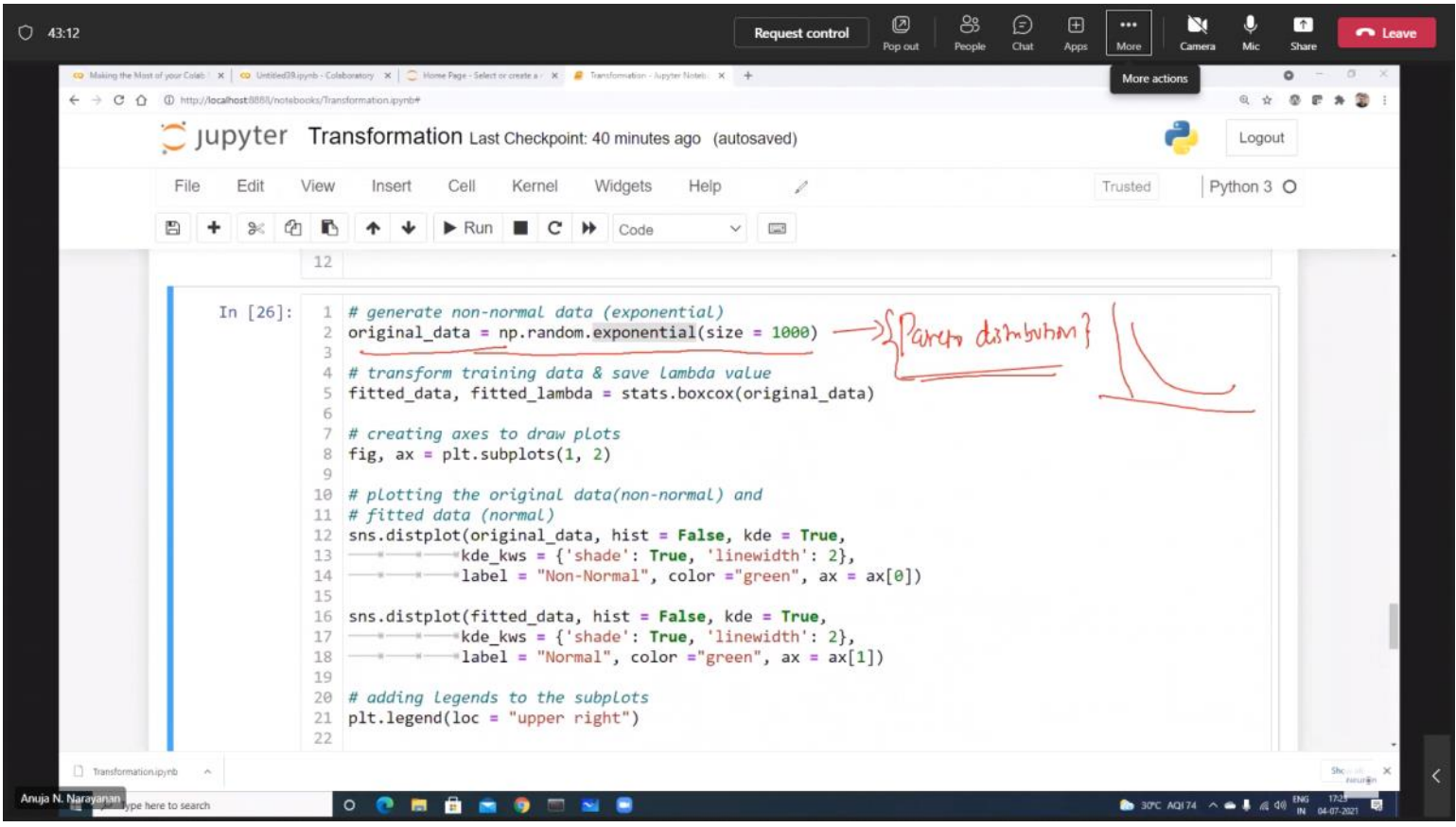
```
# the graph look different than shown below.

# Random data points generated
# using np.lognormal() method
gfg = np.random.lognormal(0, 1, 100)
sm.qqplot(np.log(gfg), line ='45')
py.show()
```



Gaussian Distributed

Anuja N. Narayanan


41:34
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
jupyter Transformation Last Checkpoint: 38 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
# plotting modules
import seaborn as sns
import matplotlib.pyplot as plt
In [26]:
# generate non-normal data (exponential)
original_data = np.random.exponential(size = 1000)
Docstring:
exponential(scale=1.0, size=None)
Draw samples from an exponential distribution.
plotting the original data(non-normal) and
fitted data (normal)
sns.distplot(original_data, hist = False, kde = True,
kde_kws = {'shade': True, 'linewidth': 2},
label = "Non-Normal", color ="green", ax = ax[0])
sns.distplot(fitted_data, hist = False, kde = True,
kde_kws = {'shade': True, 'linewidth': 2},
label = "Normal", color ="green", ax = ax[1])
Transformation.ipynb
Anuja N. Narayanan
Type here to search
30°C AQI 74
ENG IN
17:28
04-07-2021


43:12
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
jupyter Transformation Last Checkpoint: 40 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [26]:
# generate non-normal data (exponential)
original_data = np.random.exponential(size = 1000)
Pareto distribution?
transform training data & save lambda value
fitted_data, fitted_lambda = stats.boxcox(original_data)
creating axes to draw plots
fig, ax = plt.subplots(1, 2)
plotting the original data(non-normal) and
fitted data (normal)
sns.distplot(original_data, hist = False, kde = True,
kde_kws = {'shade': True, 'linewidth': 2},
label = "Non-Normal", color ="green", ax = ax[0])
sns.distplot(fitted_data, hist = False, kde = True,
kde_kws = {'shade': True, 'linewidth': 2},
label = "Normal", color ="green", ax = ax[1])
# adding legends to the subplots
plt.legend(loc = "upper right")
Transformation.ipynb
Anuja N. Narayanan
Type here to search
30°C AQI 74
ENG IN
17:29
04-07-2021

43:56

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Transformation Last Checkpoint: 41 minutes ago (unsaved changes)

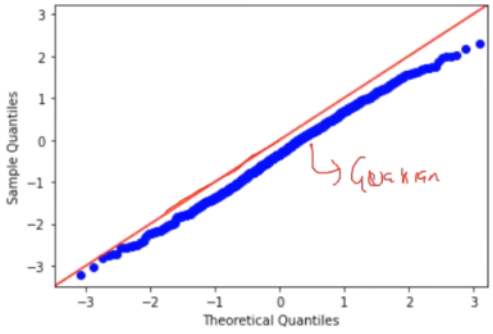
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

```
gfg = np.random.lognormal(0, 1, 100)
sm.qqplot(fitted_data, line ='45')
py.show()
```



In [ ]:

```
1
```

Transformation.ipynb

Anuja N. Narayanan

48:52

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

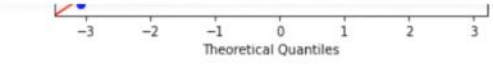
jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

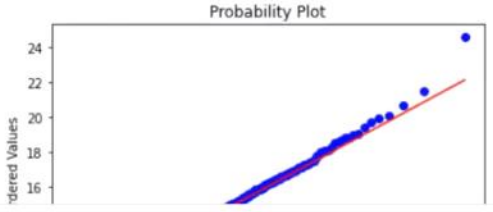
Trusted | Python 3

Run | Code



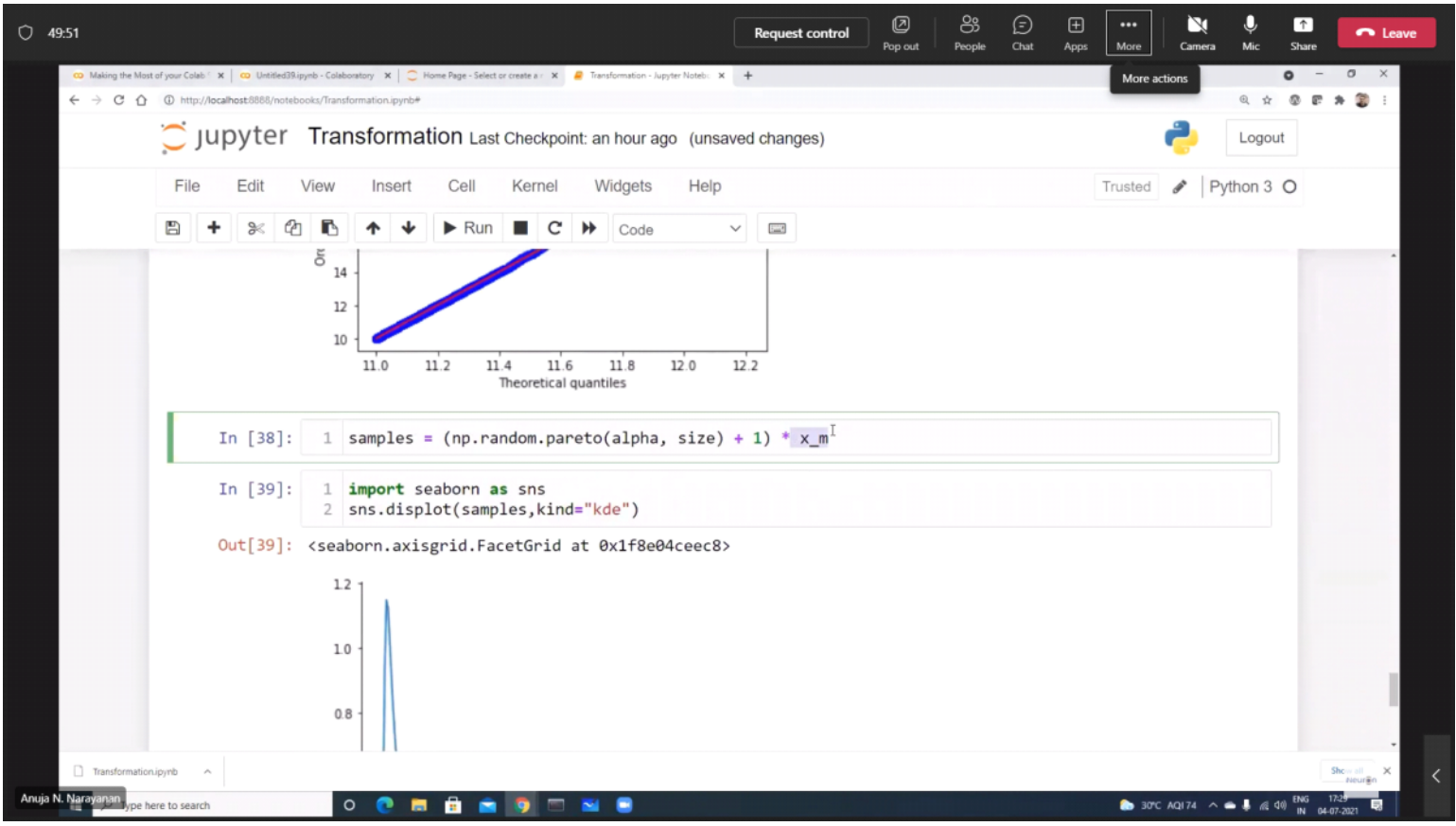
In [35]:

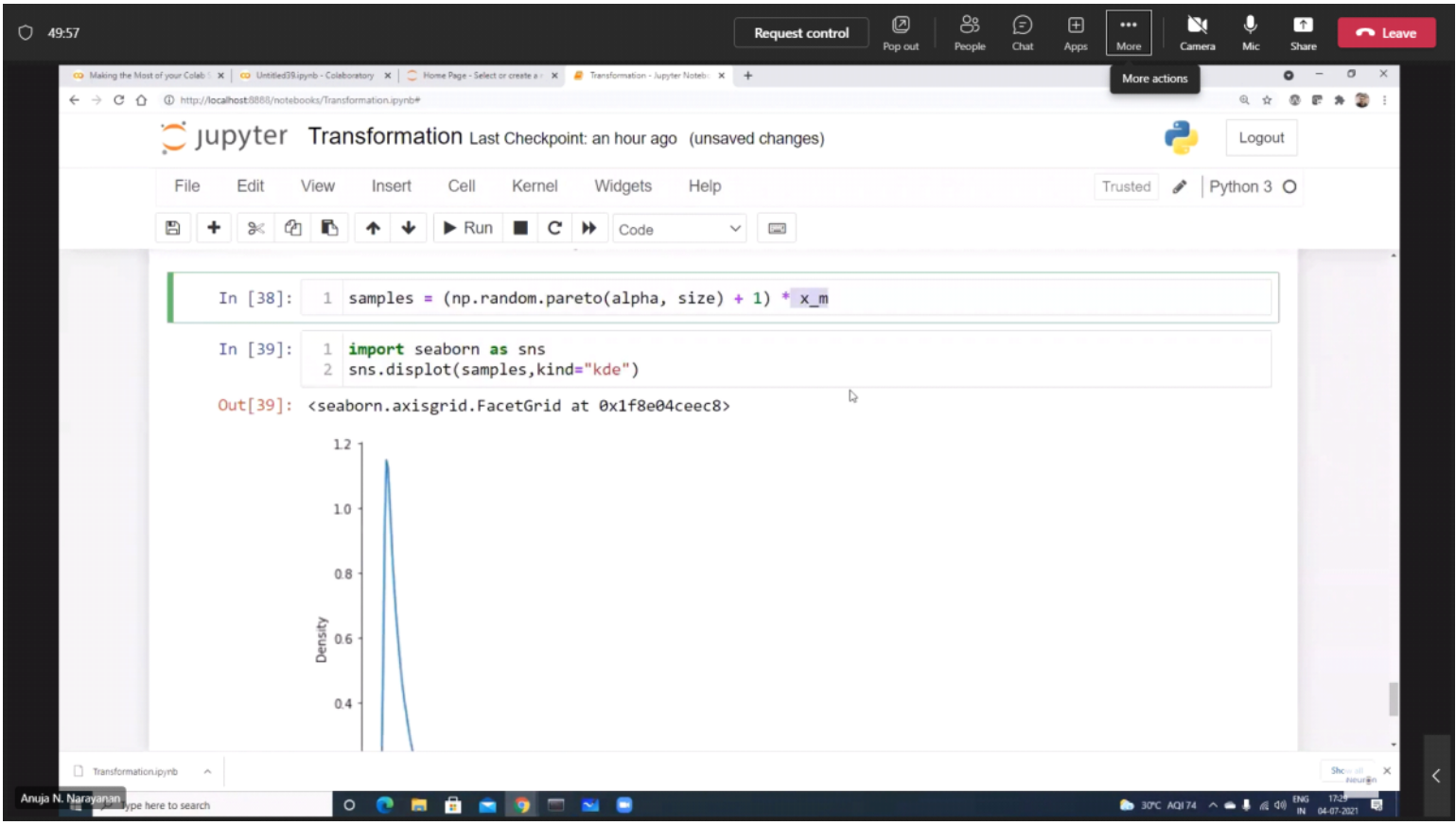
```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pylab
x_m = 10
alpha = 15
size = 100000 #the size of the sample(no. of random samples)
samples = (np.random.pareto(alpha, size) + 1) * x_m
stats.probplot(samples, dist='pareto', sparams=(15, 10), plot=pylab)
plt.show()
```



Transformation.ipynb

Anuja N. Narayanan


49:51
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
14
12
10
11.0
11.2
11.4
11.6
11.8
12.0
12.2
Theoretical quantiles
In [38]: samples = (np.random.pareto(alpha, size) + 1) * x_m
In [39]: import seaborn as sns
sns.displot(samples,kind="kde")
Out[39]: <seaborn.axisgrid.FacetGrid at 0x1f8e04ceec8>
1.2
1.0
0.8
Anuja N. Narayanan


49:57
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [38]: samples = (np.random.pareto(alpha, size) + 1) * x_m
In [39]: import seaborn as sns
sns.displot(samples,kind="kde")
Out[39]: <seaborn.axisgrid.FacetGrid at 0x1f8e04ceec8>
1.2
1.0
0.8
0.6
0.4
Density
Anuja N. Narayanan

50:19

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)

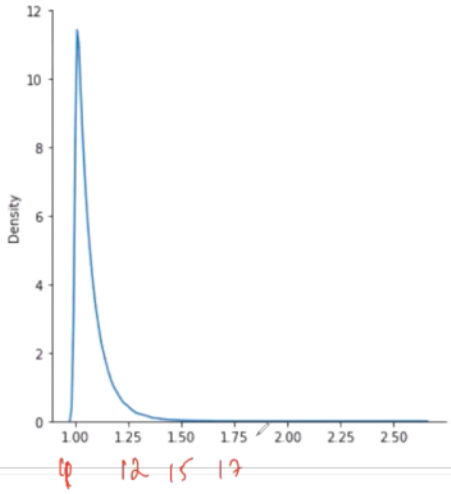
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

Out[41]: <seaborn.axisgrid.FacetGrid at 0x1f8e052ea48>



Anuja N. Narayanan

50:32

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)

Logout

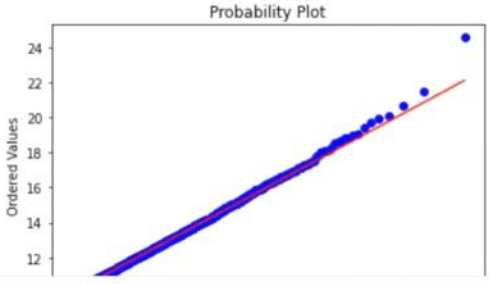
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

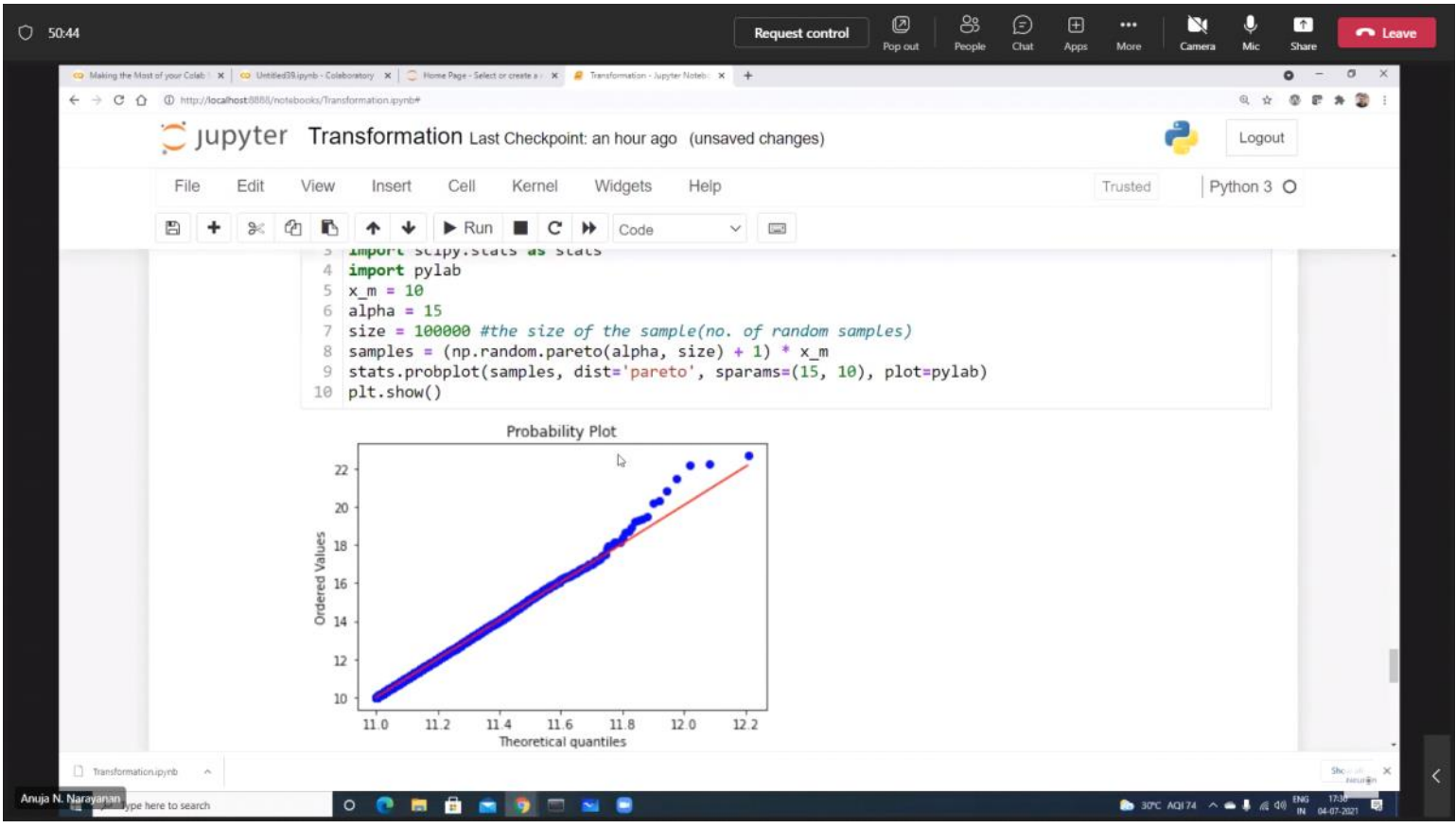
Run | Code

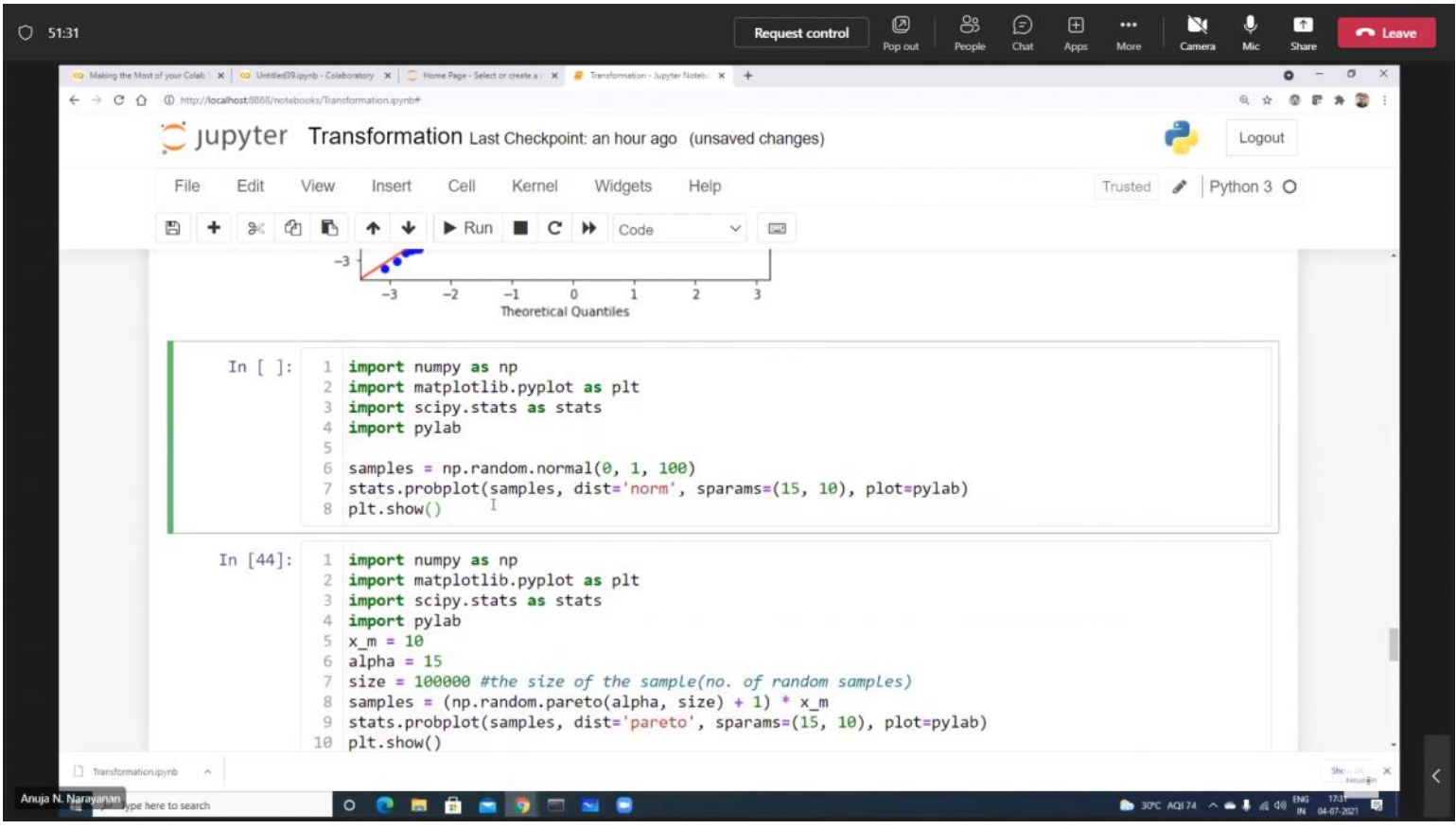
In [35]:

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pylab
x_m = 10
alpha = 15
size = 100000 #the size of the sample(no. of random samples)
samples = (np.random.pareto(alpha, size) + 1) * x_m
stats.probplot(samples, dist='pareto', sparams=(15, 10), plot=pylab)
plt.show()
```



Anuja N. Narayanan


50:44
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
import scipy.stats as stats
import pylab
x_m = 10
alpha = 15
size = 100000 #the size of the sample(no. of random samples)
samples = (np.random.pareto(alpha, size) + 1) * x_m
stats.probplot(samples, dist='pareto', sparams=(15, 10), plot=pylab)
plt.show()
Probability Plot
Ordered Values
Theoretical quantiles
Anuja N. Narayanan


51:31
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Theoretical Quantiles
In [ ]:
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pylab
samples = np.random.normal(0, 1, 100)
stats.probplot(samples, dist='norm', sparams=(15, 10), plot=pylab)
plt.show()
In [44]:
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pylab
x_m = 10
alpha = 15
size = 100000 #the size of the sample(no. of random samples)
samples = (np.random.pareto(alpha, size) + 1) * x_m
stats.probplot(samples, dist='pareto', sparams=(15, 10), plot=pylab)
plt.show()
Anuja N. Narayanan

51:36

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

jupyter Transformation Last Checkpoint: an hour ago (unsaved changes)

Logout

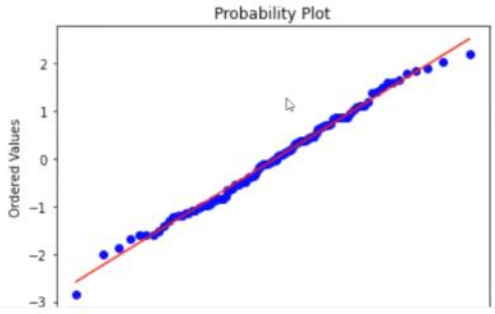
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

In [45]:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 import scipy.stats as stats
4 import pylab
5
6 samples = np.random.normal(0, 1, 100)
7 stats.probplot(samples, dist='norm', sparams=(15, 10), plot=pylab)
8 plt.show()
```



Anuja N. Narayanan

41:00

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

More actions

jupyter Transformation Last Checkpoint: 37 minutes ago (unsaved changes)

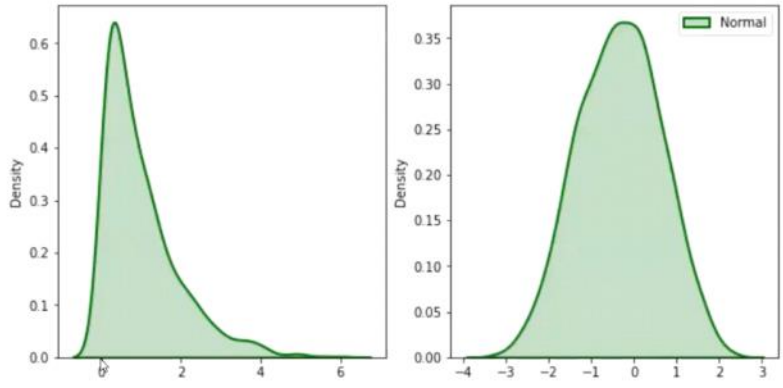
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

```
unction for kernel density plots).
  warnings.warn(msg, FutureWarning)
```



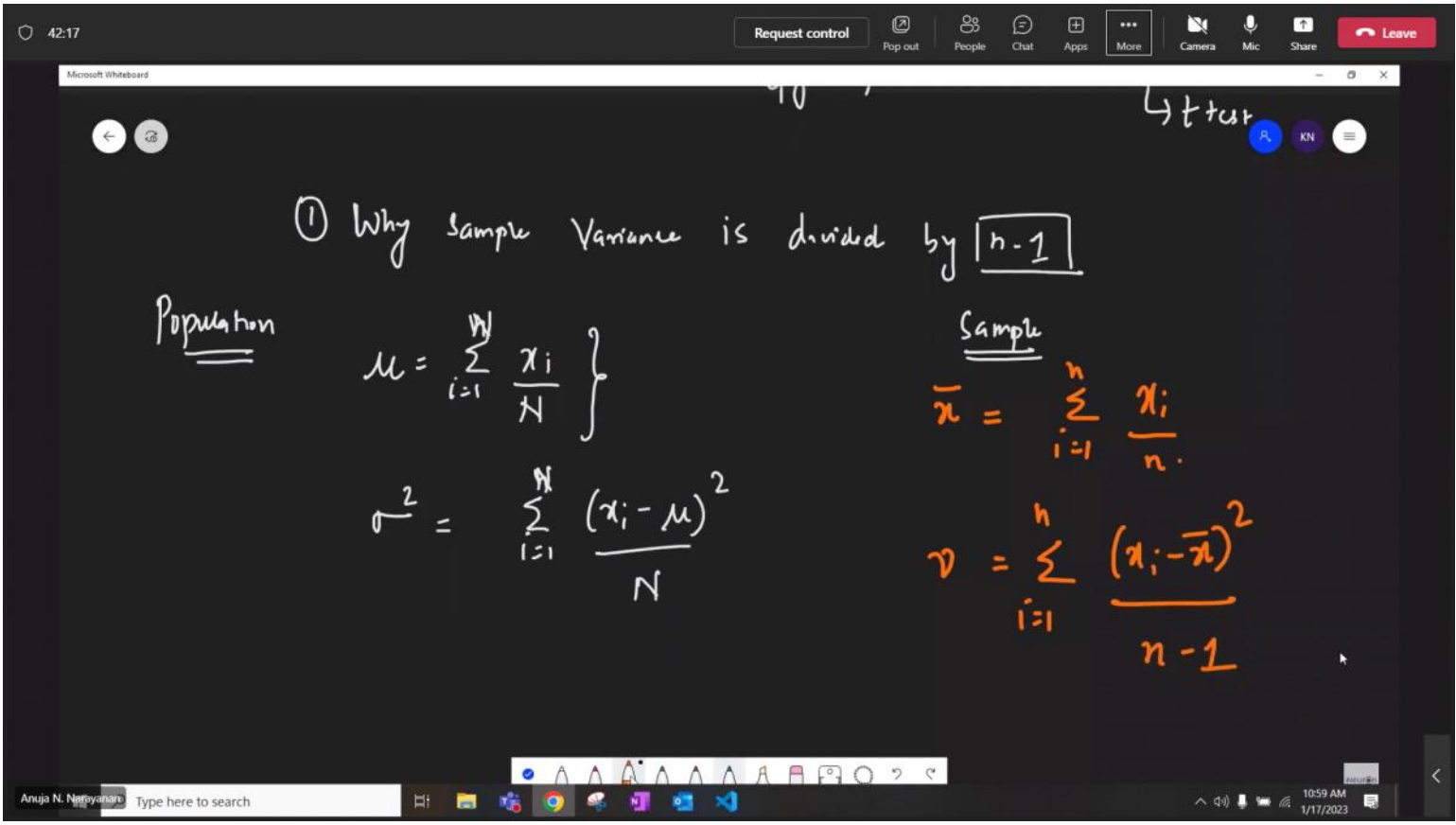
In [ ]:

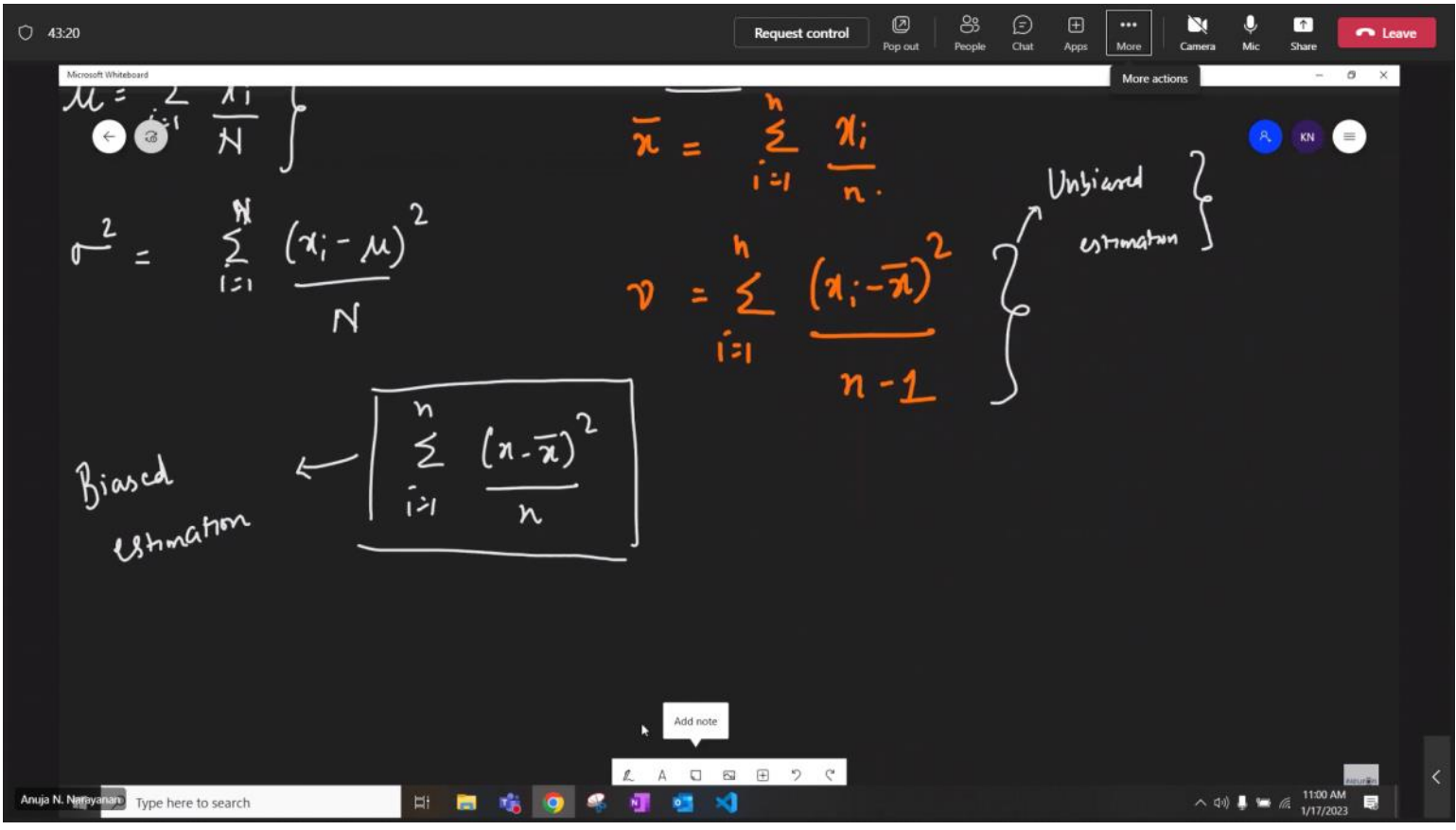
```
1
```

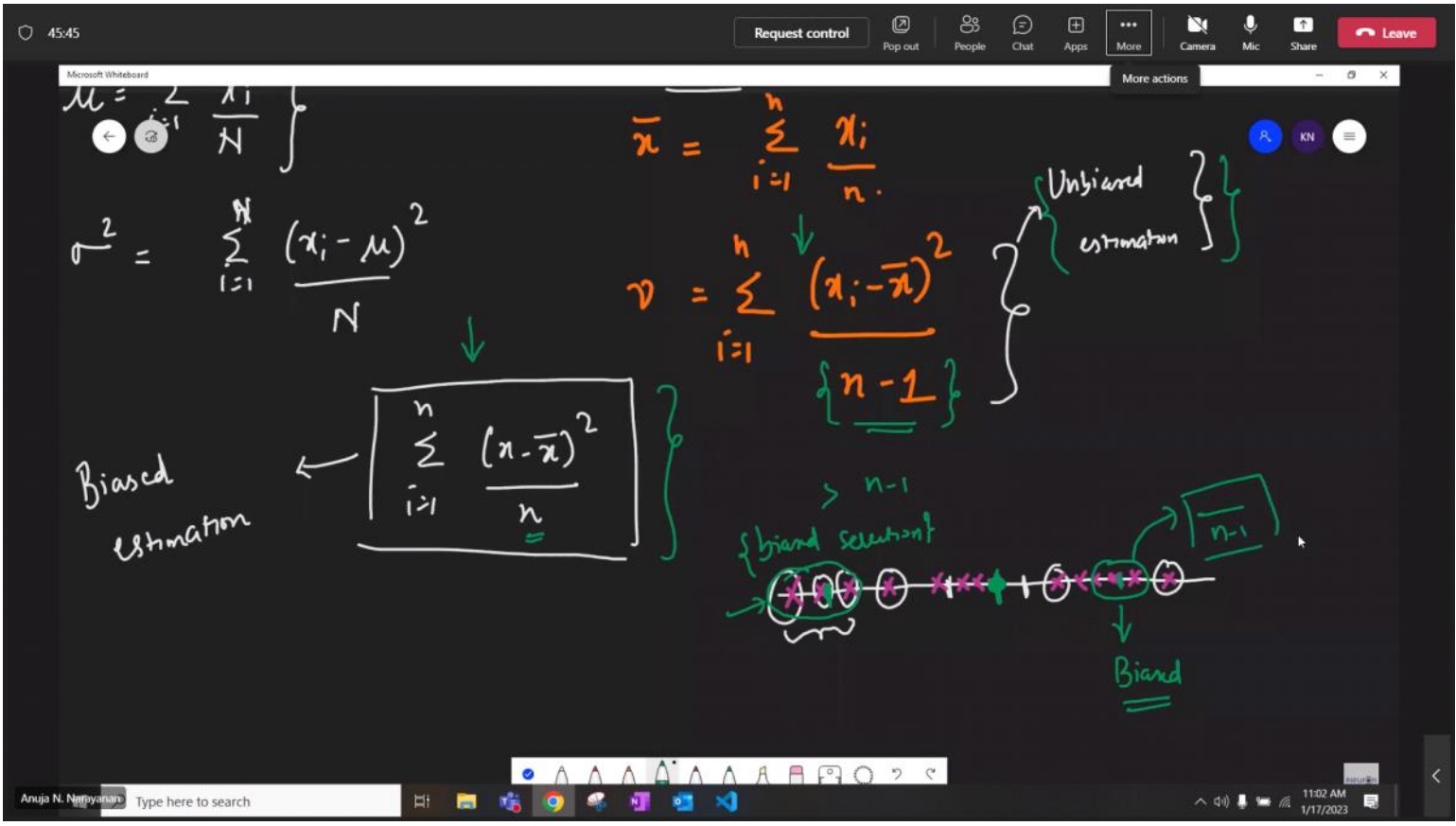
Anuja N. Narayanan

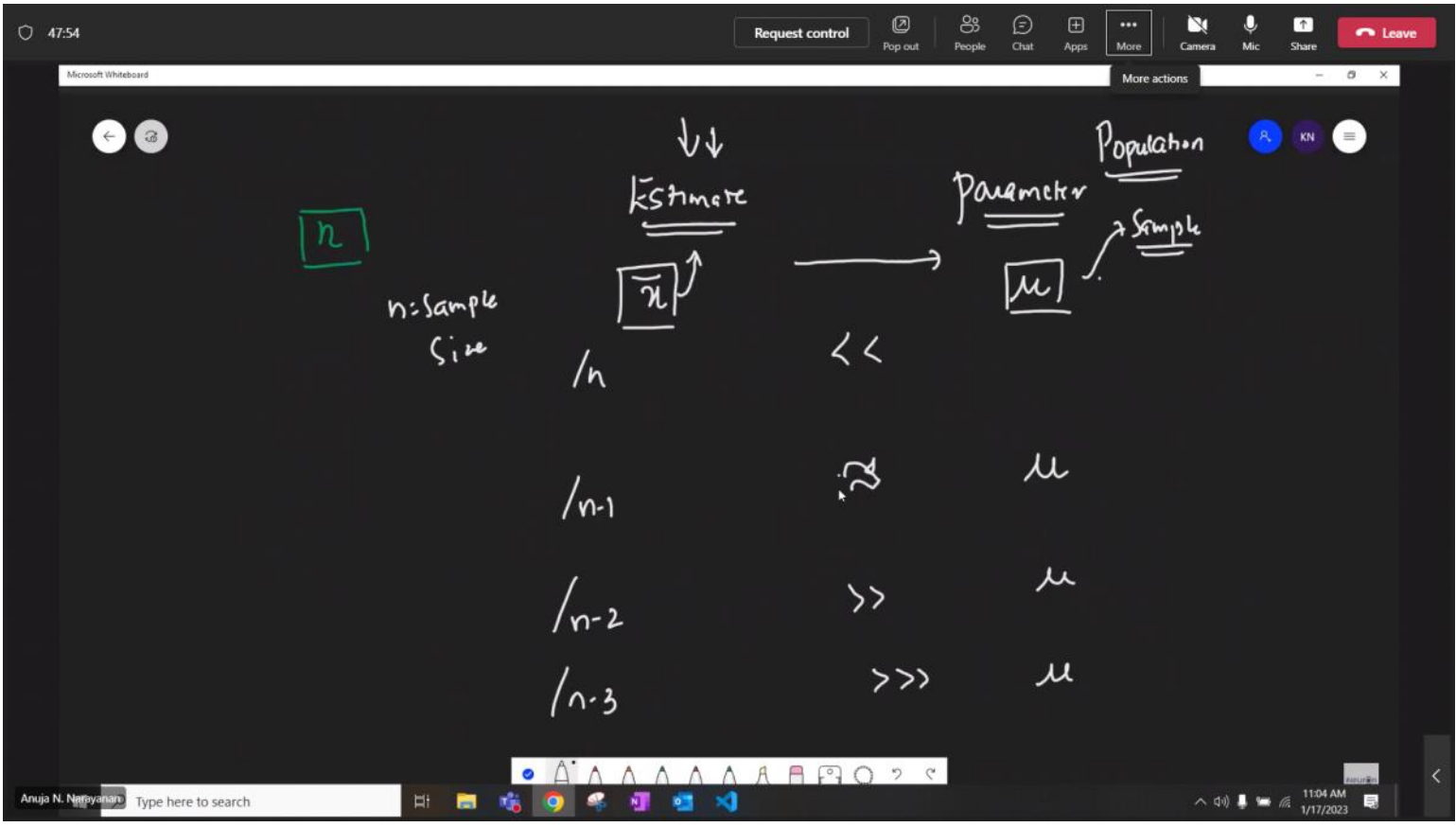
KN(lec-6)

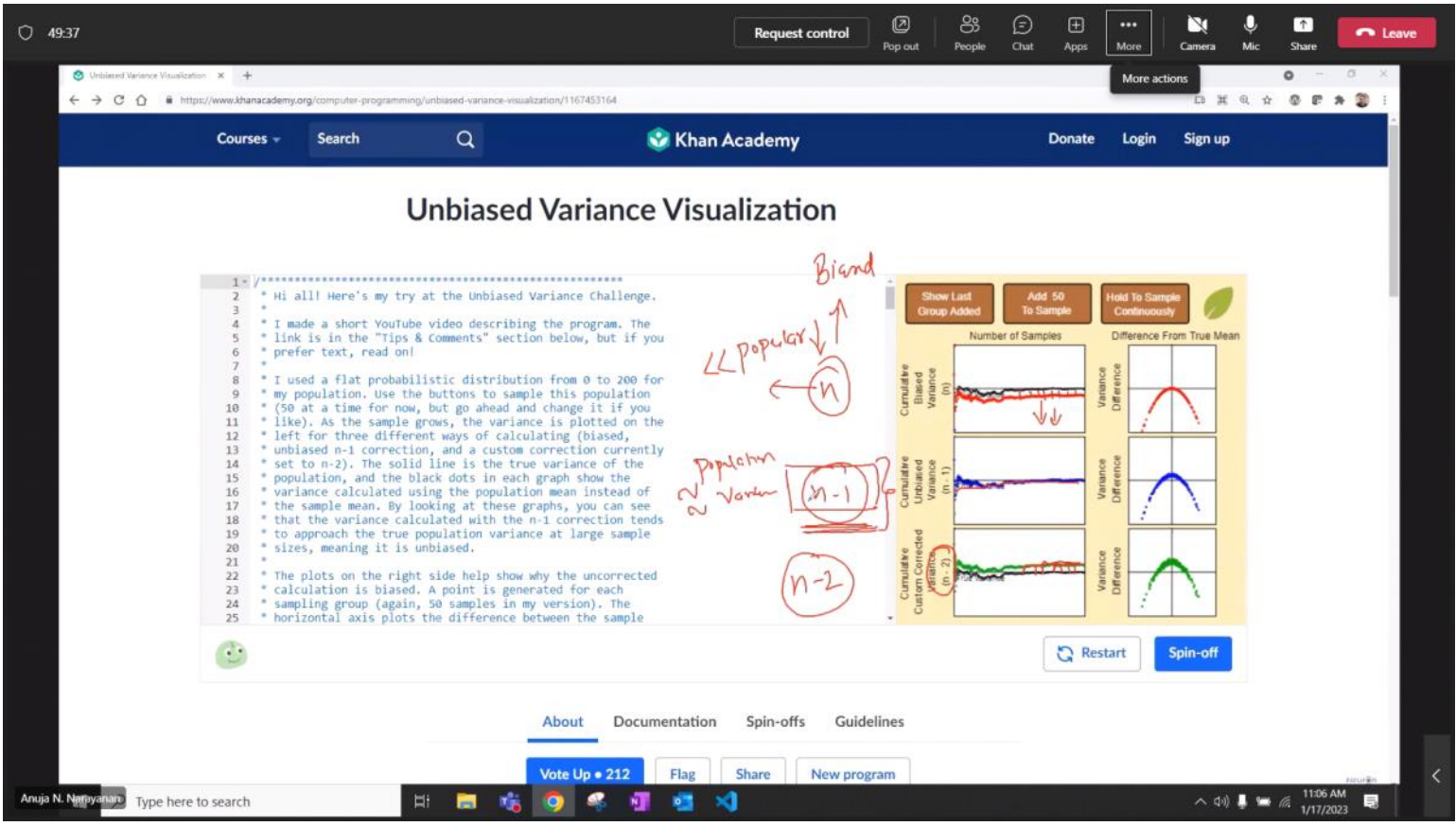
Tuesday, January 17, 2023 10:34 AM

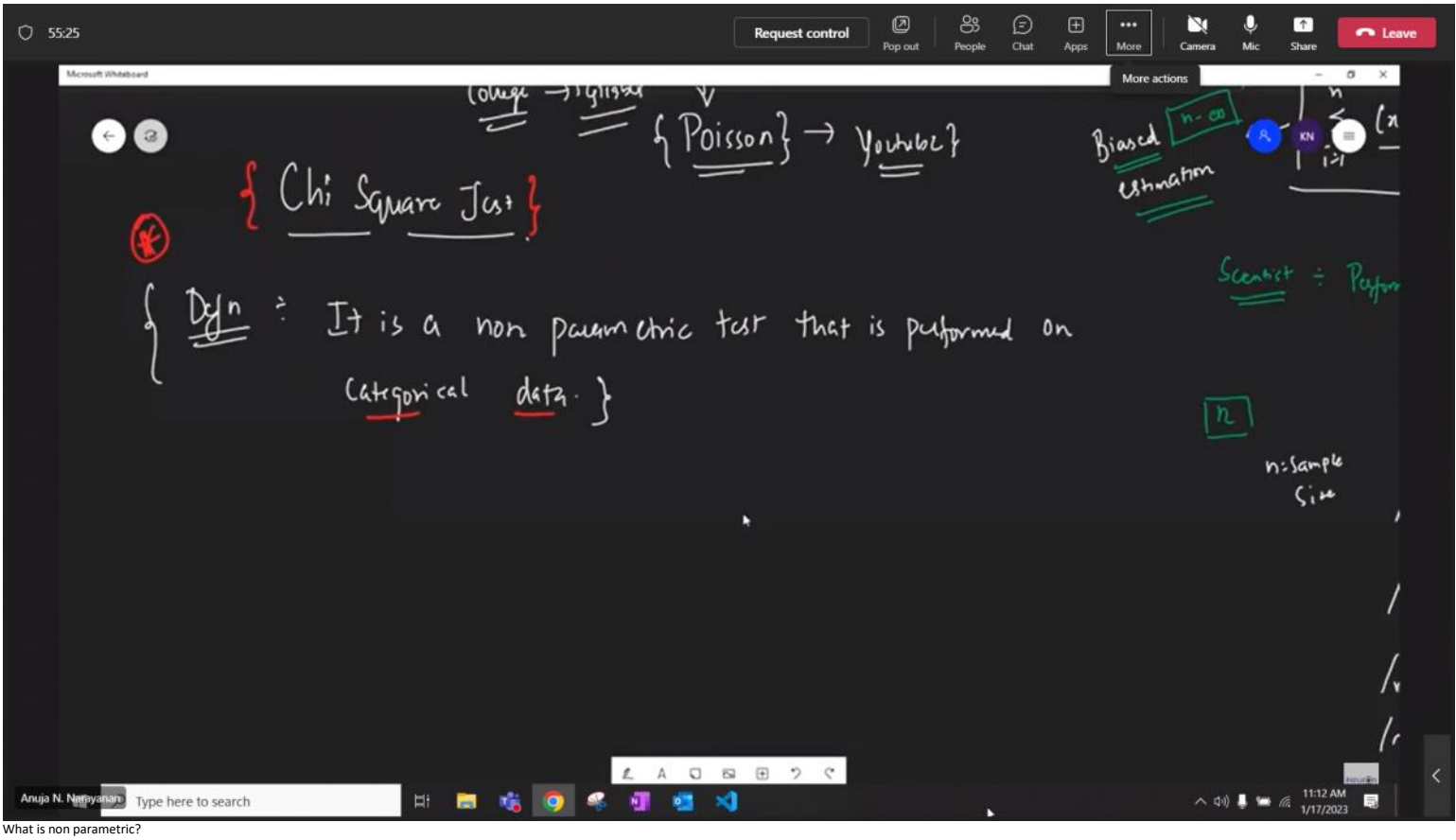





45:45
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
Unbiased estimation
Biased estimation
Biasd
Type here to search
11:02 AM
1/17/2023


47:54
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
Estimate
Parameter
Population
Sample
n: Sample Size
Type here to search
11:04 AM
1/17/2023





What is non parametric?

57:36

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

Eg: In the 2000 US Census the age of individual in a small town were found to be the following:

Less than 18	18-35	Greater than 35
20%	30%	50%

In 2010, ages of $n = 500$ individuals were sampled Below are the results.

Less than 18	18-35	Greater than 35
121	288	91

Using $\alpha = 0.05$, can you conclude the distribution of ages has been changed in 10 years?

Anuja N. Narayanan — Type here to search — 11:14 AM 1/17/2023

59:58

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

Greater than 35
91

n of ages has

① Null Hypothesis

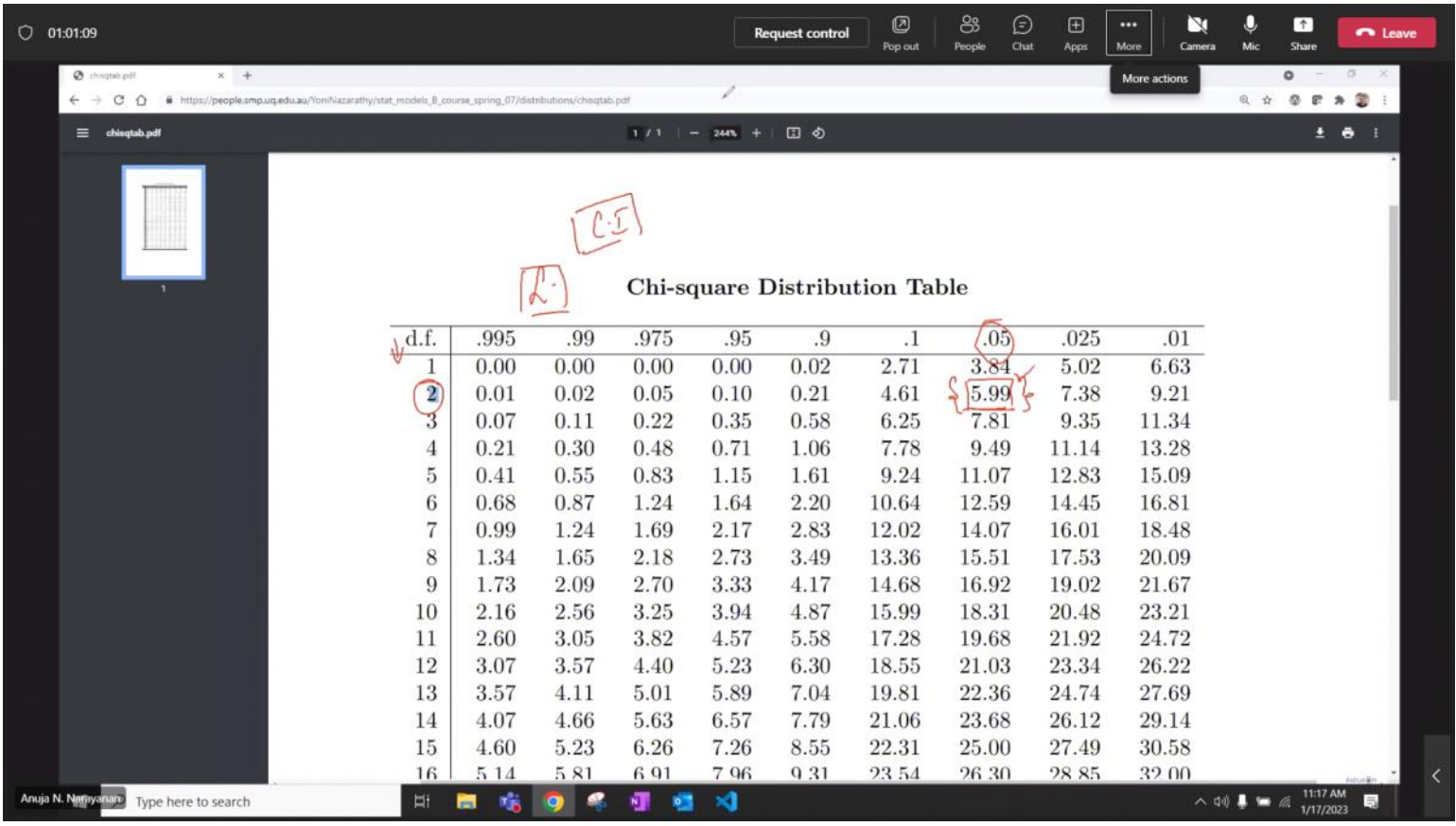
H_0: The data meets the age distribution

H_a: Distribution of ages has changed.

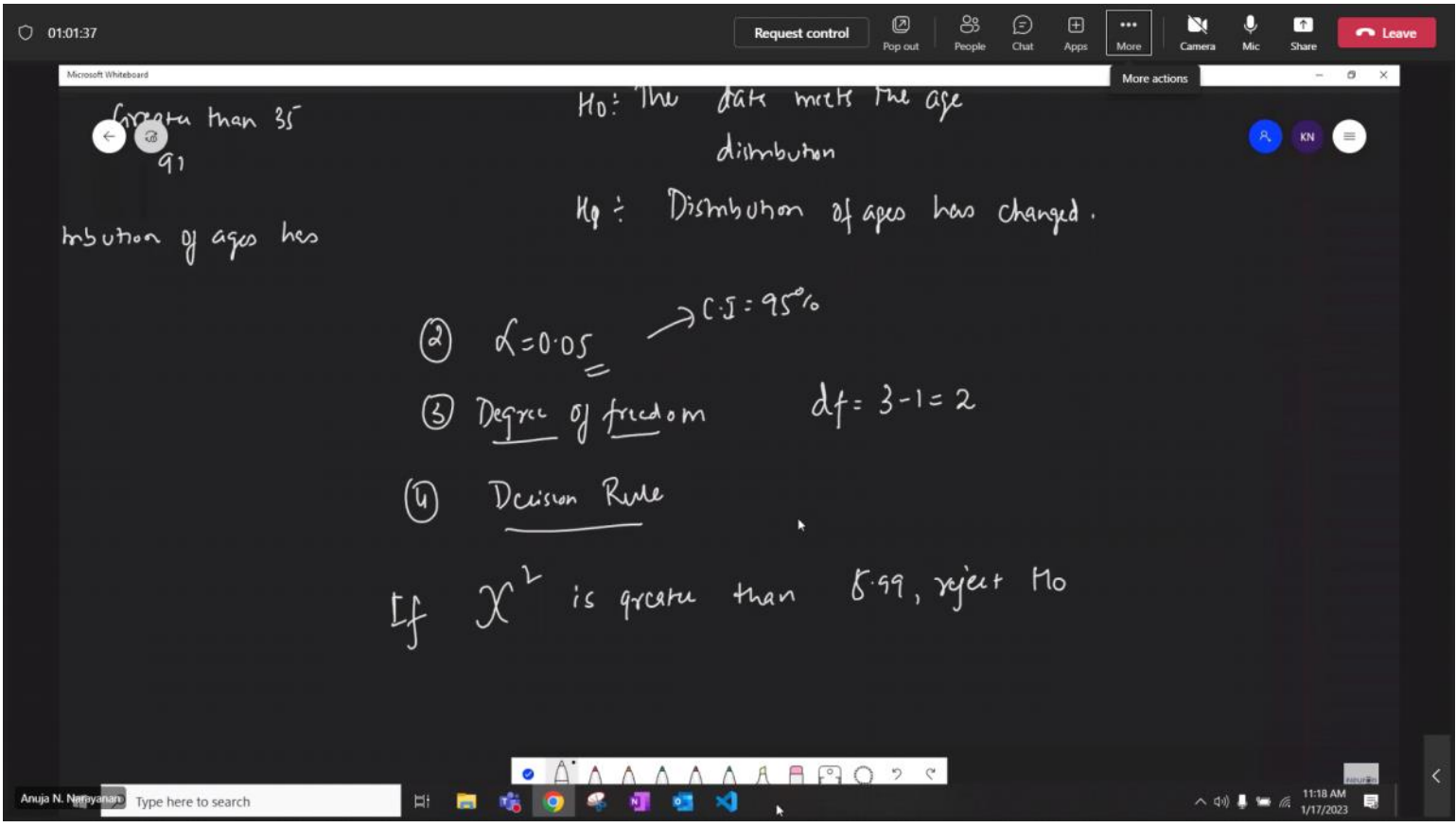
② $\alpha = 0.05$ → C.I = 95%

③ Degree of freedom $df = 3 - 1 = 2$

Anuja N. Narayanan — Type here to search — 11:16 AM 1/17/2023



d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00



01:03:20

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

121 288

500

Using $\alpha = 0.05$, can you conclude the distribution of ages has been changed in 10 years?

	Less than 18	18-35	>35
f_o (Observed) ⟶	121	288	91
f_e (Expected) ⟶	100	150	250

(2) (3) (4) $\sum f$

Anuja N. Narayanan — Type here to search

11:20 AM 1/17/2023

01:04:15

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

f_e (Expected) ⟶ 100 150 250

KN

$\sum f$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = 232.94$$

Anuja N. Narayanan — Type here to search

11:21 AM 1/17/2023

01:04:45

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

f_1 (Expected) $\longrightarrow$ 100 150 250

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = 232.94$$

Conclusion $\chi^2 > 5.99$

{ Distribution of Edge has got changed }.

Anuja N. Narayanan — Type here to search — 11:21 AM 1/17/2023

01:08:23

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

(2). Evaluate the relationship between 2 or more categorical Variable.

(*) 500 elementary school boys and girls are asked which are their favorite color, blue, green or pink?

Chi Square Test

$(f_e)^2$

Anuja N. Narayanan — Type here to search — 11:25 AM 1/17/2023

01:09:34

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

... between 2 or more Categorical Variable.

(#) 500 elementary school boys and girls are asked which are their favorite color, blue, green or pink? Results are shown below

	Blue	Green	Pink	
Boys	100	150	20	270
Girls	20	30	180	230
	120	180	200	

are Test

$(fe)^2$

Anuja N. Narayanan — Type here to search

11:26 AM 1/17/2023

01:10:18

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

their favorite color, blue, green or pink? Results are shown

	Blue	Green	Pink	
Boys	100	150	20	
Girls	20	30	180	
	120	180	200	500 n

are Test

$(fe)^2$

{ Using $\alpha = 0.05$, would you conclude there is a relationship between gender & color }

Anuja N. Narayanan — Type here to search

11:27 AM 1/17/2023

01:12:02

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

f_e 2 | 120 | 180 | 200 | 500 | n

{ Using $\alpha = 0.05$, would you conclude there is a relationship between gender & color }

↳ H_0: Gender & color are not related

H_1: ↓

Anuja N. Narayanan — Type here to search — 11:28 AM 1/17/2023

01:12:48

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

↳ H_0: Gender & color are not related

H_1: ↓

Step 2 : $\alpha = 0.05$

Step 3 : Calculate df

$df = (\text{rows} - 1) * (\text{columns} - 1)$

$= (2-1) * (3-1) = 2$

Anuja N. Narayanan — Type here to search — 11:29 AM 1/17/2023

01:14:20

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

$\alpha = 0.10$ ✓ 0.05

Step 4 :

State Decision Rule

$\chi^2 = \{5.99\}$ $\chi^2 > 5.99$ Reject H_0}

-1)

Step 5

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Anuja N. Narayanan | Type here to search | 11:31 AM 1/17/2023

01:16:24

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard — More actions

their favorite color, blue, green or pink? Results are shown below

	Blue	Green	Pink	
Boys	100	150	20	270
Girls	20	30	180	
	120	180	200	500 n

$$f_e = \frac{f_c f_r}{n}$$

$$(\text{Boy, Blue}) = \frac{270 \times 120}{500} = 54$$

{ Using $\alpha = 0.05$, would you conclude there is a relationship between gender & color }

↳ H_0: Gender & color are not related

$\alpha = 0.10$ ✓ 0.05

Anuja N. Narayanan | Type here to search | 11:33 AM 1/17/2023

favorite color, blue, green or pink? Results are shown below

	Blue	Green	Pink	
Boys	100	150	20	270
Girls	20	30	180	
	120	180	200	500 n

$f_e = \frac{f_c f_r}{n}$

(Boys, Green) $= \frac{150 \times 270}{500} = 81$

(Boy, Blue) $= \frac{100 \times 270}{500} = 54$

(Boys, Pink) $= \frac{20 \times 270}{500} = 10.8$

{ Using $\alpha = 0.05$, would you conclude there is a relationship between gender & color }

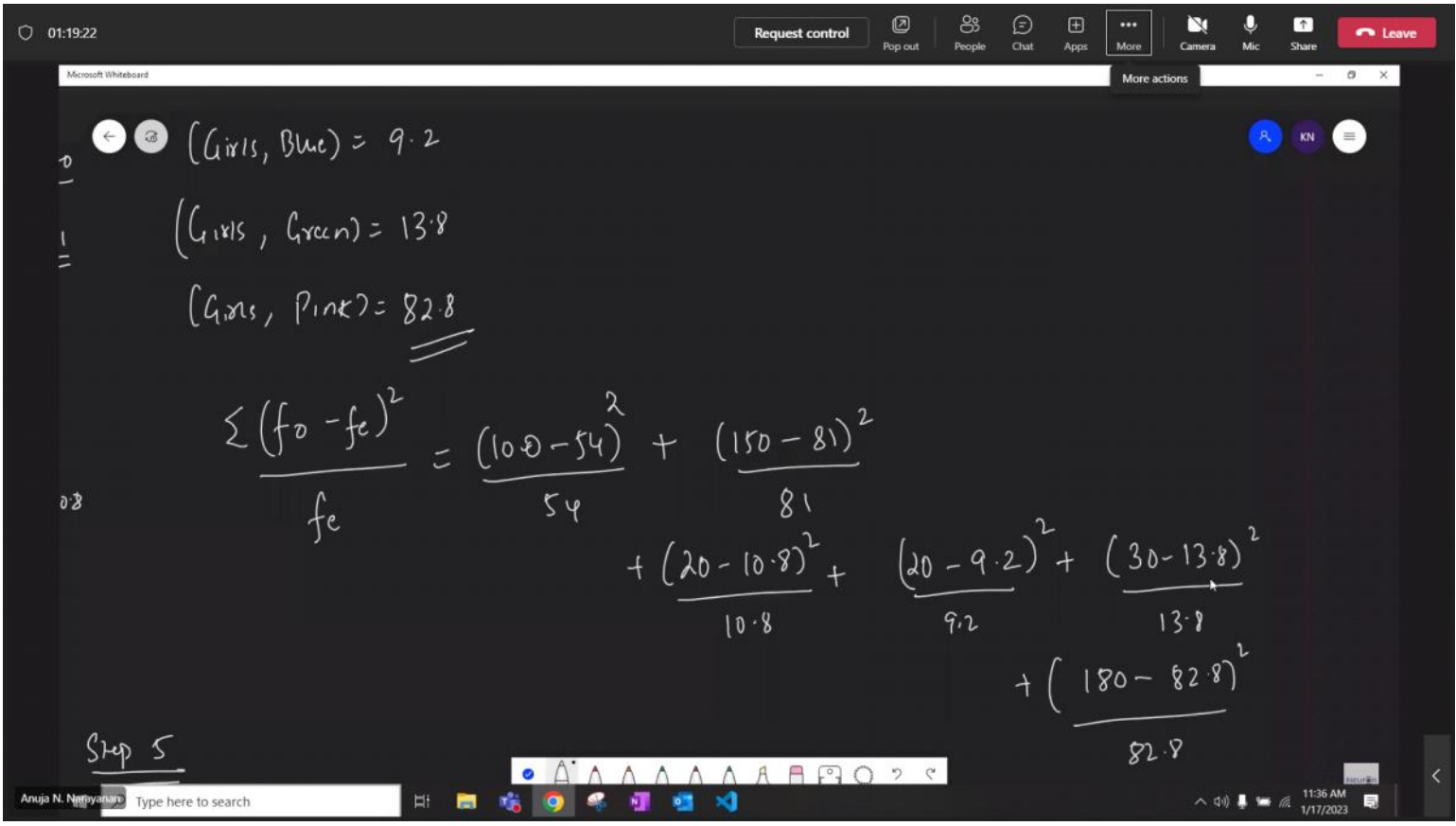
↳ H_0: Gender & color are not related

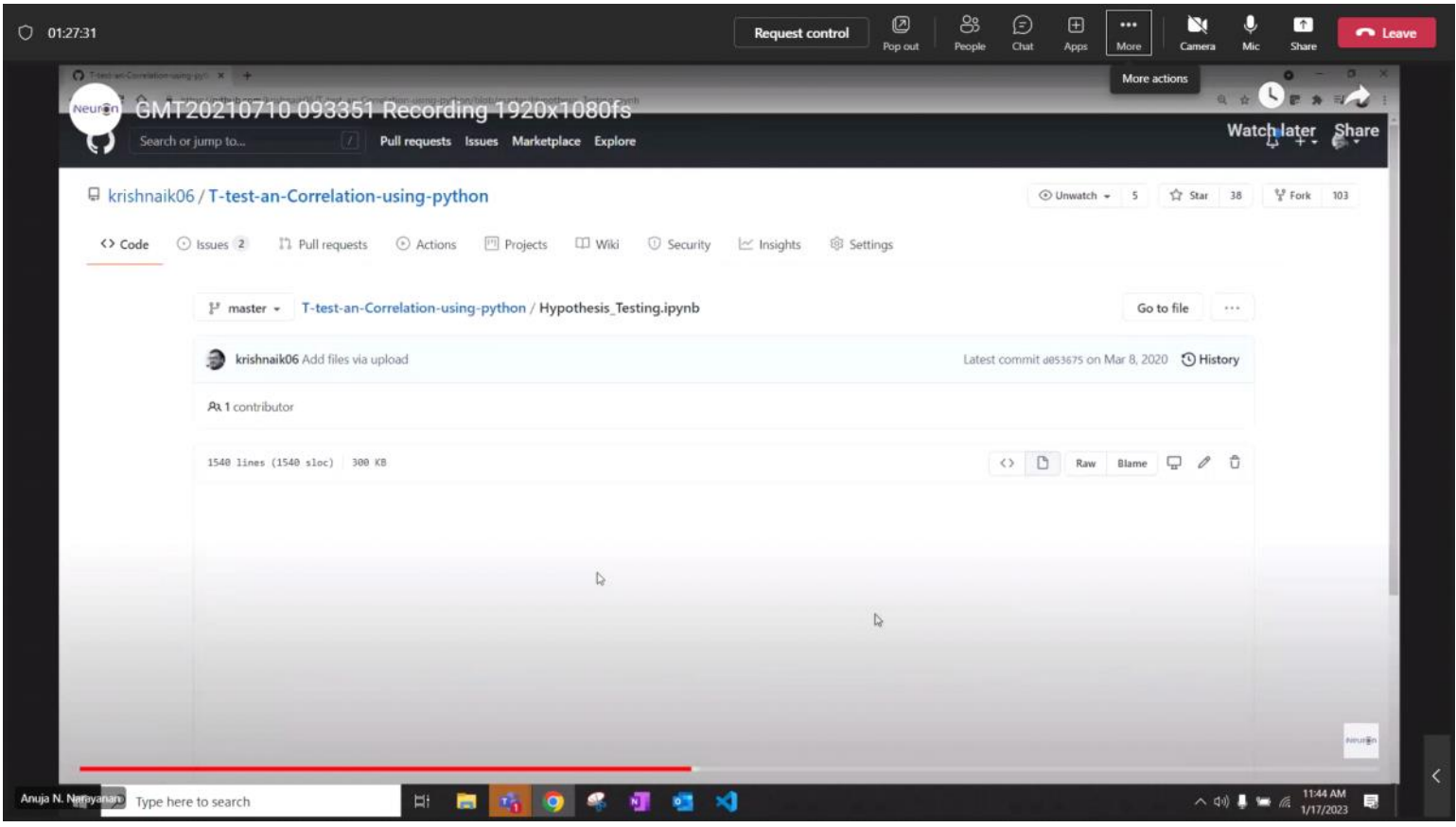
$\alpha = 0.10$ ✓ 0.05

(Girls, Blue) = 9.2

(Girls, Green) = 13.8

(Girls, Pink) = 82.8


01:19:22
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
More actions
(Girls, Blue) = 9.2
(Girls, Green) = 13.8
(Girls, Pink) = 82.8
$\frac{\Sigma (f_o - f_e)^2}{f_e} = \frac{(100-54)^2}{54} + \frac{(150-81)^2}{81} + \frac{(20-10.8)^2}{10.8} + \frac{(20-9.2)^2}{9.2} + \frac{(30-13.8)^2}{13.8} + \frac{(180-82.8)^2}{82.8}$
Step 5
Anuja N. Narayanan
Type here to search
11:36 AM
1/17/2023


01:27:31
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
More actions
GMT20210710 093351 Recording 1920x1080fs
Watch later
Share
Search or jump to...
Pull requests
Issues
Marketplace
Explore
krishnaik06 / T-test-an-Correlation-using-python
Unwatch 5
Star 38
Fork 103
Code
Issues 2
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
master
T-test-an-Correlation-using-python / Hypothesis_Testing.ipynb
Go to file
krishnaik06 Add files via upload
Latest commit d053675 on Mar 8, 2020
History
1 contributor
1540 lines (1540 sloc)
300 KB
Raw
Blame
Anuja N. Narayanan
Type here to search
11:44 AM
1/17/2023

01:32:00

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

jupyter Hypothesis_Testing (autosaved)

File Edit View Insert Cell Kernel Widgets Help | Not Trusted | Python 3

Run | Code

Anova Test(F-Test)

The t-test works well when dealing with two groups, but sometimes we want to compare more than two groups at the same time.

For example, if we wanted to test whether petal_width age differs based on some categorical variable like species, we have to compare the means of each level or group the variable

One Way F-test(Anova) :-

It tell whether two or more groups are similar or not based on their mean similarity and f-score.

Example : there are 3 different category of iris flowers and their petal width and need to check whether all 3 group are similar or not

In [283]: 1 import seaborn as sns

Anuja N. Narayanan | Type here to search | 11:48 AM 1/17/2023

01:38:48

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

KN

(8) Researchers want to test a anti anxiety medication. They split participants in 3 condition (0mg, 5mg, 100mg), then anxiety level is checked on scale 1-10. Are there any differences between the 3 condition $\alpha = 0.05$

0mg	50mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

$H_0 = \mu_{0mg} = \mu_{50mg} = \mu_{100mg}$

$H_1 = \mu_{0mg} \neq \mu_{50mg} \neq \mu_{100mg}$.

Anuja N. Narayanan | Type here to search | 11:55 AM 1/17/2023

01:40:23

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

)mg, 100mg), then anxiety level is checked on scale 1-10. Are there any differences between the 3 condition $\alpha = 0.05$

0mg	50mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

Step 2 : $\alpha = 0.05$

Step 3 : Calculate Degree of freedom

$$\begin{cases} H_0 : \mu_{0mg} = \mu_{50mg} = \mu_{100mg} \\ H_1 : \mu_{0mg} \neq \mu_{50mg} \neq \mu_{100mg} \end{cases}$$

$df_{Between} = \alpha - 1 = 3 - 1 = 2$

$df_{within} = N - \alpha = 21 - 3 = 18$

$df_{Total} = N - 1 = 21 - 1 = 20$

Anuja N. Narayanan | Type here to search | 11:57 AM 1/17/2023

01:42:16

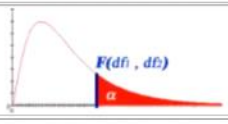
Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

F-Distribution Tables

Not secure | http://www.socr.ucla.edu/Applets.dir/F_Table.html

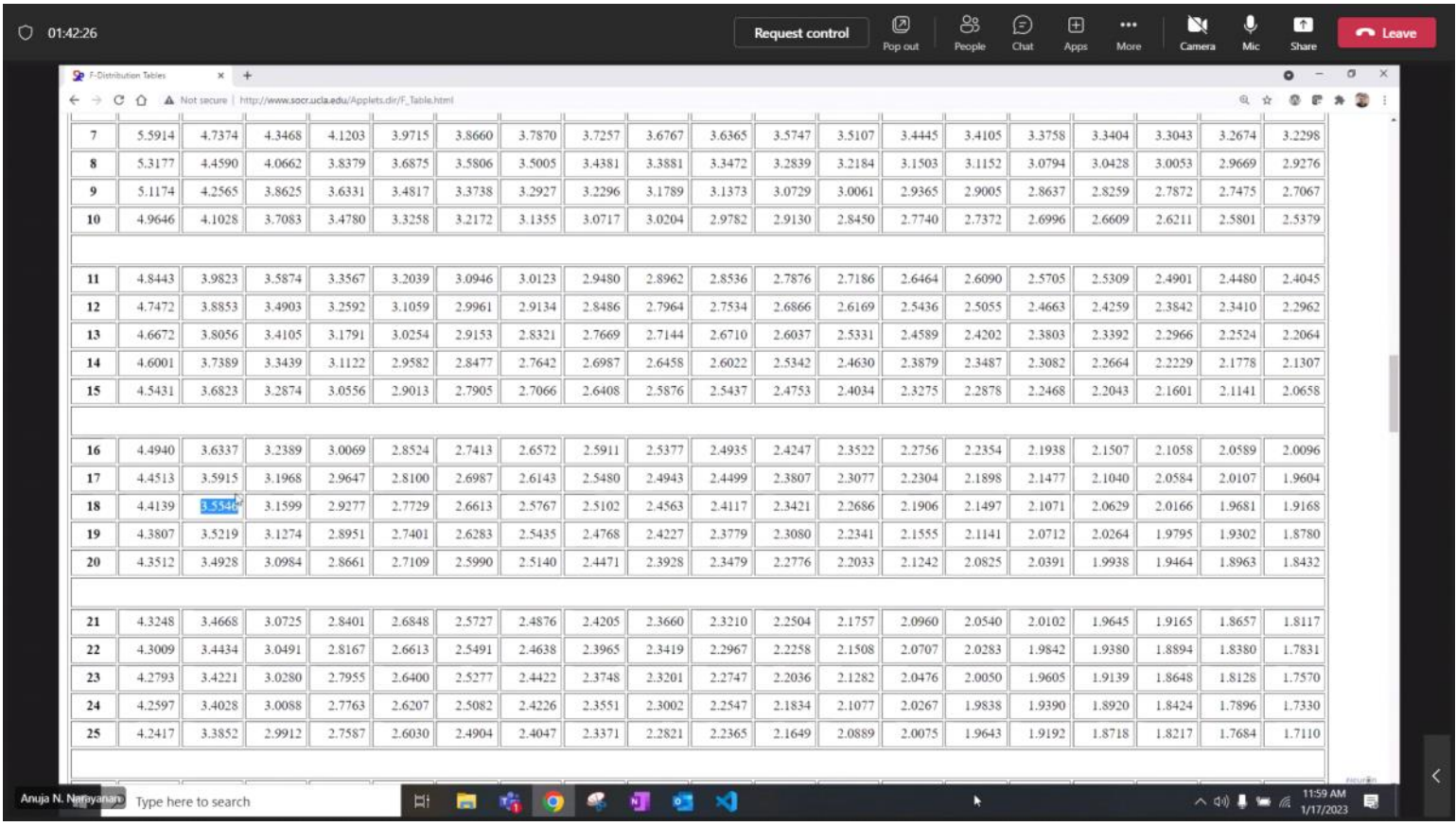
∞	2.70554	2.30259	2.08380	1.94486	1.84727	1.77411	1.71672	1.67020	1.63152	1.59872	1.54578	1.48714	1.42060	1.38318	1.34187	1.29513	1.23995	1.16860	1.00000

F Table for α = 0.05

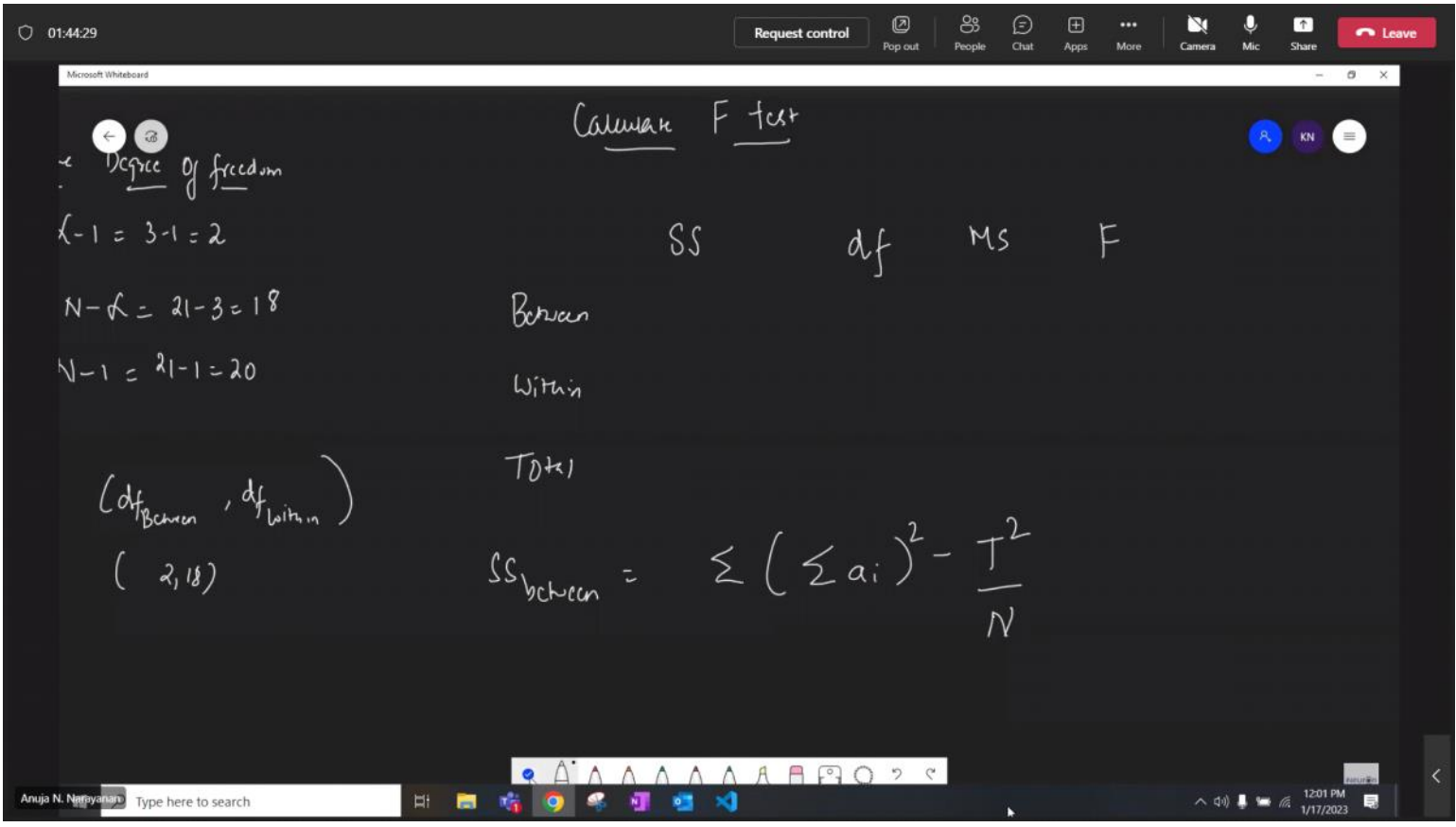


/	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	245.9499	248.0131	249.0518	250.0951	251.1432	252.1957	253.2529	254.3144
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	19.4291	19.4458	19.4541	19.4624	19.4707	19.4791	19.4874	19.4957
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.5720	8.5494	8.5264
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7744	5.7459	5.7170	5.6877	5.6581	5.6281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.5272	4.4957	4.4638	4.4314	4.3985	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.7398	3.7047	3.6689
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.4105	3.3758	3.3404	3.3043	3.2674	3.2298
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.9669	2.9276
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.7872	2.7475	2.7067
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7372	2.6996	2.6609	2.6211	2.5801	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.6090	2.5705	2.5309	2.4901	2.4480	2.4045
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.3842	2.3410	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.2966	2.2524	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.4630	2.3879	2.3487	2.3082	2.2664	2.2229	2.1778	2.1307

Anuja N. Narayanan | Type here to search | 11:59 AM 1/17/2023



7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.4105	3.3758	3.3404	3.3043	3.2674	3.2298
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.9669	2.9276
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.7872	2.7475	2.7067
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7372	2.6996	2.6609	2.6211	2.5801	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.6090	2.5705	2.5309	2.4901	2.4480	2.4045
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.3842	2.3410	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.2966	2.2524	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.4630	2.3879	2.3487	2.3082	2.2664	2.2229	2.1778	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034	2.3275	2.2878	2.2468	2.2043	2.1601	2.1141	2.0658
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.2354	2.1938	2.1507	2.1058	2.0589	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1898	2.1477	2.1040	2.0584	2.0107	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1497	2.1071	2.0629	2.0166	1.9681	1.9168
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.1141	2.0712	2.0264	1.9795	1.9302	1.8780
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.9464	1.8963	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.2504	2.1757	2.0960	2.0540	2.0102	1.9645	1.9165	1.8657	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	2.0283	1.9842	1.9380	1.8894	1.8380	1.7831
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	2.0050	1.9605	1.9139	1.8648	1.8128	1.7570
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9838	1.9390	1.8920	1.8424	1.7896	1.7330
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9643	1.9192	1.8718	1.8217	1.7684	1.7110



01:44:52

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

0mg	50mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

$(9+8+7+8+8+9+8)^2$

$\Sigma(a_i)^2$

Step 2 : $\alpha = 0.05$

Cartesian

Step 3 : Calcula

$H_0 : \mu_{0mg} = \mu_{50mg} = \mu_{100mg}$

$H_1 : \mu_{0mg} \neq \mu_{50mg} \neq \mu_{100mg}$.

$df_{Between} =$

$df_{within} =$

$df_{Total} =$

Anuja N. Narayanan | Type here to search | 12:01 PM 1/17/2023

01:46:49

Request control | Pop out | People | Chat | Apps | More | Camera | Mic | Share | Leave

Microsoft Whiteboard

$J-1 = 21-1 = 20$

Within

Total

$(df_{Between}, df_{within})$

$(2, 18)$

(Σa_i)

$SS_{between} = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$

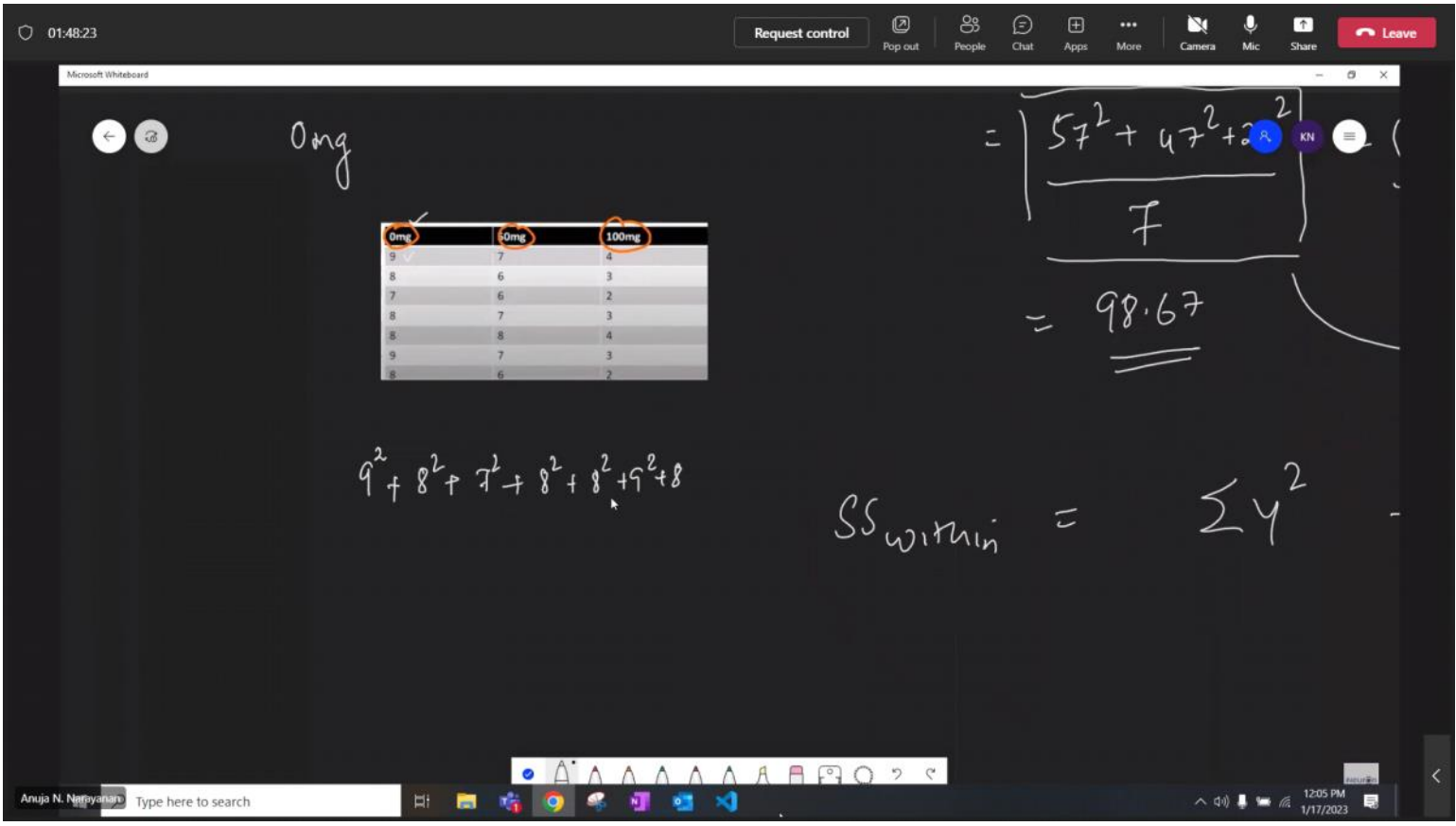
$[57+47+21] \Rightarrow T$

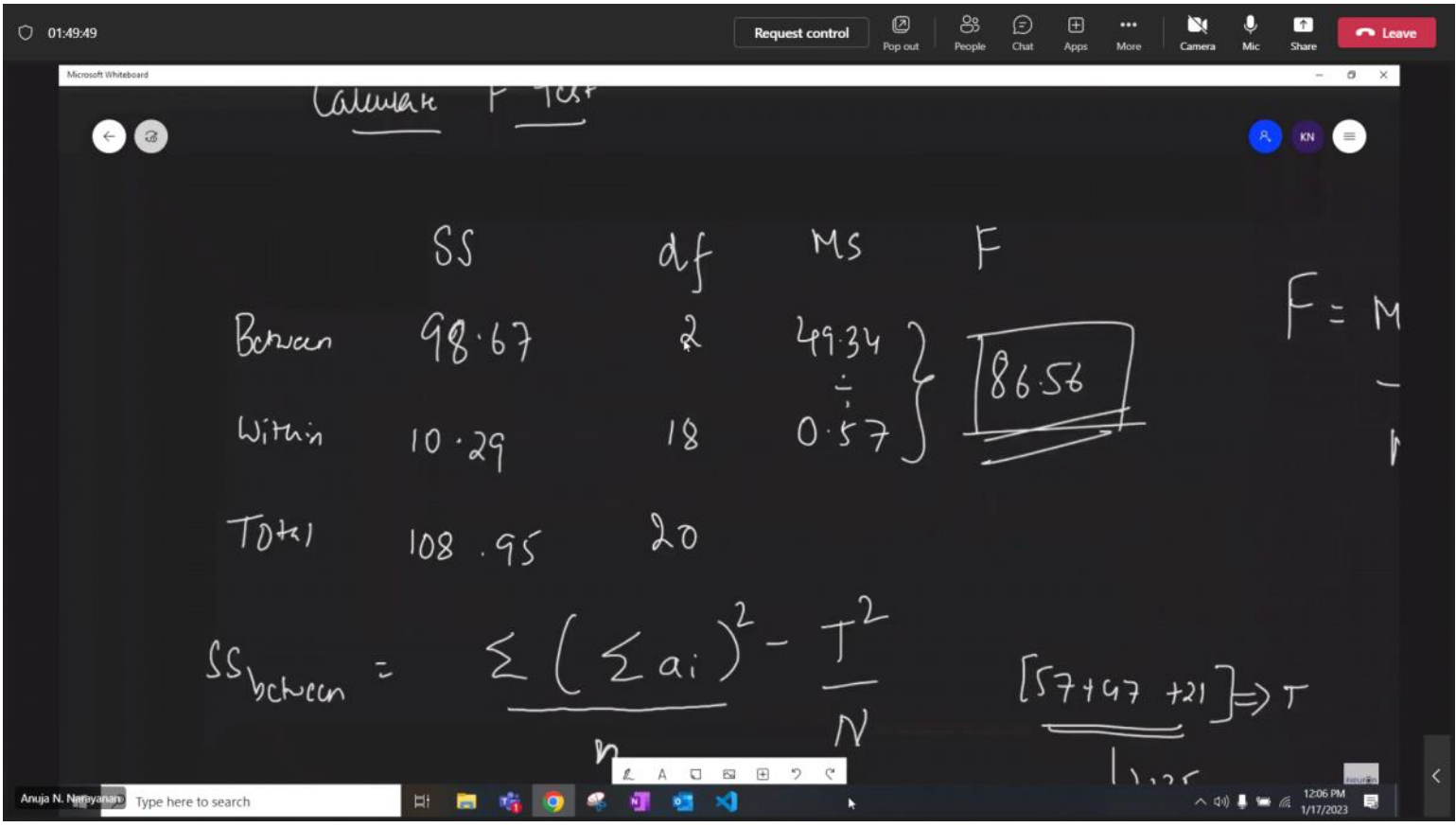
$\downarrow 125$

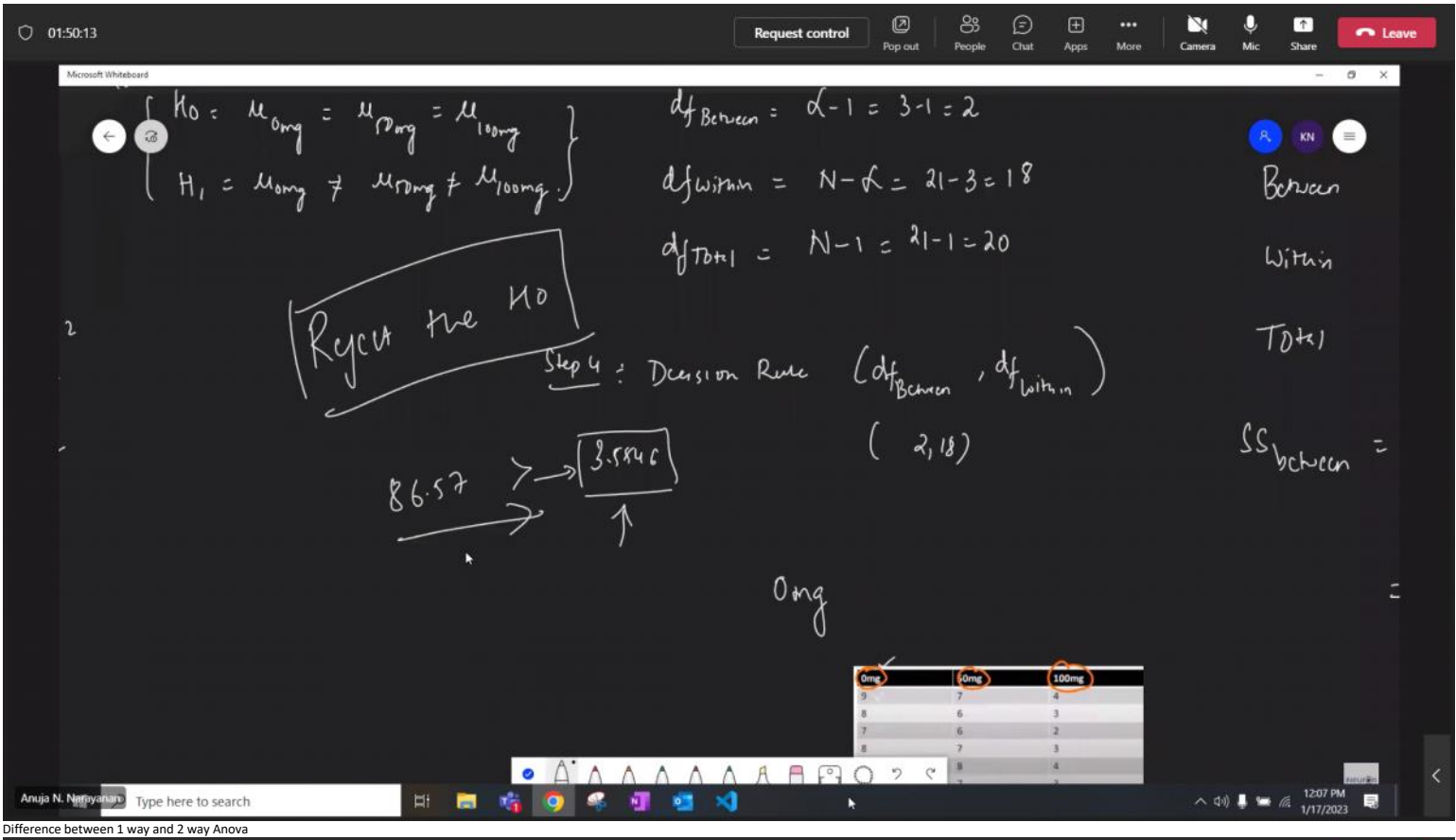
$= \frac{57^2+47^2+21^2}{7} - \frac{(125)^2}{21}$

$= 98.67$

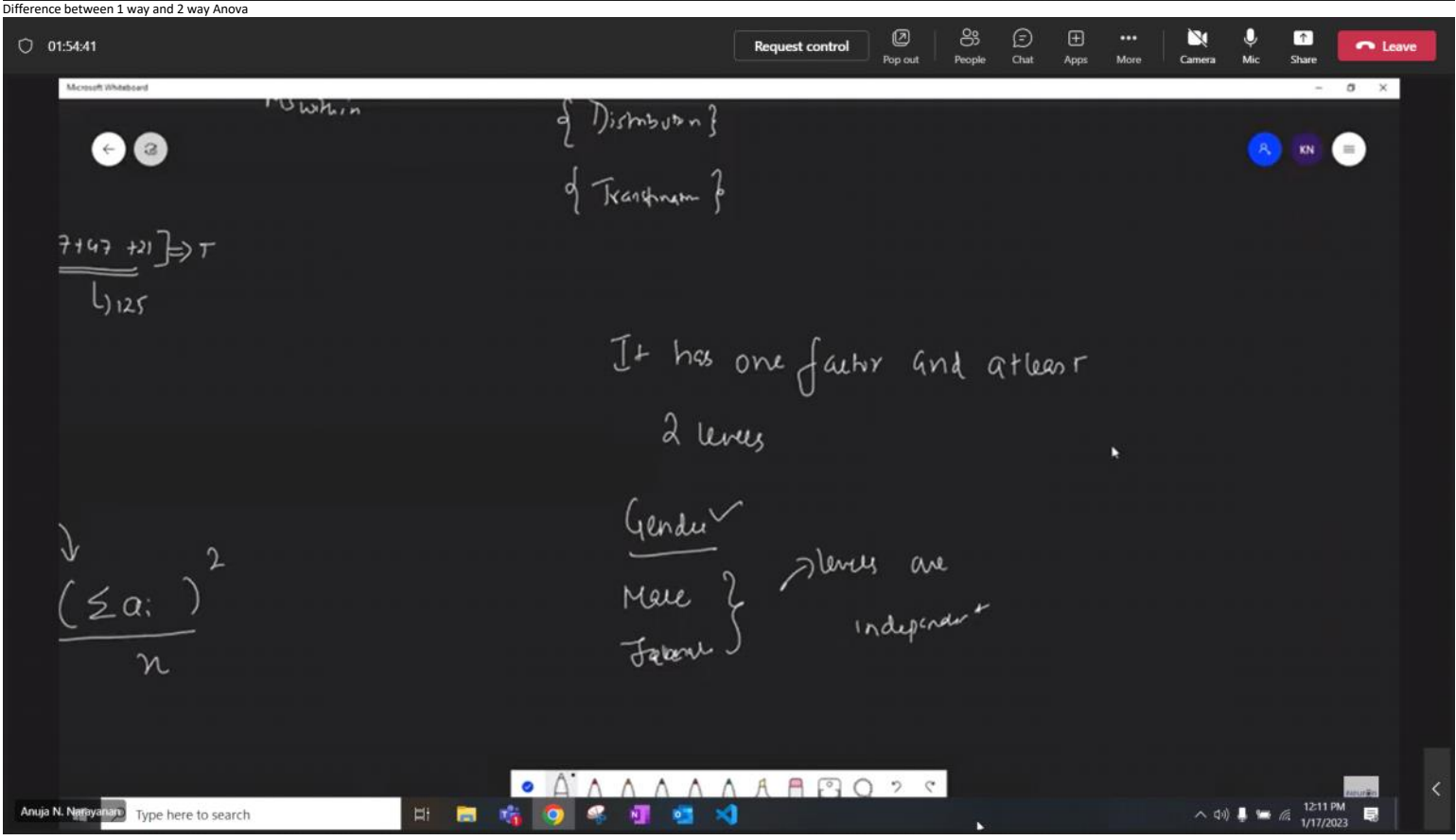
Anuja N. Narayanan | Type here to search | 12:03 PM 1/17/2023

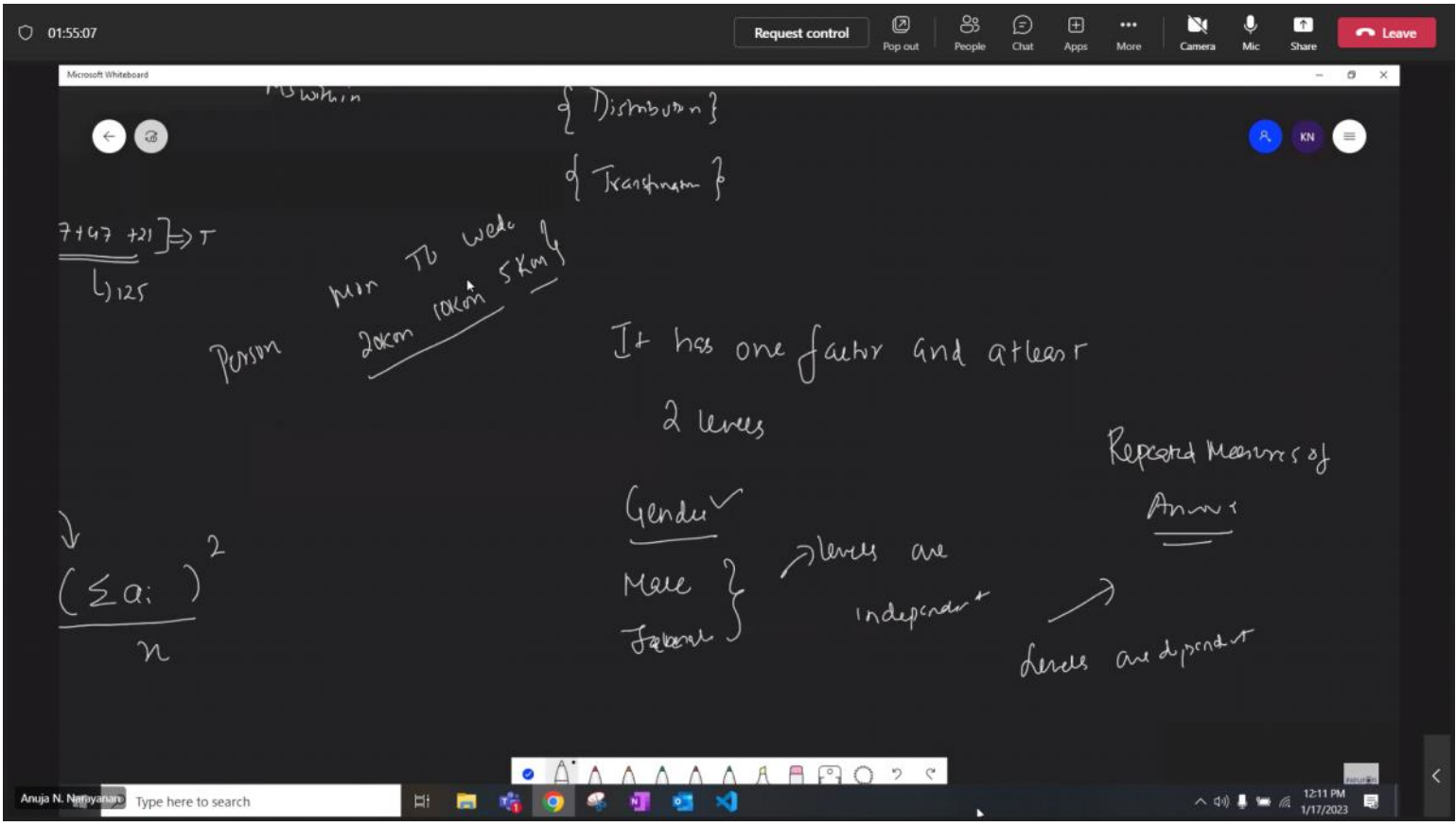

01:48:23
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
0mg
50mg
100mg
9 7 4
8 6 3
7 6 2
8 7 3
8 8 4
9 7 3
8 6 2
$= \frac{57^2 + 47^2 + 21^2}{7}$
$= 98.67$
$9^2 + 8^2 + 7^2 + 8^2 + 8^2 + 9^2 + 8$
$SS_{within} = \Sigma y^2$
Anuja N. Narayanan
Type here to search
12:05 PM
1/17/2023


01:49:49
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
Calculate F test
SS df MS F
Between 98.67 2 49.34 86.56
Within 10.29 18 0.57
Total 108.95 20
$SS_{between} = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$
$[57 + 47 + 21] \Rightarrow T$
Anuja N. Narayanan
Type here to search
12:06 PM
1/17/2023



Difference between 1 way and 2 way Anova




01:55:07
Request control
Pop out
People
Chat
Apps
More
Camera
Mic
Share
Leave
Microsoft Whiteboard
{ Distribution }
{ Transformation }
7+47 +21 ⇒ T
125
Person
Min To Walk
2km 10km 5Km
It has one factor and atleast
2 levels
Gender
Male
Female
levels are
independent
levels are dependent
Repeated Measures of
Anxiety
$\frac{(\Sigma a_i)^2}{n}$
Type here to search
12:11 PM
1/17/2023

Stats Lect 1 Notes

Wednesday, January 18, 2023 10:51 AM

Statistics definition : A branch of science which involves collecting, analyzing , exploring, visualizing data in large quantities, so that you can come up with some meaningful information and solving various use cases and getting conclusions

Types of Stats

- Descriptive Statistics : Analyzing, Exploring, Visualizing, and other techniques to **understand the data** i.e. Organizing and Summarizing the data
  Tools : Histograms, Bar graph, Pie graph, PDF function, Scatter plots
  5 point summary: min, 25%, median(50%), 75%, max

- Inferential Statistics : From the population data we take the sample data , perform certain experiments and based on the results we make conclusions about the population.
  Tools: Hypothesis testing, p value, t test, F test, Chi square test, Z test

Population (N) and Sample (n)

A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from, sample is a subset of population.
The size of the sample is always less than the size of the population. There are various sampling methods to select the subset of N that can be considered as n for our experiments.

Sampling Methods

1. **Random Sampling** : Samples are chosen randomly.
   Advantages
   - Easy to implement.
   - Each member of the population has an equal chance of being chosen.
   - Free from bias.

   Disadvantages
   - If the sampling frame is large random sampling may be impractical.
   - A complete list of the population may not be available.
   - Minority subgroups within the population may not be present in sample.
   - Issues of overlapping in the samples
   - For a specific use case this doesn't work like if you want to do survey related to data sceience

   Example : Exit polls

2. **Stratified Sampling** : Population is divided into strata/ subgroups based on specific characteristics, such as age, gender or race. Within the strata random sampling is used to choose the sample.
   Advantages
   - Strata can be proportionally represented in the final sample.
   - It is easy to compare subgroups.

   Disadvantages
   - Information must be gathered before being able to divide the population into subgroups.

   Example : Sales of cosmetics male vs female

3. **Systematic Sampling** : All data is sequentially numbered and every nth piece of data is chosen. The number n is chosen by n=size of population / desired population size.
   Advantages
   - Easy to select.
   - Identified easily.
   - Evenly spread over the entire population.

   Disadvantages
   - May be biased where the pattern used for the samples coincides with a pattern in the population.

4. **Clustering Sampling** : Data is divided into clusters and random sampling is used to select whole clusters. The sample will be obtained from a collection of entire cluster groups. It is usually used with naturally occurring groups of individuals for example classrooms, city blocks or postcodes.
   Advantages
   - Cuts down the cost and time by collecting data from only a limited number of groups.
   - Can show grouped variations.

   Disadvantages
   - It is not a genuine random sample.
   - The sample size is smaller and from thus the sample is likely to be less representative of the population.

   Example : E commerce website providing different offers for different groups of the data like new users , legacy users, customer segmentation

This sampling technique will help in train test split while building ML models.

Measures of Central Tendency

1. Mean x(bar) = $x_1 + x_2 + x_3 + ...+ x_n$ / n
   Whenever we have outliers do not use mean as a measure of central tendency. The mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero. One main disadvantage of the mean is its susceptibility to the influence of outliers

Example calculation and Formula - both population and sample
Ex1 : data = [1,2,3,4,5], mean = 3
Ex 2: data = [1,2,3,4,5, 100], mean = 19.16
From above example 1 and 2 we can say that the mean has shifted when there is a outlier in the data. So the conclusion here is don't use mean when there is a outlier in the data.

2. Median :
The median is less affected by outliers and skewed data. This property makes it a better option than the mean as a measure of central tendency.
ex1 : data = [1,2,3,4,5], median = 3 : len(data) is odd so the median would be the middle element
ex2: data = [1,2,3,4,5,100], median = (3+4)/2 when len(data) is even then the median would be the avg of 2 middle element.
So the conclusion here is that the median is not impacted by the outliers.
Note: (1st step is to sort the data and then calculate the median)

3. Mode:
The mode has an advantage over the median and the mean because it can be computed for both numerical and categorical (non-numerical) data. However, the mode has its limitations too. In some distributions, the mode may not reflect the center of the distribution very well. The presence of more than one mode can limit the ability of the mode to describe the center or typical value of the distribution because a single value cannot be identified to describe the center. In some cases, particularly where the data are continuous, the distribution may have no mode (i.e., if all values are different). In such cases, it may be better to consider using the median or mean, or group the data into appropriate intervals, and find the modal class.
Example calculation (odd and even)

- The mean is
  - A single value that is intended to represent an entire set of data.
  - It attempts to identify a central position (middle) within a data set.
  - The mean is sensitive to outliers or skewed data.
- The Median is
  - The middle value of a sorted set of numbers.
  - To determine the median we need to first sort or arrange the values of a data set in order of magnitude.
  - The median is computed differently in the case of odd and even numbers of values in a data set.
  - The median is not sensitive to outliers or skewed data. It is a better measure of central tendency when there are extremely l arge or small values in a data set.
- The Mode is
  - The most frequently occurring value in a data set.
  - In a data set, a value can occur more than once or many times. In such a case the data set will have 2 or more modes. This data set is then described as bi-modal or multi-modal.

Random Variable
A **random variable** (also called **random quantity**, **aleatory variable**, or **stochastic variable**) is a mathematical formalization of a quantity or object which depends on random events. It is a mapping or a function from possible outcomes (e.g., the possible upper sides of a flipped coin such as heads H and tails T) in a sample space (e.g., the set {H, T} to a measurable space, often the real numbers (e.g., in which 1 corresponding to H and -1 corresponding to T).

Categorical RV / Qualitative :
1. Nominal : {Male, Female} , {Heads , Tail}
2. Ordinal Ex : customer rating (5/4/3/2/1) - sequence of the variables matters, Ranking is important

Continuous RV / Quantitative :
1. Continuous Quantitative Variables
   Height : {170.23, 163.52, 152}
2. Discrete Quantitative Variables
   Age : {72, 54, 86, 32}

Independent Samples & Dependent Samples
Dependent variables are nothing but the variable which holds the phenomena which we are studying. Independent variables are the ones which through we are trying to explain the value or effect of the output variable (dependent variable) by creating a relationship between an independent and dependent variable.

CO2 | Gases | ....| --> AQI
Dependent variable AQI is derived from the set of independent variables CO2, Gases etc.

--Done--

Stats Lect 2 Notes

Wednesday, January 18, 2023 10:51 AM

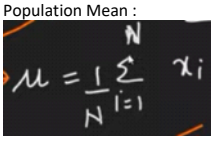
Measures of Dispersion

1. Variance : Measuring the spread of the data
2. Standard Deviation : Measure how far an element is away from the mean

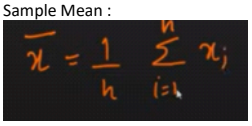
Mathematical Formula

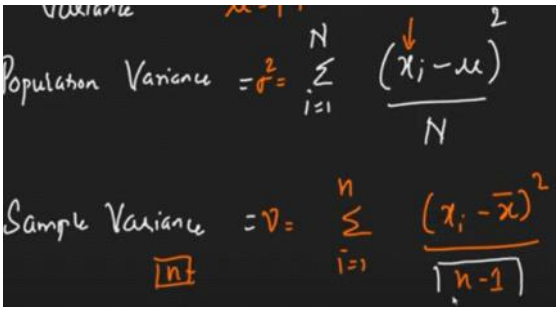
X is the random variable : {x1, x2, x3, x4, …. xn}

Population Mean :



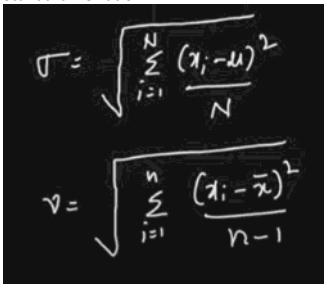
Sample Mean :





WHY n-1 : --> IMP

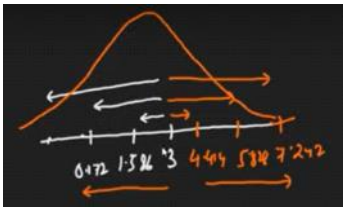
Standard Deviation :



Example : {1,2,3,4,5}
Mean (mu) = 3
Variance (sigma square) = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 / 5
= 4+1+0+1+4 / 5 =10/5 = 2

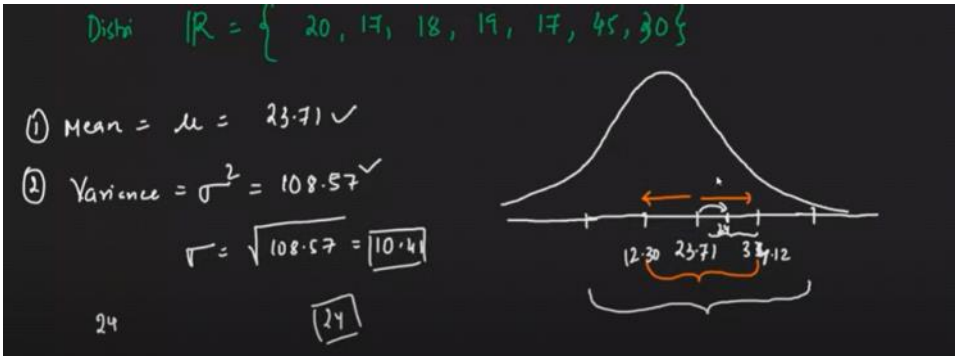
Hence sigma sqrt(2) = 1.414

Value of 1 standard deviation is 1.414

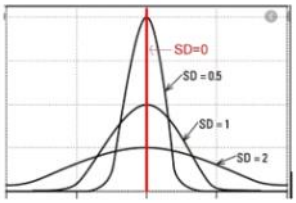


Variance defines the spread and std dev defines how much away from mean

Example 2 :

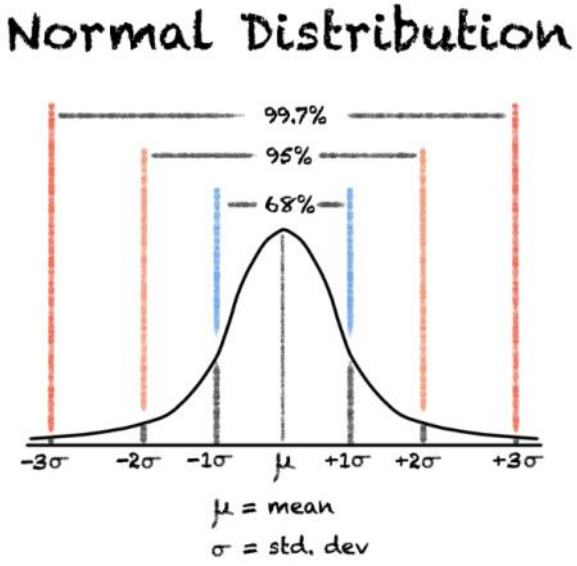


- As variance / std dev increases the spread of the distribution increases



- This distribution is called Gaussian / Normal Distribution / Bell Curve
- Most of the data in the world follows gaussian normal dist

Properties of a Gaussian Normal Distribution



Empirical rule

$\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68.27\%$
$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$
$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.73\%$

Probability of a random variable X to fall within 1 std dev to the left and 1 std dev to the right is 68%

68% of the total data points lies within 1 standard deviation of the data
95% of the total data points lies within the 2nd std dev of the data
99.7% of the total data points lies within the 3rd std dev of the data

Properties:

1. It is symmetric

A normal distribution comes with a perfectly symmetrical shape. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve.

2. The mean, median, and mode are equal

The middle point of a normal distribution is the point with the maximum frequency, which means that it possesses the most observations of the variable. The midpoint is also the point where these three measures fall. The measures are usually equal in a perfectly (normal) distribution.

3. Empirical rule

In normally distributed data, there is a constant proportion of distance lying under the curve between the mean and specific number of standard deviations from the mean. For example, 68.25% of all cases fall within +/- one standard deviation from the mean. 95% of all cases fall within +/- two standard deviations from the mean, while 99% of all cases fall within +/- three standard deviations from the mean.

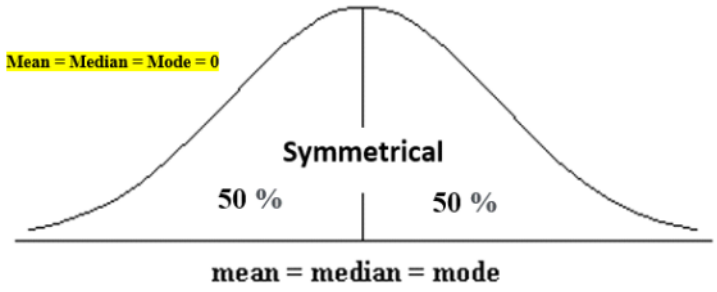
Shape of data: Skewness and Kurtosis

Skewness:

It's an important statistical technique that helps to determine asymmetrical behavior than of the frequency distribution, or more precisely, the lack of symmetry of tails both left and right of the frequency curve. A distribution or dataset is symmetric if it looks the same to the left and right of the center point.

The normal distribution helps to know a skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central

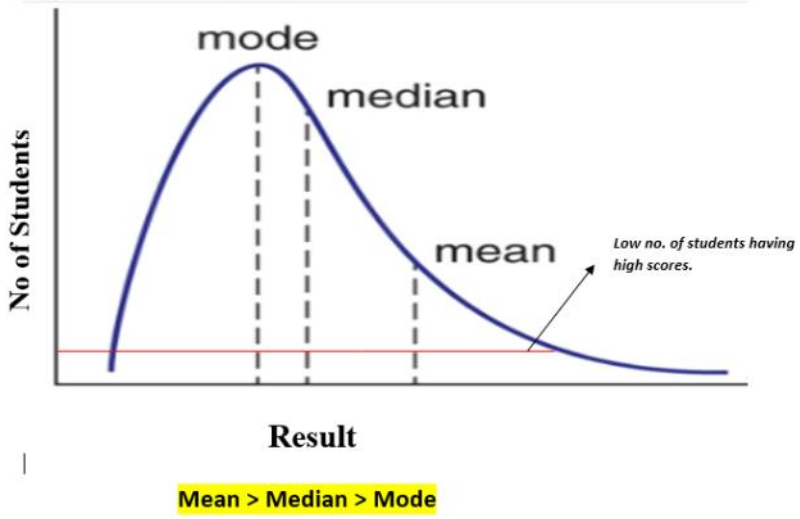
tendency lies in the middle.



When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness comes into the picture.

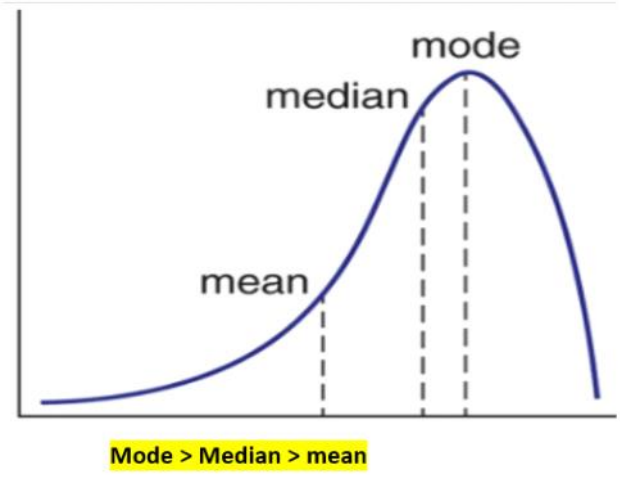
Types of skewness

1. Positive skewed or right-skewed



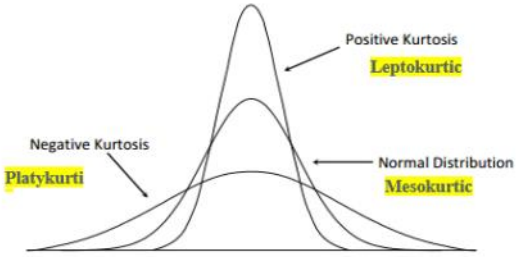
In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the highest value.
The extreme positive skewness is not desirable for distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

2. Negative skewed or left-skewed



Kurtosis

Kurtosis refers to the degree of presence of outliers in the distribution.
Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.



In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

Excess Kurtosis

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

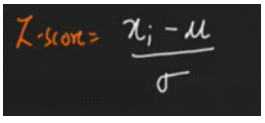
Excess kurtosis = Kurt – 3

Types of excess kurtosis

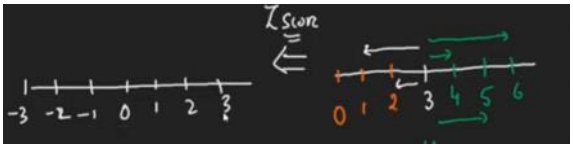
1. Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution), Leptokurtic (kurtosis > 3)
2. Mesokurtic (kurtosis same as the normal distribution), Mesokurtic (kurtosis = 3)
3. Platykurtic or short-tailed distribution (kurtosis less than normal distribution), platykurtic (kurtosis < 3)

Z Score

X = {1,2,3,4,5}
Mu mean = 3
Consider std dev sigma = 3



Examples : https://courses.lumenlearning.com/introstats1/chapter/the-standard-normal-distribution/

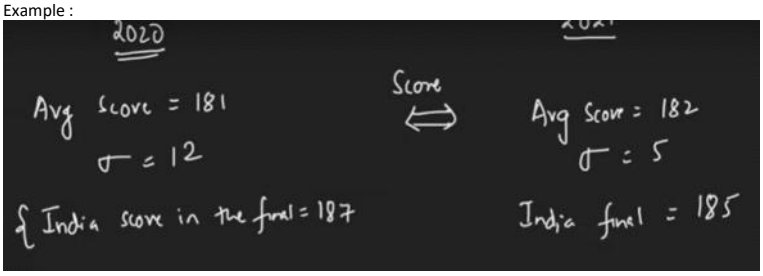


The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1. Any normal distribution can be standardized by converting its values into z scores. Z scores tell you how many standard deviations from the mean each value lies.

Why Z Score

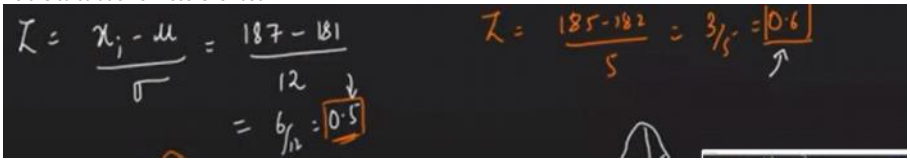
1. Standardization is used to bring all the features to the same scale. It is done using Z Score
2. Comparing scores of different distributions

Example :



Which year did we score better?

Do the calculation of Z score for both



Higher z score denotes better performance 0.5(2020) < 0.6(2021) hence in 2021 they performed better.

Probability
Experiment : Tossing a coin
Sample space : {H, T}

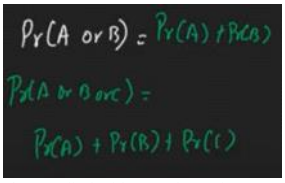
Experiment : Rolling a dice
Sample space : {1,2,3,4,5,6}

Pr = 1/6

Mutually Exclusive and Non Mutually Exclusive

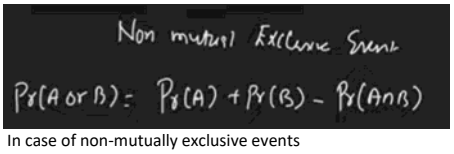
Two events are mutually exclusive or disjoint if they cannot both occur at the same time
Above both are ME examples

Additive rule :



This is called additive rule applies to mutually exclusive events

Example 2 : Taking out a card from a deck - which is K or Heart



In case of non-mutually exclusive events

Calculate the above.

Independent and Dependent Events
Dependent events influence the probability of other events – or their probability of occurring is affected by other events. Independent events do not affect one another and do not increase or decrease the probability of another event happening.

Multiplicative rule :

Dependent Event
Probability of getting a 1 in first throw and getting a 6 in the second throw in a rolling dice experiment

P(A and B) = P(A) * P(B)

Dependent Event
What is the probability of taking out the white marble and then the red marble from a bag of marbles having 5 marbles of diff colors.

P(A and B) = P(A) * P(B/A) --> Probability of B given A / Conditional Probability

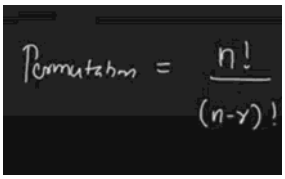
Permutation and Combination
A permutation is an act of arranging objects or numbers in order. Combinations are the way of selecting objects or numbers from a group of objects or collections, in such a way that the order of the objects does not matter.

A numeral of permutations when 'r' components are positioned out of a total of 'n' components is $_nP_r = n! / (n - r)!$.

Number of combinations when 'r' components are chosen out of a total of 'n' components is, $_nC_r = n! / [(r!) \times (n - r)!]$

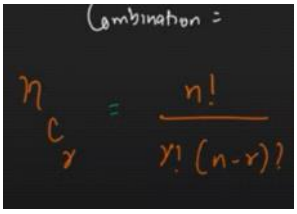
Example : You went to a zoo and there are 6 types of animals there --> Tiger , Lion, Chimpanzee, Giraffe, Monkey, Snake.
You are given a piece of paper and told to write down the first three animals that you see

First name --> 6 possible names
Second name --> 5 possible names
Third name --> 4 possible names
= 6*5*4 = 120 ways to fill the three names (order matters here)



n = total samples = 6
r = options to fill = 3

Combination : Lion Tiger Monkey is same as Tiger Monkey Lion and considered as 1 combination



--Done--

Stats Lect 3 Notes

Thursday, January 19, 2023 9:01 AM

5 Number summary

1. Min
2. First quartile (25%)
3. Median
4. Third quartile (75%)
5. Max

Percentile:
X percentile means X% of the entire distribution is less than X

Percentile of value X = Number of values below (less than) X / Sample size

Eg :{ 2,2,3,4,5,5,6,7,8,8,8,9,9,10}

n=14
What is the percentile of 7

Percentile of 7 = 7 / 14 = 0.5 *100 = 50 percentile (i.e. median)

What is the 25th percentile of the given distribution

Value = (Percentile / 100)* (n+1) = (25/100) * (14+1) = 3.75

Removing Outliers

Consider the sorted distribution : {1,2,2,2,3,3,4,5,5,6,6,6,6,7,8,8,9,27}

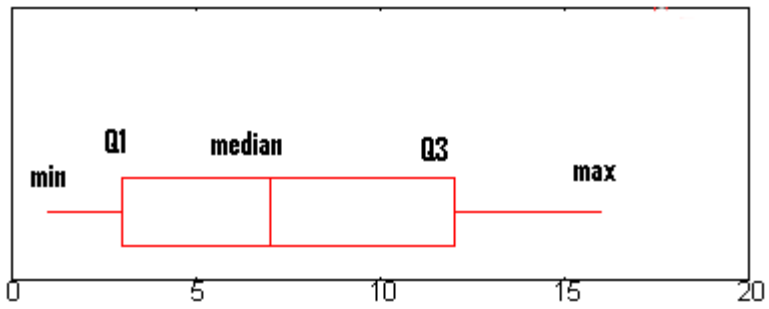
First , sort the dataset

Lower Fence = Q1 - 1.5 IQR
Q1 = 25 percentile
IQR = Inter quartile range = Q3 - Q1
Q3 = 75 percentile

Q1 (Value at 25) = (25/100)*(18+1) = 4.75 th position = 5th position = 3
Q3 (Value at 75) = (75/100)*(18+1) = 14.25 th position = 14th position = 7
IQR = Q3 - Q1 = 7 - 3 = 4

Lower Fence = 3 - (1.5*4) = -3
Higher Fence = 7 + (1.5*4) = 13

No values x < -3 but we have one value x > 13 i.e. 27 hence 27 is an outlier



COVARIANCE

Covariance **quantifies the relationship** between two random variables - independent vs dependent or two / more independent variables.

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

- Solve an example

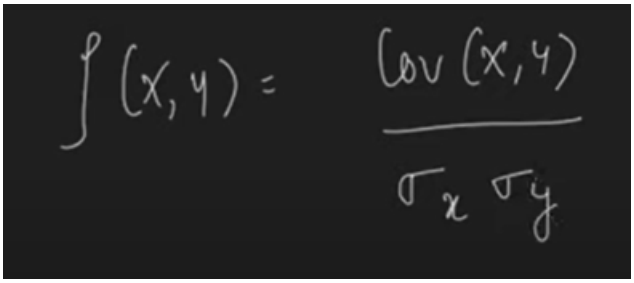
1. Positive covariance value --> positively correlated
2. Negative covariance value --> Negatively correlated
3. Nearing zero/ Zero covariance value --> Not correlated

Disadvantages of covariance

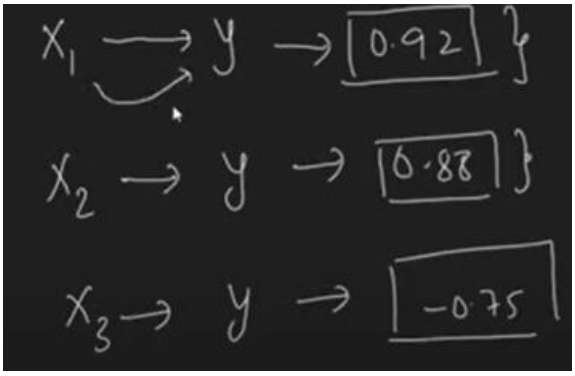
- Able to quantify the relation (+ve or -ve) but not able to calculate how much is the correlation value.

Restricting the correlation value between -1 to 1 that signifies that closer to +1 it is highly correlated positively and closer to -1 is highly correlated negatively and by what value is given by **Pearson's correlation**

Pearson's correlation coefficient is given by :

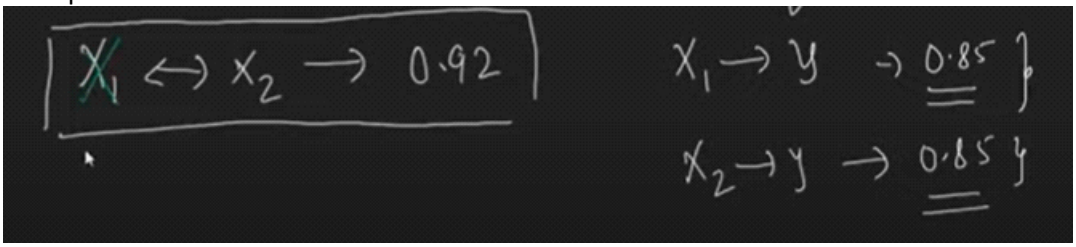


Example :

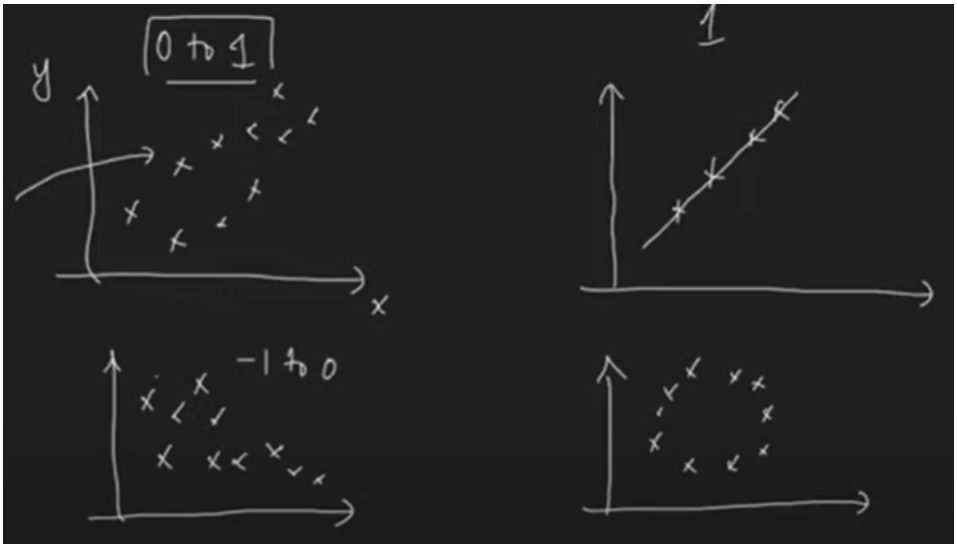


>> X1 is highly correlated to y as compared to X2

Example :

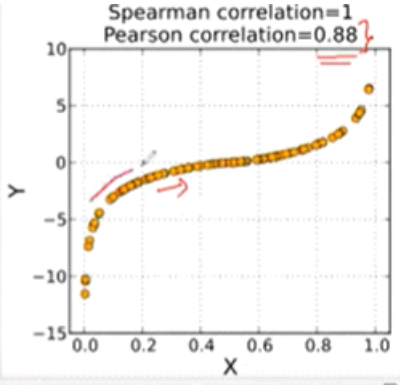


>> In this case since X1 and X2 are mutually correlated and X1 and X2 are correlated with y as well, we can omit either X1 or X2 since there is no value addition using both the features.



>> last graph is no correlation data

Spearman's Rank Correlation



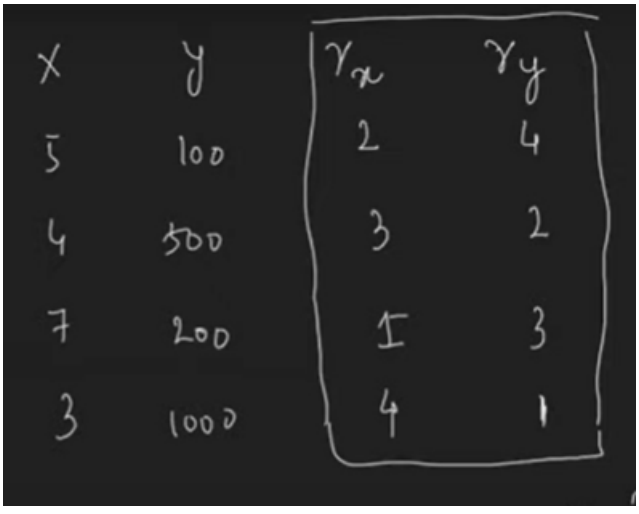
This correlation captures the more non - linear properties
Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

For a sample of size *n*, the *n* raw scores X_i, Y_i are converted to ranks $\mathrm{R}(X_i), \mathrm{R}(Y_i)$, and r_s is computed as

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}},$$

where

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,
$\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))$ is the covariance of the rank variables,
$\sigma_{\mathrm{R}(X)}$ and $\sigma_{\mathrm{R}(Y)}$ are the standard deviations of the rank variables.



When to use Spearman ? Which one is better - Pearson or Spearman ?

Sometimes we need to capture the non-linear properties in the data and we use Spearman for that. But there is no underlying rule and we should try to apply both the correlation metrices and compare results.

-- Done --

Stats Lect 4 Notes

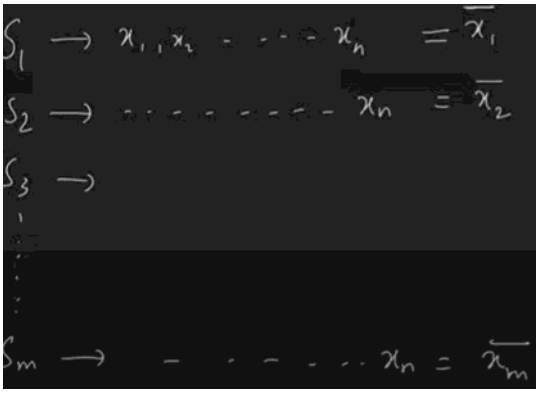
Thursday, January 19, 2023 8:15 AM

Inferential Statistics

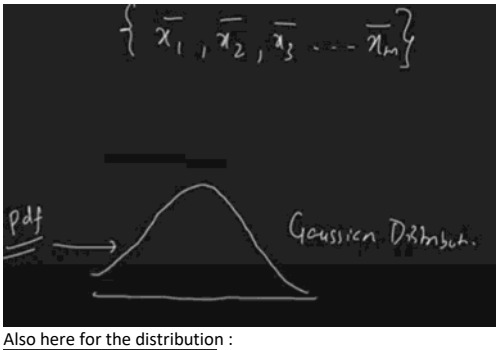
Performing tests on sample data to derive conclusions about the population.

Central Limit Theorem: If we have random variable, it may or may not belong to a gaussian distribution that has a mean mu and standard deviation sigma.

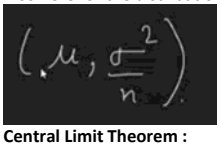
Suppose we have m samples from the population, each sample having n random variables and compute the mean of each sample



Next for a set of data of the computed means, if we plot them , we get a gaussian distribution , it is called a pdf function and it follows a curve as shown below.



Also here for the distribution :



Central Limit Theorem :

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
- Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
- Sample mean is approximately equal to Population mean
- Sample std dev is equal to sigma square / n where sigma is population std dev
- A sufficiently large sample size can predict the characteristics of a population more accurately.

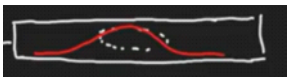
Hypothesis Testing

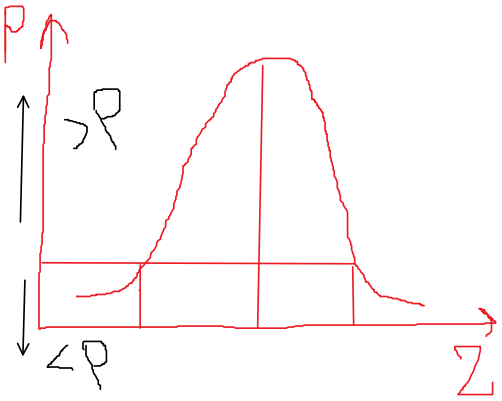
- Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.
- First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0.
- An alternative hypothesis (denoted Ha), which is the opposite of what is stated in the null hypothesis, is then defined.
- The hypothesis-testing procedure involves using sample data to determine whether or not H0 can be rejected.
- If H0 is rejected, the statistical conclusion is that the alternative hypothesis Ha is true.

cc

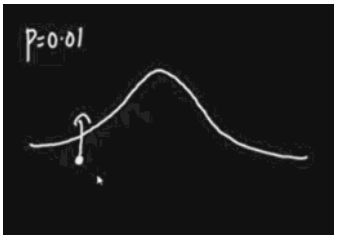
Significance value and P value

Example : Consider a case of space bar and the distribution of touches of the space bar. The distribution will look somewhat like below, where majority of the touches tend to happen at the center and less touches towards the end.

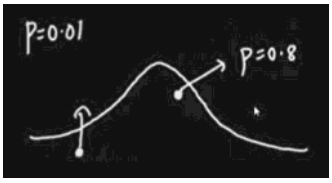




This distribution of touches on a space bar follow a gaussian distribution
Now, if p-value is 0.01 it means that out of 100 touches 1 time we are touching in the marked region

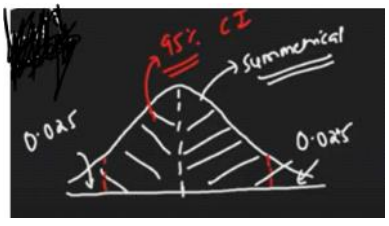


Obviously, the center portion will have a higher p-value, so p=0.8 meaning out of 100 touches 80 times we are touching in the marked point on the space bar



Consider a case where significance value denoted by alpha a = 0.05
Now consider the gaussian distribution, it is symmetrical around the mean.

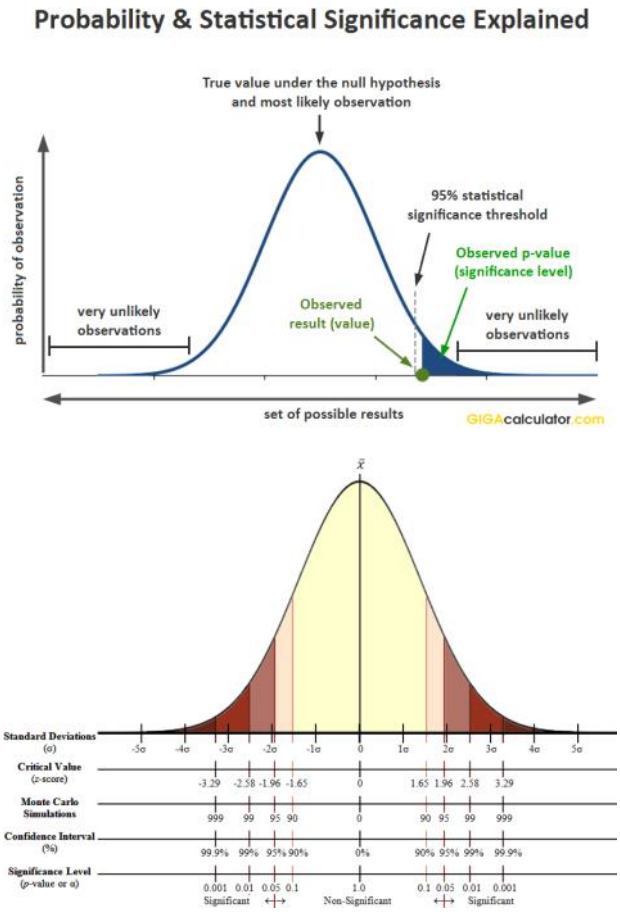
Considering the area under the gaussian as 1,
This denotes that the confidence interval = (1-0.05)*100 = 95 %



Now let us assume we did some statistical tests and got the p value as 0.07, which are would it lie??
--> It would lie in the main region (not the tail part) indicating that p value of 0.07 is within the confidence interval of 95%

If p = 0.001 --> falls in tail part

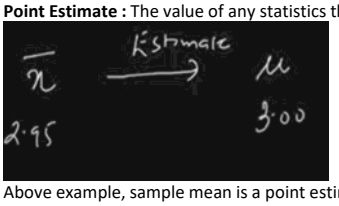
If p value is within the confidence interval then we "ACCEPT THE NULL HYPOTHESIS"
And hence we "REJECT THE ALTERNATE HYPOTHESIS"



Types of Tests:

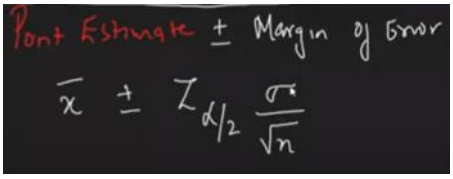
1. Z test {Comparison of Mean}
2. t Test {Comparison of Mean}
3. F test {Comparison of Variance}
4. ANOVA test { Analysis of Variance}
5. Chi Square {Comparison between two categorical variables}

Point Estimate : The value of any statistics that estimates the value of a parameter is called point estimate



Above example, sample mean is a point estimate of population mean.

If population standard deviation is given , the formula for Confidence Interval is given by
CI =



n = sample population
Alpha = significance value

Example 1: On the verbal section of CAT exam, the standard dev is 100 (population std dev). The sample of 25 test takers is taken, they have a mean of 520. Construct a 95% CI about the mean.

n = 25
Sigma = 100
Alpha = 0.05
CI = 95%

$$520 + Z_{0.025}\left(\frac{100}{\sqrt{25}}\right)$$

To get Z0.025

1. Subtract 0.025 from 1 --> 1-0.025 = 0.975
2. Go to the Z table (online) and check for 0.975 in the table
3. Map to the right and top to get the number --> 1.9 and 0.06
4. The Z score is 1.96 and -1.96 since the gaussian dist is symmetrical

$$520 + Z_{0.025}\left(\frac{100}{\sqrt{25}}\right) = 520 + (1.96)\left(\frac{100}{\sqrt{25}}\right) = 559.2$$

$$520 - (1.96)\left(\frac{100}{\sqrt{25}}\right) = 480.8$$

So from any hypo testing the mean value should lie between 559.2 and 480.8 , to accept the null hypothesis with 95% confidence interval.

$$[480.8 \longleftrightarrow 559.2]$$

↳ to accept Null Hypothesis

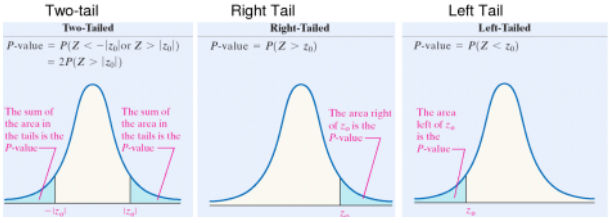
Solution :

H0 : Mean = 520 (mu)
Ha : Mean != 520

Since mean of 25 test takers is 520, and it falls in the range of mean , we "ACCEPT THE NULL HYPOTHESIS"

P-Value Approach

Assume that the null hypothesis is true.
The P-Value is the probability of observing a sample mean that is as or more extreme than the observed.
How to compute the P-Value for each type of test:
Step 1: Compute the test statistic $z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$



Example 2: On the verbal section of CAT, a sample of 25 test takers has a mean of 520 with standard deviation of 80.
Construct 95% confidence interval about the mean
(Population Std Dev not given but Sample std dev is given)

So we use t test instead of Z test

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

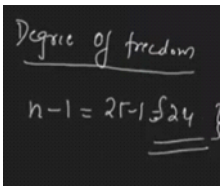
Sample sd = s

$\bar{x} = 520$, $n = 25$, $s = 80$, $\alpha = 0.05$

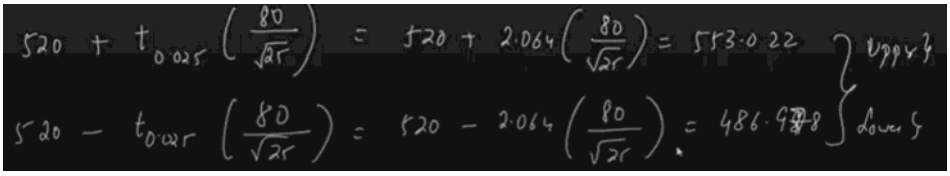
$$520 + t_{0.025}\left(\frac{80}{\sqrt{25}}\right)$$

To compute t0.025:

1. When std dev of population is not given, we have to compute "degree of freedom"



2. DoF = n - 1 = 25 - 1 = 24
3. Check for the t table online
4. Refer the two tail on the column titles and check for 0.05 and Dof in thr rows as 24
5. The corresponding value in the table is called t value --> 2.064



Accept null hypothesis
"In Z test n value should be greater than or equal to 30" --> mandatory

ToDo: Examples

1] One Sample Z Test

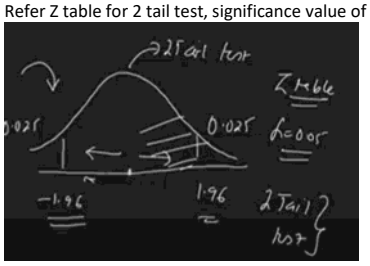
The average IQ is 100 and std dev is 15. A company test a new medication to check whether it increases or decreases the IQ. After taking the meds, for a sample of 30 participants the IQ was checked and the mean IQ = 140. Did medication affect intelligence? Significance value alpha = 0.05

Z test will be used because
1. Population std dev is given
2. Sample size is >= 30

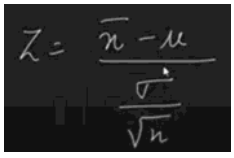
H0 : Medication does not affect the intelligence Mean = 100
Ha : Medication affects the intelligence ; Mean != 100

Since intelligence change means either it increases or decreases it is a 2 tail test

3. Refer Z table for 2 tail test, significance value of 0.05,

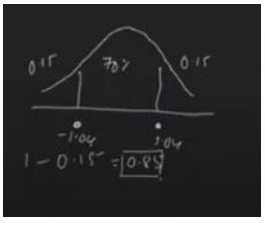


4. Calculate Z test statistics



Z = 14.60
5. Range is -1.96 to 1.96 and we got Z = 14.60 , so we reject the null hypothesis
6. Hence, there is an effect of medication on the intelligence.

Example 2 Test the same with alpha = 0.30
CI = 70%



Still the test will fail --> reject null hypothesis

Example 3 : The average IQ is 100 and std dev is 15. A company test a new medication to check whether it increases or

decreases the IQ. After taking the meds, for a sample of 30 participants the IQ was checked and the mean IQ = 115. Did medication affect intelligence? Significance value alpha = 0.05

2] One Sample t Test

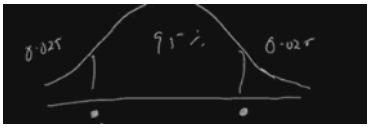
{Population std dev is not given}
{sample number need not be more than or equal to 30}

Example 1: In the population the average IQ is 100. A team of scientists want to test a new medication to see if it has a positive or negative effect on intelligence, or no effect at all. A sample of 25 participants have taken the medication and has a mean IQ of 140 with std dev as 20. Did the medication affect the intelligence? Alpha 0.05

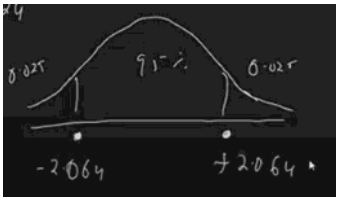
H0 : Mean = 100
Ha : Mean != 100

Degree of freedom : 24.

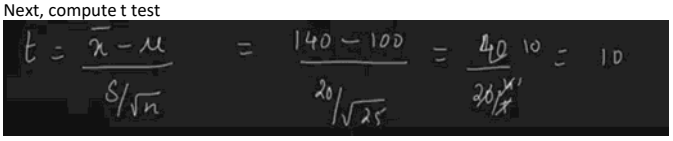
It will be a two tail test



Check for the points on x axis from the t table



Next, compute t test

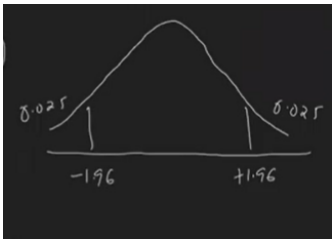


T value is not in the range of -2.064 to + 2.064 so null hypothesis is rejected and the intelligence has increased (since > 2.064)

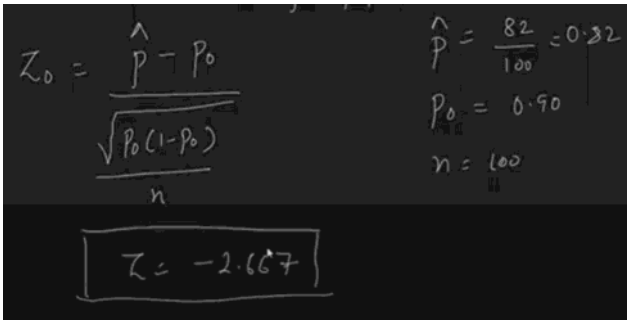
3] One Sample Z test for proportion

Example 1: A survey claims that 9 out of 10 doctors recommend aspirin for their patients with headache. To test this claim, a random sample of 100 doctors is obtained, out of this 100 doctors 82 indicate that they recommend Aspirin. Is this claim accurate? Using alpha = 0.05

H0 : p = 9/10 = 0.9
Ha : p != 0.9



Calculate Z statistics



-2.667 < -1.96
So we reject the null hypothesis and accept the alternate hypothesis

So the claim is inaccurate.

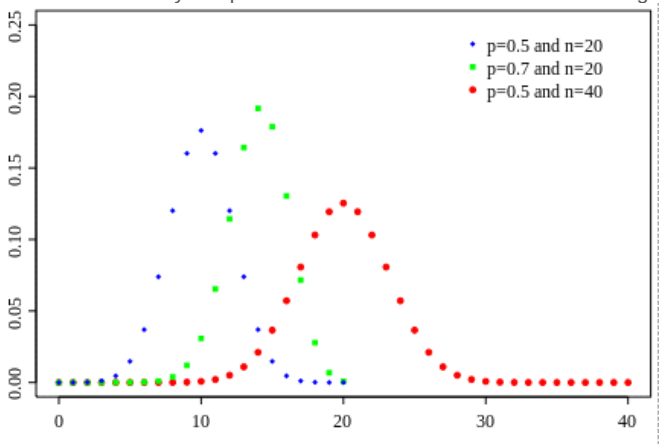
--Done--

Stats Lect 5 Notes - pending

Monday, January 23, 2023 9:21 AM

1. **Bernoulli Distribution**
   It's a discrete probability distribution of a random variable which takes the value 1 with probability **p**
   and the value 0 with probability q = 1-p

2. **Binomial Distribution**
   A **binomial distribution** can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has **two possible outcomes** (the prefix "bi" means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.



- The first variable in the binomial formula, n, stands for the number of times the experiment runs.
- The second variable, p, represents the probability of one specific outcome.

Binomial distributions must also meet the following three criteria:

a. **The number of observations or trials is fixed.** In other words, you can only figure out the probability of something happening if you do it a certain number of times. This is common sense—if you toss a coin once, your probability of getting a tails is 50%. If you toss a coin a 20 times, your probability of getting a tails is very, very close to 100%.
b. **Each observation or trial is** independent. In other words, none of your trials have an effect on the probability of the next trial.
c. The **probability of success** (tails, heads, fail or pass) is **exactly the same** from one trial to another.

The binomial distribution formula is:

$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}$

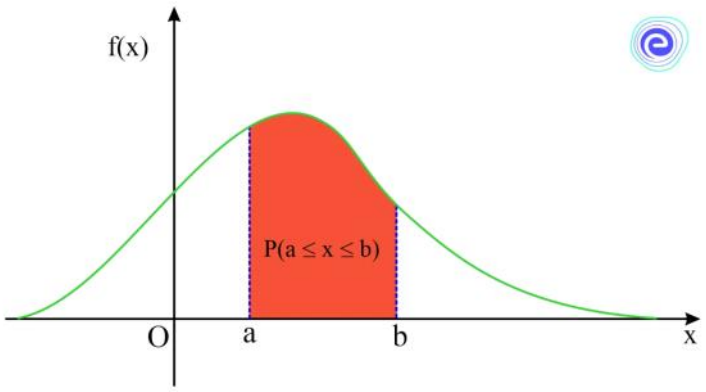
$$P(X) = \frac{n!}{(n-X)!\,X!} \cdot (p)^X \cdot (q)^{n-X}$$

Where:
b = binomial probability
x = total number of "successes" (pass or fail, heads or tails etc.)
P = probability of a success on an individual trial
n = number of trials

Example 1

Q. A coin is tossed 10 times. What is the probability of getting exactly 6 heads?
I'm going to use this formula: $b(x; n, P) - {}_nC_x * P_x * (1 - P)_{n-x}$
The number of trials (n) is 10
The odds of success ("tossing a heads") is 0.5 (So 1-p = 0.5)
x = 6
$P(x=6) = {}_{10}C_6 * 0.5\text{^}6 * 0.5\text{^}4 = 210 * 0.015625 * 0.0625 = 0.205078125$

3. **PDF(Probability Density Function) and CDF(cumulative distribution function) (Kernel density estimators)**
   Probability Density Function: A probability density function calculates the likelihood that the value of a random variable will fall within a specified range.
   For continuous random variables, the probability density function is used.
   Like the probability density function, the probability mass function is used for discrete random variables.
   The shape of the graph of a probability density function is a bell curve.



The ***probability density function*** **(pdf)**, denoted f, of a continuous random variable X satisfies the following:

1. $f(x) \geq 0$, for all $x \in \mathbb{R}$
2. f is piecewise continuous
3. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
4. $P(a \leq X \leq b) = \int_{b}^{a} f(x)\,dx$

Example

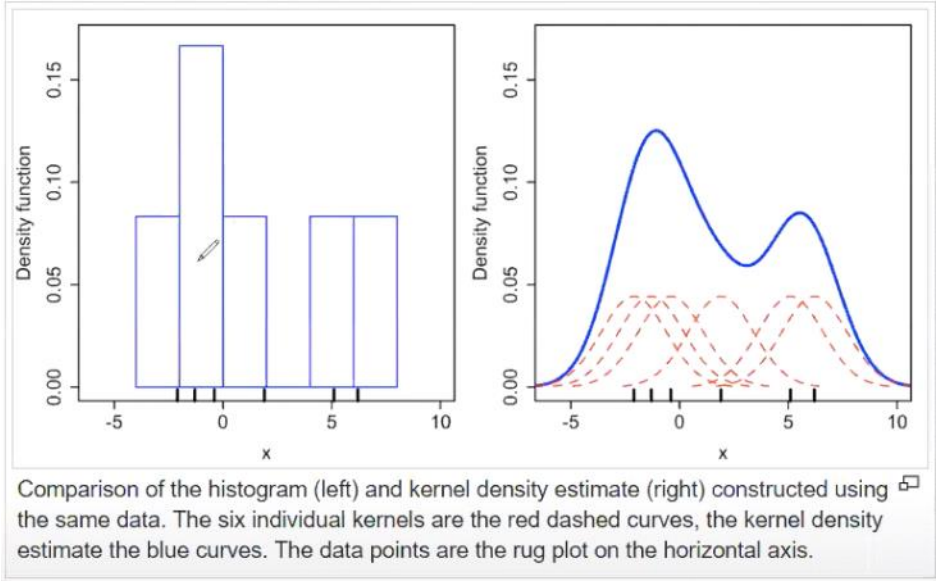
Let us consider a probability density function of some continuous random variable in f(x)=2x–1, when 0 < x ≤2.0
If we need to find the P(0.5<X<1),P(0.5<X<1), we need to follow the steps given below:

- First, integrate the given function
  $\int_{0.5}^{1}(2x-1) = [x^2]_{0.5}^{1}$
- Now, apply the limits 0.5 and 1 and subtract the values obtained
  $(1)^2-(0.5)^2 = 1-0.25 = 0.75$

Thus, the result obtained gives the probability that the continuous random variable is within the given limits.
Here, the probability of the given continuous random variable lying between 0.5 and 1 is 0.75

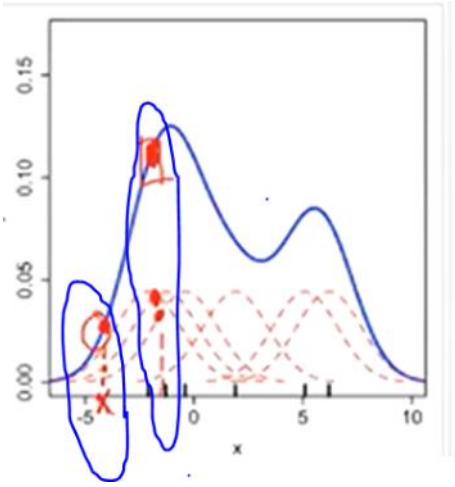
Kernel Density Estimator

Kernel density estimator helps to get smoother version of histogram

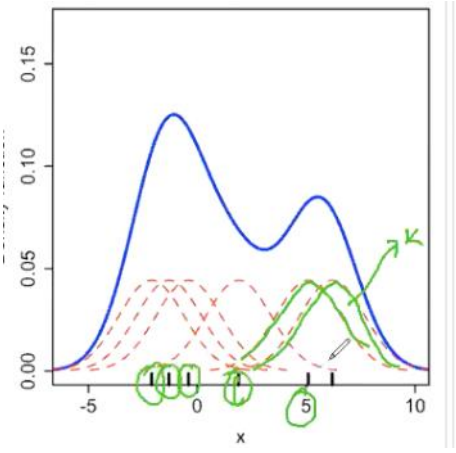


Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The six individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

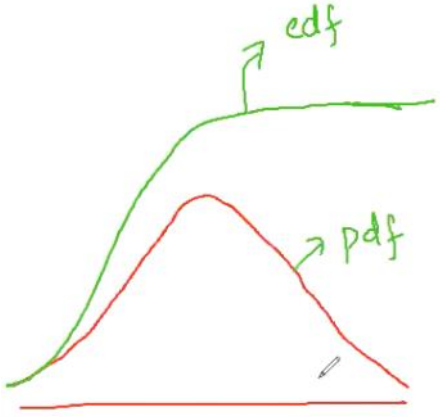
In right image we see each dotted gaussian curve is created for each point considering that point as a mean
And the blue line is constructed by taking the sum of all the values available for the particular point



All the small curves present in the diagram are called kernels and for every point there is a different kernel



CDF cumulative distributed function



Cumulative distributed function is an addition of PDF.
So for first half curve of PDF, the CDF value is increasing exponentially and for second half curve of PDF the CDF value has small increment and the graph looks steady.

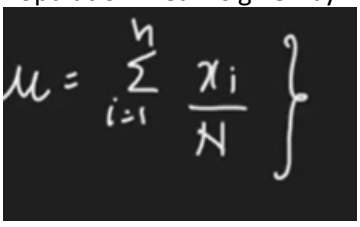
4. **Power Law (Pareto Distribution) Box Cox transformation**
5. **Chebyshev Inequality**
6. **Q-Q plot**

Stats Lect 6 Notes

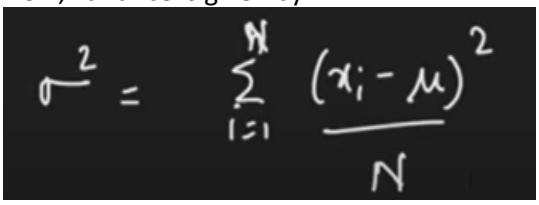
Monday, January 23, 2023 3:33 PM

Why Sample variance is divided by n-1

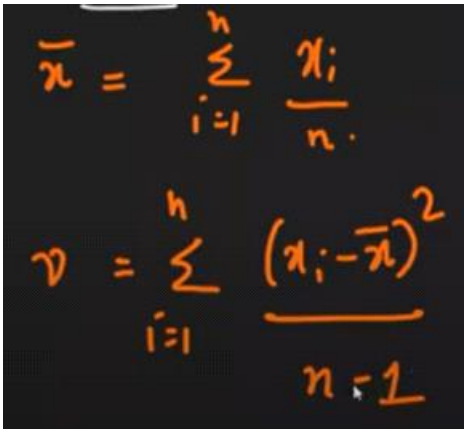
Population mean is given by



Now, variance is given by



Same for sample ,



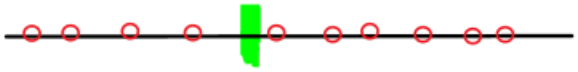
Here the question is why is variance for sample divided by n-1 and n.

- Dividing variance by n-1 is called unbiased estimation.
- Dividing by n is called biased estimation.

Consider a population and its mean as below

The red points are the points in the population and the green mark is the population mean

Now consider we select a biased sample from the population.



Here the sample mean will be far away from the population mean.
So we divide by n-1 to correct the bias.

Here n-1 is tried and tested for many cases and was identified as the best correction by mathematicians.

Paramteric vs Nonparametric

- A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.
- A model where the number of parameters is not determined prior to training. Nonparametric does not mean that they have NO parameters! On the contrary, nonparametric models (can) become more and more complex with an increasing amount of data.

From <https://github.com/jayinai/ml-interview>

Chi Square
It is a non-parametric (not involving any assumptions as to the form or parameters of a frequency distribution.) test that is performed on categorical data.

Example : In the 2000 US Census the age of an individual in a small town were found to be the following:

Less than 18	18-35	Greater than 35
20%	30%	50%

In 2010, ages of n=500 individuals were sampled and below are the results.

Less than 18	18-35	Greater than 35
121	288	91

Using alpha = 0.05 , can you conclude the distribution of ages has been changed in 10 years?

Solution:
Null Hypothesis: The data meets the age distribution, no change in the past 10 years
Alternate Hypothesis : Distribution of age has changed

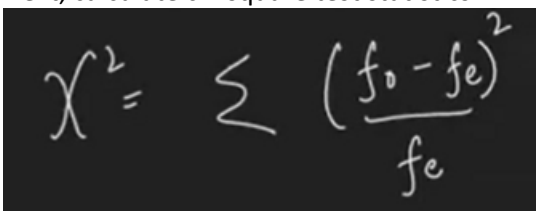
Alpha = 0.05 hence Confidence Interval = 95%
Degree of Freedom = 3-1 (No. of categories -1) = 2

Now, refer the chi square table df = 2 and 0.05 significance value = 5.99

Decision rule:

Id chi square is greater than 5.99 , then we reject the null hypothesis.

Next, calculate chi square test statistics



fe= frequency of expected
fo = frequency of observed

	Less than 18	18-35	Greater than 35
fo	121	288	91
fe	= (500*20%) = 100	=(500*30%) =150	=(500*50%) =250
fo-fe	121-100 = 21	288-150 =138	91-250 = -159

Chi Square = [(21*21)/100 + (138*138)/150 + (-159 * -159)/250] = 232.94 > 5.99
Hence we reject the null hypothesis

"Distribution of age has changed"

Example : 500 elementary school girls and boys are asked which is their fav color?
Results are as below:

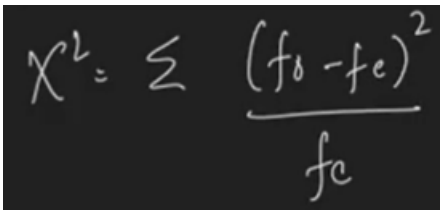
	Blue	Green	Pink	
Boys	100	150	20	270
Girls	20	30	180	230
	120	180	200	500 = n

Using alpha = 0.05, would you conclude that there is a relationship between gender and color.

H0 : Gender and fav color are not related
H1 : Gender and fav color are related

Degree of Freedom = (rows -1)*(columns-1) = (2-1)*(3-1) = 2

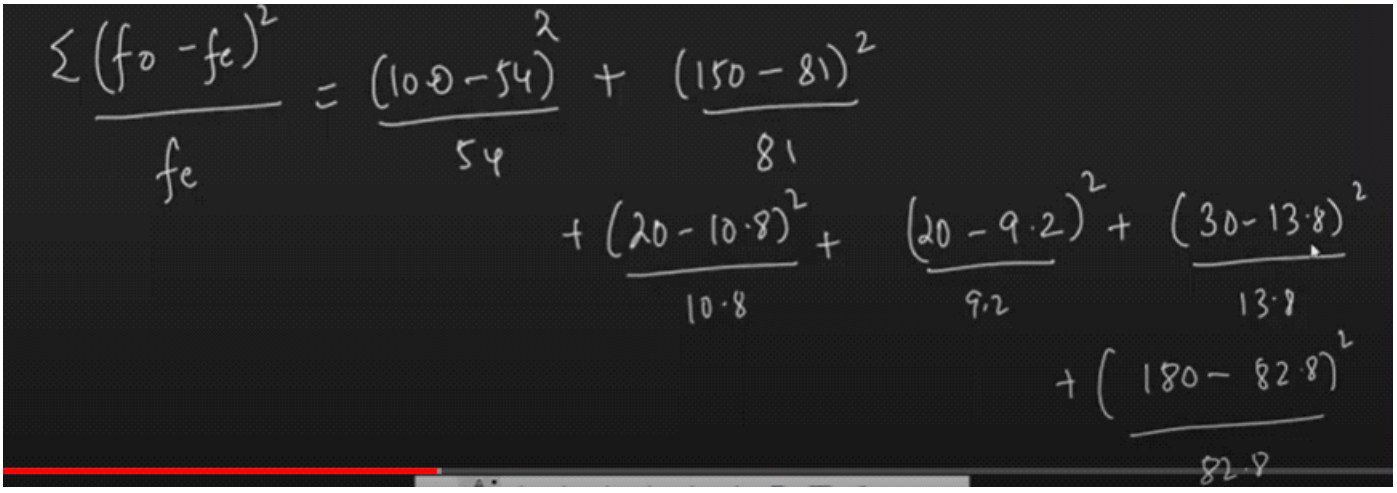
Decision rule : Chi square > 5.99 then reject the null hypothesis.



Here fe = fc*fr/n { f_column * f_row / n }

fe(Boys, Blue) = 100*270 / 500 = 54
fe(Boys, Green) = 150 * 270 / 500 = 81
fe(Boys, Pink) = 20 * 270 / 500 = 10.8

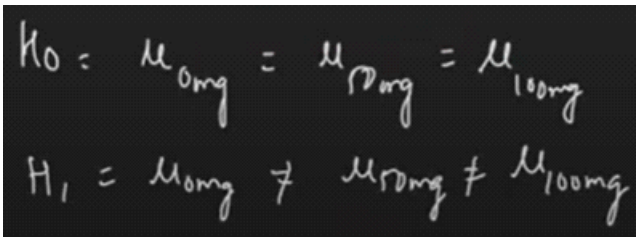
fe(Girls, Blue) = 20 * 230 / 500 = 9.2
fe(Girls, Green) = 30 * 230 / 500 = 13.8
fe(Girls, Pink) = 180 * 230 / 500 = 82.8



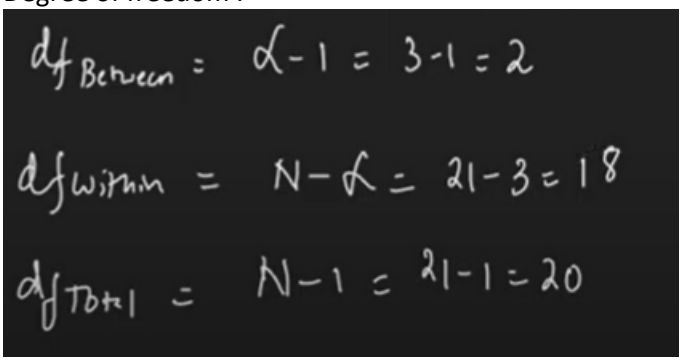
Chi square = 251.6 > 5.99
Hence null hypothesis is rejected
Hence gender and age are related.

Example : Researchers wants to test an anti-anxiety medication. The split participants in three condition (0mg, 5mg, 100mg) then anxiety level is checked on scale 1-10, are there any differences between the three conditions
Alpha = 0.05

0mg	50mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2



Degree of freedom :



Decision rule : (df_between, df_within) = (2,18)
Open f distribution table in the browser and look for df1 = 2 and df2 = 8 @ alpha = 0.05 --> 3.5546

Calculate F statistics

--- complicated - skipping.

Monday, January 23, 2023 12:22 PM

Data visualization 33 hrs 11 days
EDA 13.5 hrs 5 days

16 days

Probability questions with solutions

Wednesday, January 25, 2023 9:18 AM

1. **Harvard Law School courses often have assigned seating to facilitate the "Socratic method." Suppose that there are 100 first year Harvard Law students, and each takes two courses: Torts and Contracts. Both are held in the same lecture hall (which has 100 seats), and the seating is uniformly random and independent for the two courses.**

   (a) Find the probability that no one has the same seat for both courses (exactly; you should leave your answer as a sum).

   Let N be the number of students in the same seat for both classes. The problem has the same structure as the *de Montmort matching problem* from lecture. Let E_j be the event that the j^{th} student sits in the same seat in both classes. Then

$$P(N=0) = 1 - P\left(\bigcup_{j=1}^{100} E_j\right)$$

   By symmetry, inclusion-exclusion gives

$$P\left(\bigcup_{j=1}^{100} E_j\right) = \sum_{j=1}^{100} (-1)^{j-1} \binom{100}{j} P\left(\bigcap_{k=1}^{j} E_k\right)$$

   The j-term intersection event represents j particular students sitting pat throughout the two lectures, which occurs with probability $(100-j)!/100!$. So

$$P\left(\bigcup_{j=1}^{100} E_j\right) = \sum_{j=1}^{100} (-1)^{j-1} \binom{100}{j} \frac{(100-j)!}{100!} = \sum_{j=1}^{100} (-1)^{j-1}/i!$$

$$P(N=0) = 1 - \sum_{j=1}^{100} \frac{(-1)^{j-1}}{j!} = \sum_{j=0}^{100} \frac{(-1)^j}{j!}.$$

   (b) Find a simple but accurate approximation to the probability that no one has the same seat for both courses.

Define I_i to be the indicator for student i having the same seat in both courses, so that $N = \sum_{i=1}^{100} I_i$. Then $P(I_i = 1) = 1/100$, and the I_i are weakly dependent because

$$P((I_i = 1) \cap (I_j = 1)) = \left(\frac{1}{100}\right)\left(\frac{1}{99}\right) \approx \left(\frac{1}{100}\right)^2 = P(I_i = 1)P(I_j = 1)$$

So N is close to Pois(λ) in distribution, where $\lambda = E(N) = 100EI_1 = 1$. Thus,

$$P(N = 0) \approx e^{-1}1^0/0! = e^{-1}.$$

This agrees with the result of (a), which we recognize as the Taylor series for e^x, evaluated at $x = -1$.

(c) Find a simple but accurate approximation to the probability that at least two students have the same seat for both courses.

Using a Poisson approximation, we have

$$P(N \geq 2) = 1 - P(N = 0) - P(N = 1) \approx 1 - e^{-1} - e^{-1} = 1 - 2e^{-1}.$$

2. **There are 100 passengers lined up to board an airplane with 100 seats (with each seat assigned to one of the passengers). The first passenger in line crazily decides to sit in a randomly chosen seat (with all seats equally likely). Each subsequent passenger takes his or her assigned seat if available, and otherwise sits in a random available seat. What is the probability that the last passenger in line gets to sit in his or her assigned seat?**

   This is a common interview problem, and a beautiful example of the power of symmetry.) Hint: Call the seat assigned to the jth passenger in line "seat j" (regardless of whether the airline calls it seat 23A or whatever). What are the possibilities for which seats are available to the last passenger in line, and what is the probability of each of these possibilities?

   The seat for the last passenger is either seat 1 or seat 100; for example, seat 42 can't be available to the last passenger since the 42nd passenger in line would have sat there if possible. Seat 1 and seat 100 are equally likely to be available to the last passenger, since the previous 99 passengers view these two seats symmetrically. So the probability that the last passenger gets seat 100 is 1/2.

3. **Raindrops are falling at an average rate of 20 drops per square inch per minute. What would be a reasonable distribution to use for the number of raindrops hitting a particular region measuring 5 inches2 in t minutes? Why? Using your chosen distribution, compute the probability that the region has no rain drops in a given 3 second time interval. A reasonable choice of distribution is P**

   A reasonable choice of distribution is Pois(λt), where λ = 20 · 5 = 100 (the average number of raindrops per minute hitting the region). Assuming this distribution, P(no raindrops in 1/20 of a minute) = $e^{-100/20}(100/20)^0/0! = e^{-5} \approx 0.0067$

4. **Let X be a random day of the week, coded so that Monday is 1, Tuesday is 2, etc. (so X takes values 1, 2,..., 7, with equal probabilities). Let Y be the next day after X (again represented as an integer between 1 and 7). Do X and Y have the same distribution? What is P(X)**

   Yes, X and Y have the same distribution, since Y is also equally likely to represent any day of the week.

However, X is likely to be less than Y . Specifically, P(X < Y) = P(X != 7) = 6 / 7 .

In general, if Z and W are independent r.v.s with the same distribution, then P(Z < W) = P(W < Z) by symmetry.

Here though, X and Y are dependent, and we have P(X < Y) = 6/7, P(X = Y) = 0, P(Y < X) = 1/7

5. **For a group of 7 people, find the probability that all 4 seasons (winter, spring, summer, fall) occur at least once each among their birthdays, assuming that all seasons are equally likely.**

Let A_i be the event that there are no birthdays in the ith season. The probability that all seasons occur at least once is $1 - P(A_1 \cup A_2 \cup A_3 \cup A_4)$. Note that $A_1 \cap A_2 \cap A_3 \cap A_4 = \emptyset$. Using the inclusion-exclusion principle and the symmetry of the seasons,

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup A_4) &= \sum_{i=1}^{4} P(A_i) - \sum_{i=1}^{3} \sum_{j>i} P(A_i \cap A_j) \\ &\quad + \sum_{i=1}^{3} \sum_{j>i} \sum_{k>j} P(A_i \cap A_j \cap A_k) \\ &= 4P(A_1) - 6P(A_1 \cap A_2) + 4P(A_1 \cap A_2 \cap A_3). \end{aligned}$$

We have $P(A_1) = (3/4)^7$. Similarly,

$$P(A_1 \cap A_2) = \frac{1}{2^7} \text{ and } P(A_1 \cap A_2 \cap A_3) = \frac{1}{4^7}.$$

Therefore, $P(A_1 \cup A_2 \cup A_3 \cup A_4) = 4(\frac{3}{4})^7 - \frac{6}{2^7} + \frac{4}{4^7}$. So the probability that all 4 seasons occur at least once is $1 - \left(4(\frac{3}{4})^7 - \frac{6}{2^7} + \frac{4}{4^7}\right) \approx 0.513$.

6. **Alice attends a small college in which each class meets only once a week. She is deciding between 30 non-overlapping classes. There are 6 classes to choose from for each day of the week, Monday through Friday. Trusting in the benevolence of randomness, Alice decides to register for 7 randomly selected classes out of the 30, with all choices equally likely. What is the probability that she will have classes every day, Monday through Friday?**

Direct Counting Method: There are two general ways that Alice can have class every day: either she has 2 days with 2 classes and 3 days with 1 class, or she has 1 day with 3 classes, and has 1 class on each of the other 4 days. The number of possibilities for the former is $\binom{5}{2}\binom{6}{2}^2 6^3$ (choose the 2 days when she has 2 classes, and then select 2 classes on those days and 1 class for the other days). The number of possibilities for the latter is $\binom{5}{1}\binom{6}{3}6^4$. So the probability is

$$\frac{\binom{5}{2}\binom{6}{2}^2 6^3 + \binom{5}{1}\binom{6}{3}6^4}{\binom{30}{7}} = \frac{114}{377} \approx 0.302.$$

Inclusion-Exclusion Method: we will use inclusion-exclusion to find the probability of the complement, which is the event that she has at least one day with no classes. Let $B_i = A_i^c$. Then

$$P(B_1 \cup B_2 \cdots \cup B_5) = \sum_i P(B_i) - \sum_{i<j} P(B_i \cap B_j) + \sum_{i<j<k} P(B_i \cap B_j \cap B_k)$$

(terms with the intersection of 4 or more B_i's are not needed since Alice must have classes on at least 2 days). We have

$$P(B_1) = \frac{\binom{24}{7}}{\binom{30}{7}}, P(B_1 \cap B_2) = \frac{\binom{18}{7}}{\binom{30}{7}}, P(B_1 \cap B_2 \cap B_3) = \frac{\binom{12}{7}}{\binom{30}{7}}$$

and similarly for the other intersections. So

$$P(B_1 \cup \cdots \cup B_5) = 5\frac{\binom{24}{7}}{\binom{30}{7}} - \binom{5}{2}\frac{\binom{18}{7}}{\binom{30}{7}} + \binom{5}{3}\frac{\binom{12}{7}}{\binom{30}{7}} = \frac{263}{377}.$$

Therefore,

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \frac{114}{377} \approx 0.302.$$

7. **Is it possible that an event is independent of itself? If so, when?**

Let A be an event. If A is independent of itself, then P(A) = P(A∩A) = P(A) 2 , so P(A) is 0 or 1. So this is only possible in the extreme cases that the event has probability 0 or 1.

8. **Is it always true that if A and B are independent events, then Ac and Bc are independent events? Show that it is, or give a counterexample.**

Yes, because we have

$$P(A^c \cap B^c) = 1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A \cap B));$$

since A and B are independent, this becomes

$$1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P(A^c)P(B^c).$$

9. **Give an example of 3 events A, B, C which are pairwise independent but not independent. Hint: find an example where whether C occurs is completely determined if we know whether A occurred and whether B occurred, but completely undetermined if we know only one of these things.**

Consider two fair, independent coin tosses, and let A be the event that the first toss is Heads, B be the event that the second toss is Heads, and C be the event that the two tosses have the same result. Then A, B, C are dependent since $P(A \cap B \cap C) = P(A \cap B) = P(A)P(B) = 1/4 \neq 1/8 = P(A)P(B)P(C)$, but they are pairwise independent: A and B are independent by definition; A and C are independent since $P(A \cap C) = P(A \cap B) = 1/4 = P(A)P(C)$, and similarly B and C are independent.

10. A bag contains one marble which is either green or blue, with equal probabilities. A green marble is put in the bag (so there are 2 marbles now), and then a random marble is taken out. The marble taken out is green. What is the probability that the remaining marble is also green?

Let A be the event that the initial marble is green, B be the event that the removed marble is green, and C be the event that the remaining marble is green.

We need to find P(C|B). There are several ways to find this; one natural way is to condition on whether the initial marble is green:

$P(C|B) = P(C|B, A)P(A|B) + P(C|B, A^c)P(A^c|B) = 1P(A|B) + 0P(A^c|B)$.

To find P(A|B), use Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{1/2}{P(B|A)P(A) + P(B|Ac)P(Ac)} = \frac{1/2}{1/2 + 1/4} = 2/3.$$

So $P(C|B) = 2/3$.

11. A group of n 2 people decide to play an exciting game of Rock-Paper Scissors. As you may recall, Rock smashes Scissors, Scissors cuts Paper, and Paper covers Rock (despite Bart Simpson saying "Good old rock, nothing beats that!"). Usually, this game is played with 2 players, but it can be extended to more players as follows. If exactly 2 of the 3 choices appear when everyone reveals their choice, say a, b 2 {Rock, P aper, Scissors} where a beats b, the game is decisive: the players who chose a win, and the players who chose b lose. Otherwise, the game is indecisive and the players play again. For example, with 5 players, if one player picks Rock, two pick Scissors, and two pick Paper, the round is indecisive and they play again. But if 3 pick Rock and 2 pick Scissors, then the Rock players win and the Scissors players lose the game. 1 Assume that the n players independently and randomly choose between Rock, Scissors, and Paper, with equal probabilities. Let X, Y, Z be the number of players who pick Rock, Scissors, Paper, respectively in one game.
(a) Find the joint PMF of X, Y, Z.
(b) Find the probability that the game is decisive. Simplify your answer (it should not involve a sum of many terms).
(c) What is the probability that the game is decisive for n = 5? What is the limiting probability that a game is decisive as n ! 1? Explain briefly why your answer makes sense.

12. A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase "free money" is used, whereas this phrase is only used in 1% of non-spam emails. A new email has just arrived, which does mention "free money". What is the probability that it is spam?

Let S be the event that an email is spam and F be the event that an email has the "free money" phrase.

By Bayes' rule, $P(S|F) = \frac{P(F|S)P(S)}{P(F)} = \frac{0.1 * 0.8}{0.1 * 0.8 + 0.01 * 0.2} = \frac{80/1000}{82/1000} = 80/82 \approx 0.9756.$

13. A crime is committed by one of two suspects, A and B. Initially, there is equal evidence against both of them. In further investigation at the crime scene, it is found that the guilty party had a blood type found in 10% of the population. Suspect A does match this blood type, whereas the blood type of Suspect B is unknown.

(a) Given this new information, what is the probability that A is the guilty party?

Let M be the event that A's blood type matches the guilty party's and for brevity, write A for "A is guilty" and B for "B is guilty".

By Bayes' Rule, $P(A|M) = \frac{P(M|A)P(A)}{P(M|A)P(A) + P(M|B)P(B)} = \frac{1/2}{1/2 + (\frac{1}{10})(\frac{1}{2})} = 10 / 11$.

(We have P(M|B) = 1/10 since, given that B is guilty, the probability that A's blood type matches the guilty party's is the same probability as for the general population.)

b) Given this new information, what is the probability that B's blood type matches that found at the crime scene?

Let C be the event that B's blood type matches, and condition on whether B is guilty.

This gives $P(C|M) = P(C|M, A)P(A|M) + P(C|M, B)P(B|M) = \frac{1}{10} \cdot \frac{10}{11} + \frac{1}{11} = 2 / 11$.

14. You are going to play 2 games of chess with an opponent whom you have never played against before (for the sake of this problem). Your opponent is equally likely to be a beginner, intermediate, or a master. Depending on

(a) What is your probability of winning the first game?

a) Let Wi be the event of winning the ith game. By the law of total probability, P(W1) = (0.9 + 0.5 + 0.3)/3 = 17/30.

(b) Congratulations: you won the first game! Given this information, what is the probability that you will also win the second game

a) We have P(W2|W1) = P(W2, W1)/P(W1). The denominator is known from (a), while the numerator can be found by conditioning on the skill level of the opponent:

$P(W1, W2) = \frac{1}{3} P(W1, W2|\text{beginner}) + \frac{1}{3} P(W1, W2|\text{intermediate}) + \frac{1}{3} P(W1, W2|\text{expert})$.

Since W1 and W2 are conditionally independent given the skill level of the opponent, this becomes

$P(W1, W2) = (0.9^2 + 0.5^2 + 0.3^2)/3 = 23/60$.

So $P(W2|W1) = \frac{23/60}{17/30} = 23/34$.

(c) Explain the distinction between assuming that the outcomes of the games are independent and assuming that they are conditionally independent given the opponent's skill level. Which of these assumptions seems more reasonable, and why?

Independence here means that knowing one game's outcome gives no information about the other game's outcome, while conditional independence is the same statement where all probabilities are conditional on the opponent's skill level. Conditional independence given the opponent's skill level is a more reasonable assumption here. This is because winning the first game gives information about the opponent's skill level, which in turn gives information about the result of the second game. That is, if the opponent's skill level is treated as fixed and known, then it may be reasonable to assume independence of games given this information; with the opponent's skill level random, earlier games can be used to help infer the opponent's skill level, which affects the probabilities for future games.

15. A chicken lays n eggs. Each egg independently does or doesn't hatch, with probability p of hatching.

For each egg that hatches, the chick does or doesn't survive (independently of the other eggs), with probability s of survival. Let N ⇠Bin(n, p) be the number of eggs which hatch, X be the number of chicks which survive, and Y be the number of chicks which hatch but don't survive (so X + Y = N). Find the marginal PMF of X, and the joint PMF of X and Y . Are they independent?

Marginally we have $X \sim \text{Bin}(n, ps)$, as shown on a previous homework problem using a story proof (the eggs can be thought of as independent Bernoulli trials with probability ps of success for each). Here X and Y are *not* independent, unlike in the chicken-egg problem from class (where N was Poisson). This follows immediately from thinking about an *extreme case*: if $X = n$, then clearly $Y = 0$. So they are not independent: $P(Y = 0) < 1$, while $P(Y = 0|X = n) = 1$.

To find the joint distribution, condition on N and note that only the $N = i+j$ term is nonzero: for any nonnegative integers i, j with $i + j \leq n$,

$$\begin{aligned} P(X = i, Y = j) &= P(X = i, Y = j|N = i + j)P(N = i + j) \\ &= P(X = i|N = i + j)P(N = i + j) \\ &= \binom{i+j}{i} s^i (1-s)^j \binom{n}{i+j} p^{i+j}(1-p)^{n-i-j} \\ &= \frac{n!}{i!j!(n-i-j)!}(ps)^i(p(1-s))^j(1-p)^{n-i-j}. \end{aligned}$$

If we let Z be the number of eggs which don't hatch, then from the above we have that (X, Y, Z) has a Multinomial$(n, (ps, p(1-s), (1-p))$ distribution, which makes sense intuitively since each egg independently falls into 1 of 3 categories: hatch-and-survive, hatch-and-don't-survive, and don't-hatch, with probabilities $ps, p(1-s), 1-p$ respectively.

Additional resources

Wednesday, January 25, 2023 2:49 PM

Problem/Coding based questions :

1. https://www.nicksingh.com/posts/40-probability-statistics-data-science-interview-questions-asked-by-fang-wall-street
2. Interview query 1-2 stats and probability questions : https://www.interviewquery.com/questions?searchQuery=&searchQuestionTag=&searchCompany=&ordering=Recommended&pageSize=100&page=0&tags=Statistics&tags=Probability
3. https://ravivats.github.io/2020-09-06-headstart-ml-p02-stats-interview-questions/
4. 20 probability questions : https://github.com/kojino/120-Data-Science-Interview-Questions/blob/master/probability.md

Theory based questions :

1. https://www.projectpro.io/article/statistic-and-probability-interview-questions-for-data-science/483
2. https://intellipaat.com/blog/interview-question/statistics-interview-questions/
3. https://github.com/kojino/120-Data-Science-Interview-Questions/blob/master/statistical-inference.md
4. https://www.mygreatlearning.com/blog/statistics-interview-questions/
5. https://www.ctanujit.org/uploads/2/5/3/9/25393293/data_science_interview_questions.pdf

Additional Topics:

1. AB testing
2. How to select null and alternate hypothesis ?

Interview questions

Monday, October 16, 2023 12:15 PM

What is type 1 error and type 2 error

- Type 1 Errors (or false positives) occur when we reject a hypothesis when it is actually true.
- Type 2 Errors (or false negatives) occur when we fail to reject a hypothesis when it is actually false.