Day 7

26 May 2024 22:39

Bayes theorem

Inferential techniques

- t-test
  - Df

1. Know the problem
2. Choose the right type of stat test
3. Very sure about the alpha (sig level)
   a. 5%, 1% .5% ….
4. Skewed , normally ..
5. Continuous or categorical
6. Which library , scipy …how to use
7. Return
   a. Test statistic
   b. p-value
8. Plotting
   a. Get the right distribution plotted
   b. Mark the Critical values, t_statistic
   c. Derive p-value
9. concluding

Degrees of freedom

degrees of freedom (df) represent the number of independent values or quantities in a sample that are free to vary

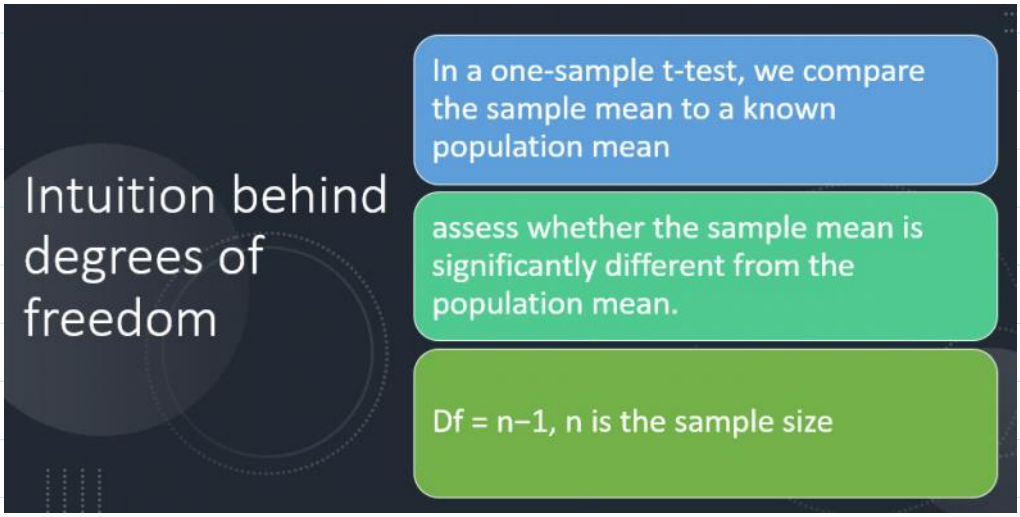
In a t-test, the degrees of freedom are calculated as the total number of observations minus the number of parameters estimated from the data. For an independent samples t-test, the degrees of freedom are equal to the sum of the sample sizes of the two groups minus 2.

The degrees of freedom (df) for an independent samples t-test is calculated using the formula:

$df = n1 + n2 - 2$

$df = 5 + 5 - 2$

$df = 8$



In a 30-seat classroom scenario, the first 29 individuals can choose any seat, but the 30th person has only one seat remaining.

Similarly, when calculating the mean of a sample of 30 numbers, the first 29 numbers can vary, but the 30th number is determined to achieve the desired sample mean.

when estimating the mean of a single population, the degrees of freedom is 29, representing the number of values that can vary in the sample.

why the t-statistic is in terms of standard deviation units

Calculation:

- is calculated by dividing the difference between the sample mean and the hypothesized population mean by the standard error of the mean.

$$t = \frac{\text{sample mean} - \text{population mean}}{\text{standard error of the mean}}$$

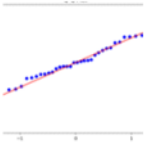
Interpretation:

- Expressing the t-statistic in terms of standard deviation units provides a meaningful interpretation.
- For example, a t-statistic of 2.0 indicates that the sample mean is 2 standard deviations away from the hypothesized population mean, suggesting that the observed difference is relatively large and unlikely to occur by random chance alone.

Assumptions for t-test

Normality:	The data should follow a normal distribution. However, this assumption is robust to violations, especially for larger sample sizes (typically n > 30).
Random Sampling:	The data points should be randomly and independently sampled from the population.
Scale of Measurement:	The dependent variable should be measured on an interval or ratio scale.
Homogeneity of Variance (Equal Variance):	For independent samples t-tests, the variances of the two groups being compared should be approximately equal.

Check the normality of the data



Stat: 254

Visual Inspection Using Q-Q Plots:

Q-Q plots (Quantile-Quantile plots) allow you to visually assess if the data follows a normal distribution.

Points should approximately lie along a straight line if the data is normally distributed.

Statistical Tests:

Shapiro-Wilk test is commonly used for checking normality.

tests the null hypothesis that the data was drawn from a normal distribution.

Verify the assumption of equal variances

Levene's Test for Equality of Variances	Levene's test is a commonly used statistical test for assessing whether the variances of two or more groups are equal. The null hypothesis of Levene's test is that the variances are equal across all groups.
Interpretation of Results:	**Levene's Test Statistic**: The test statistic measures the discrepancy between the observed and expected variances. A larger test statistic suggests greater variability between groups. **Levene's Test p-value**: If the p-value is greater than the chosen significance level (e.g., 0.05), you fail to reject the null hypothesis, indicating that the variances are approximately equal across groups. Conversely, if the p-value is less than the significance level, you reject the null hypothesis, suggesting that the variances are not equal.

Why check for assumptions?

helps ensure the validity of parametric tests

While normality checks are important, it's also essential to consider the context of the data and the specific requirements of the statistical analysis being conducted.

non-parametric alternative to the t-test

Mann-Whitney U Test (Wilcoxon Rank-Sum Test)

Purpose: Compares differences between two independent groups.

When to Use:

- Data are ordinal, interval, or ratio scale.
- When comparing the medians of two independent groups, especially when the data do not meet the assumptions of the t-test (e.g., non-normal distribution).

Mann-Whitney U test

Combine and Rank

Combine the data from both groups and assign ranks to the combined data.

- If there are ties (i.e., identical values), assign to each tied value the average of the ranks they would have received.

Sum

- Calculate the sum of the ranks for each group separately.
- Let's denote these sums as $R1$ and $R2$.

Compute

- Use the rank sums to compute the U statistic for each group.

formulas for $U1$ and $U2$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where:

- n_1 is the number of observations in group 1.
- n_2 is the number of observations in group 2.
- R_1 is the sum of ranks for group 1.
- R_2 is the sum of ranks for group 2.

The smaller value of U_1 and U_2 is used as the test statistic U.

Example

- Consider two groups with the following scores:
  - Group 1: [85, 86, 88, 75, 78]
  - Group 2: [92, 94, 89, 95, 90]
- **Combine and Rank**
  - Combined data: [75, 78, 85, 86, 88, 89, 90, 92, 94, 95]
  - Ranks: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
  - Group 1 ranks: [3, 4, 5, 1, 2] -> Sum $R1$ = 3+4+5+1+2=15
  - Group 2 ranks: [8, 9, 6, 10, 7] -> Sum $R2$= 8+9+6+10+7=40

Calculate U for Each Group

- **Determine the Smaller U**
- The smaller of the two U values is
- U=0.

Using the formulas for U_1 and U_2:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
$$U_1 = 5 \cdot 5 + \frac{5(5+1)}{2} - 15$$
$$U_1 = 25 + 15 - 15 = 25$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$
$$U_2 = 5 \cdot 5 + \frac{5(5+1)}{2} - 40$$
$$U_2 = 25 + 15 - 40 = 0$$

U statistic

U statistic can never be negative.

01

U statistic is a measure of the rank differences between two groups and is always a non-negative integer.

02

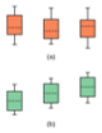
U statistic ranges from a minimum value of 0 to a maximum value of $n1 \times n2$,

- $n1$ and $n2$ are the sample sizes of the two groups being compared.

ANOVA - Analysis of variance

- Extension of t-test
- When we have more than 2 groups

One-way vs 2-way - Number of Factors



One-Way ANOVA:

one categorical independent variable (factor) and multiple levels or groups within that variable.

For example, you might have different treatment groups for a single factor like "Diet Type" with levels "A," "B," and "C."

or A	Factor B	
	B_{low}	B
	25	
	45	

Two-Way ANOVA:

two categorical independent variables (factors). It's used when you want to examine the effects of two factors simultaneously. Each factor has multiple levels, resulting in various combinations of factor levels.

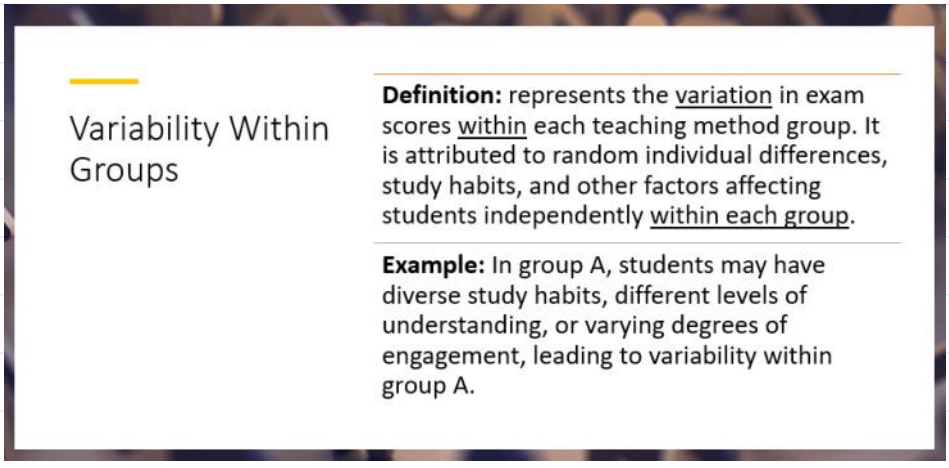
For example, you might be investigating the effects of both "Diet Type" and "Exercise Level" on weight loss.

Variability Components

- ANOVA decomposes the total variability in the data into two components:
  - variability within groups (due to random individual differences)
  - variability between groups (due to the effect of different treatments or categories).

example

- **Exam Scores in Different Teaching Methods**
  - Imagine a study comparing the effectiveness of three different teaching methods (A, B, and C) on students' exam scores.
  - exam scores are collected from three separate groups of students, each exposed to one of the teaching methods.
  - goal is to understand whether there are significant differences in exam scores between the teaching methods.



variability within groups

- is calculated by assessing the spread or dispersion of individual observations within each group.
- measures how much individual data points deviate from the mean of their respective groups.

$$\text{Variability Within Groups} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

where:

- k is the number of groups.
- n_i is the number of observations in the i^{th} group.
- X_{ij} is the j^{th} observation in the i^{th} group.
- $\bar{X}_i$ is the mean of the i^{th} group.

Calculate the Mean for Each Group ($\bar{X}_i$):

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

Calculate the Difference of Each Observation from its Group Mean:

$$X_{ij} - \bar{X}_i$$

Square the Differences for Each Observation:

$$(X_{ij} - \bar{X}_i)^2$$

Sum Up the Squared Differences Within Each Group:

$$\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

Sum Up the Squared Differences Across All Groups:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

Worked out example

Group A: [80, 85, 88, 90, 92]
Group B: [75, 78, 80, 82, 85]
Group C: [85, 88, 90, 92, 95]

Step 1: Calculate Group Means ($\bar{X}_i$):

$$\bar{X}_A = \frac{80+85+88+90+92}{5} = 87$$
$$\bar{X}_B = \frac{75+78+80+82+85}{5} = 80$$
$$\bar{X}_C = \frac{85+88+90+92+95}{5} = 90$$

Step 2: Calculate Overall Mean ($\bar{X}$):

$$\bar{X} = \frac{87+80+90}{3} = 85.67$$

...next steps

Step 3: Calculate Variability Within Groups:

$$\text{Within Group A} = (80-87)^2 + (85-87)^2 + (88-87)^2 + (90-87)^2 + (92-87)^2$$
$$\text{Within Group B} = (75-80)^2 + (78-80)^2 + (80-80)^2 + (82-80)^2 + (85-80)^2$$
$$\text{Within Group C} = (85-90)^2 + (88-90)^2 + (90-90)^2 + (92-90)^2 + (95-90)^2$$

Step 4: Sum Up Variability Within Groups:

$$\text{Within Groups} = \text{Within Group A} + \text{Within Group B} + \text{Within Group C}$$

intuition behind variability within groups

- measures the spread or dispersion of individual data points within each group.
- Intuitively, it reflects the degree of diversity or heterogeneity among the observations within a specific treatment, condition, or category.

Variability between groups

- calculated by comparing the differences in means of the groups.
- The goal is to determine whether these mean differences are statistically significant and not likely to occur by random chance.

$$\text{Variability Between Groups} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

where:

- k is the number of groups.
- n_i is the number of observations in the i^{th} group.
- $\bar{X}_i$ is the mean of the i^{th} group.
- $\bar{X}$ is the overall mean of all observations.

Calculate the Overall Mean ($\bar{X}$):

$$\bar{X} = \frac{\sum_{i=1}^{k} n_i \bar{X}_i}{N}$$

where N is the total number of observations across all groups.

Calculate the Differences from the Overall Mean for Each Group:

$$\bar{X}_i - \bar{X}$$

Square the Differences:

$$(\bar{X}_i - \bar{X})^2$$

Weight the Squared Differences by the Number of Observations in Each Group:

$$n_i(\bar{X}_i - \bar{X})^2$$

Sum Up the Weighted Squared Differences Across All Groups:

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

Worked out example

Given Data:

- **Group A:** [80, 85, 88, 90, 92]
- **Group B:** [75, 78, 80, 82, 85]
- **Group C:** [85, 88, 90, 92, 95]

Step 1: Calculate Group Means ($\bar{X}_i$):

$$\bar{X}_A = 87$$
$$\bar{X}_B = 80$$
$$\bar{X}_C = 90$$

Step 2: Calculate Overall Mean ($\bar{X}$):

$$\bar{X} = 85.67$$

Variability Between Groups:

$$\text{Between Groups} = 5 \cdot (87 - 85.67)^2 + 5 \cdot (80 - 85.67)^2 + 5 \cdot (90 - 85.67)^2$$

Detailed Calculation:

$$\text{Between Groups} = 5 \cdot (1.33)^2 + 5 \cdot (-5.67)^2 + 5 \cdot (4.33)^2$$
$$\text{Between Groups} = 5 \cdot 1.7689 + 5 \cdot 32.2689 + 5 \cdot 18.7489$$
$$\text{Between Groups} = 8.8445 + 161.3445 + 93.7445$$
$$\text{Between Groups} = 263.9335$$

So, the variability between groups is 263.9335.

intuition behind variability between groups

- The variability between groups in an ANOVA reflects the extent to which the means of different groups differ from each other.
- Intuitively, it measures the influence or impact of the factor (or treatment) being studied on the response variable.

F-value

01

test statistic for an ANOVA test, always starts from 0.

02

because the F-distribution, which is used to determine the critical value and p-value for the F-test, is defined only for non-negative values.

03

F-distribution is positively skewed and starts from 0, extending to positive infinity.

Assumptions of ANOVA

Homogeneity of Variances:

This assumption states that the variances (or standard deviations) of the dependent variable within each group are approximately equal.

Normality:

population distribution should be approximately normal. However, as mentioned earlier, ANOVA is often robust to deviations from normality, especially with larger sample sizes.

Independence:

The independence assumption states that observations within each group are independent of each other.

[No Title]

18

01-06-2024

Chi2

- **Scenario:** You want to investigate if there is a significant association between two categorical variables.
- **Example:** You have survey data from a group of people, and you want to know if there is a relationship between gender (Male/Female) and the preference for a particular product category (e.g., Books, Electronics).

Use case

- **Educational Research - Exam Performance:**
  - **Scenario:** In educational research, the Chi-square test can be used to explore the association between student performance categories and study habits.
  - **Example:** You want to examine if there is a significant relationship between study hours per week (low, moderate, high) and exam performance categories (Fail, Pass, Distinction).
  -
- **Market Research - Product Preferences:**
  - **Scenario:** In market research, the Chi-square test can be used to analyze the relationship between demographic factors and product preferences.
  - **Example:** You want to investigate if there is a significant relationship between age groups (e.g., 18-24, 25-34, 35-44) and the preferred streaming service (e.g., Netflix, Hulu, Amazon Prime).

Use case

4

01-06-2024

Worked out example

- Suppose we want to investigate whether there is a significant association between
  - the type of mobile phone operating system people use (iOS, Android, Other) and
  - their preference for mobile apps (Games, Social Media, Utilities).

	Games	Social Media	Utilities	Total
iOS	30	20	10	60
Android	15	25	20	60
Other	10	15	5	30
Total	55	60	35	150

Formulate Hypotheses

- *H*0: There is no association between the type of mobile operating system and preference for mobile apps.
- *H*1: There is a significant association between the type of mobile operating system and preference for mobile apps.

Calculate Expected Frequencies

Observed frequency

	Games	Social Media	Utilities	Total
iOS	30	20	10	60
Android	15	25	20	60
Other	10	15	5	30
Total	55	60	35	150

$$: \frac{(row\ total \times column\ total)}{grand\ total}$$

Expected frequency

	Games	Social Media	Utilities
iOS	22.0	24.0	14.0
Android	22.0	24.0	14.0
Other	11.0	12.0	7.0

Intuition Behind the Expected Frequency Formula

Proportional Allocation: If the two variables are independent, the frequency of occurrences in any particular cell should be proportional to the product of the probabilities of the corresponding row and column.

expected frequency E_{ij} for cell (i,j) (where i represents a category of A and j represents a category of B) is calculated using the formula

$$E_{ij} = \frac{(R_i \times C_j)}{N}$$

	Games	Social Media	Utilities	Total
iOS	30	20	10	60
Android	15	25	20	60
Other	10	15	5	30
Total	55	60	35	150

$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

Calculate this value for each cell and sum them up:

$\chi^2 = \frac{(30-22)^2}{22} + \frac{(20-24)^2}{24} + \frac{(10-14)^2}{14} + \ldots + \frac{(5-7)^2}{7}$

The calculated χ^2 value is then compared to the critical value from the Chi-square distribution table or used to calculate the p-value.

	Games	Social Media	Utilities
iOS	22.0	24.0	14.0
Android	22.0	24.0	14.0
Other	11.0	12.0	7.0

analysis

Determine Degrees of Freedom:

Degrees of freedom
(*rows*−1)×(*columns*−1)=(3−1)×(3−1)=4.

Consult Chi-square Table or Use Software:

At a significance level of 0.05 and 4 degrees of freedom, the critical value is approximately 9.49.

Make a Decision:

If the p-value is less than 0.05, reject the null hypothesis.

Interpret Results:

If the null hypothesis is rejected, conclude that there is evidence of an association between the type of mobile operating system and preference for mobile apps.

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$

where $x \geq 0$ and $k > 0$ is the degrees of freedom parameter.

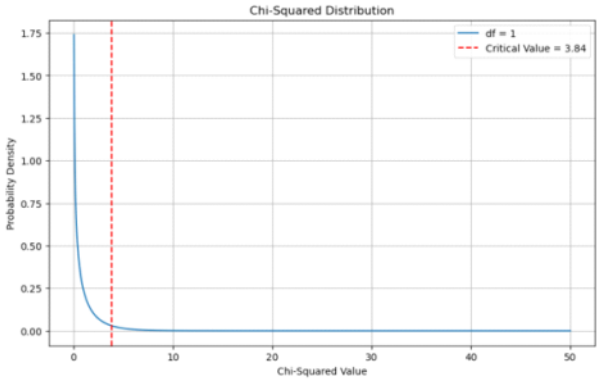
Probability Density Function (PDF)

Properties

Non-Negative	chi-squared distribution is defined for non-negative values.
Shape	shape of the chi-squared distribution depends on the degrees of freedom. skewed to the right, especially for lower degrees of freedom. As the degrees of freedom increase, the distribution becomes more symmetrical and approaches a normal distribution.
Mean and Variance	mean is k. variance is $2k$.

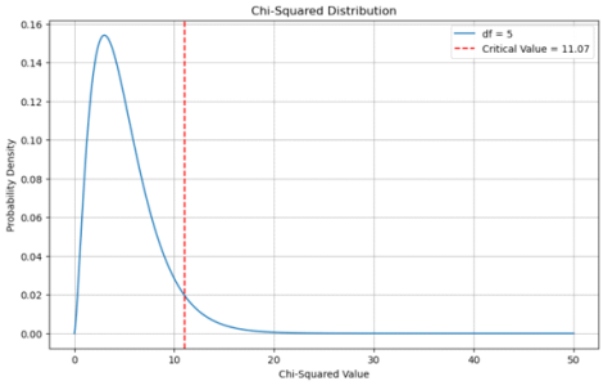
df=1

- distribution is highly skewed to the right.
- Mean = 1, Variance = 2.
- Critical value at 0.05 significance level ≈ 3.84.



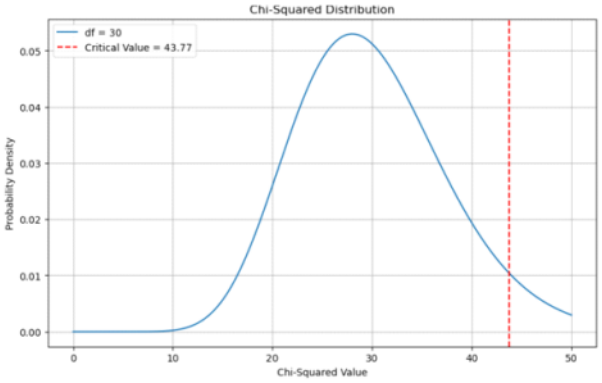
Df=5

- Less skewed, starts to look more symmetric.
- Mean = 5, Variance = 10.
- Critical value at 0.05 significance level ≈ 11.07.



Df=30

- Approximates a normal distribution.
- Mean = 30, Variance = 60.
- Critical value at 0.05 significance level ≈ 43.77.



C:\PS-DS\MathBasics\.venv\Scripts\python.exe C:\PS-DS\MathBasics\Maths\Mean.py [83, 67, 85, 74, 62, 82]
Mean value is 75.5 Variance value is 74.25
********** Standard deviation from math pack is 8.616843969807043
********** Standard deviation from numpy pack is 8.616843969807043
********** Standard deviation from stat pack is 9.439279633531363

From <https://engagex.simplilearn.com/#/zoom-meeting?meetingNumber=OTYxMjcwMDc4ODI%

3D&email=Ymh1cGVuZHJhc2luaGFAaG90bWFpbC5jb20%3D&passWord=dzZiNHUwNmY%
3D&username=Qmh1cGVuZHJhIFNpbmhh&isTrainer=true&registrantId=6qaSURzwQVirn2NoKL97Eg,
&registrantToken=g7FPuoYXPfTkv0HFDkzyDvBRYqtuYs6orkSkHuPr03s.DQYAAAAWYZ28ihY2cWFTVVJ6d1FWaXJuMk5vS0w5N0
VnAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA>

import math import numpy as np import statistics def calcaultedeviation(_meanvalue,datas): print(datas) variance = sum((x-_meanvalue)**2 for x in datas)/len(datas) print('Mean value is ', meanvalue) print('Variance value is',variance) print('**********') print('Standard deviation from math pack is',math.sqrt(variance)) print('**********') print('Standard deviation from numpy pack is',np.sqrt(variance)) print('**********') print('Standard deviation from stat pack is',statistics.stdev(datas)) basenumber = 0.85 # Calculate the power of 10 result = base_number ** 10 #print("2 raised to the power of 10 is:", result) def calulatemean(datas): mean= sum(datas)/len(datas) calcaultedeviation(mean,datas) dealer1 =[83,67,85,74,62,82] dealer2 =[83,85,82,75,69,69] dealer3 =[85,66,77,84,69,75] dealer4 =[73,91,82,85,76,83] dealer5 =[81,82,76,74,70,61] def calculatedetails(): datalsts=[dealer1] for x in datalsts: calulatemean(x) calculatedetails()

From <https://engagex.simplilearn.com/#/zoom-meeting?meetingNumber=OTYxMjcwMDc4ODI%3D&email=Ymh1cGVuZHJhc2luaGFAaG90bWFpbC5jb20%
3D&passWord=dzZiNHUwNmY%3D&username=Qmh1cGVuZHJhIFNpbmhh&isTrainer=true&registrantId=6qaSURzwQVirn2NoKL97Eg,
&registrantToken=g7FPuoYXPfTkv0HFDkzyDvBRYqtuYs6orkSkHuPr03s.DQYAAAAWYZ28ihY2cWFTVVJ6d1FWaXJuMk5vS0w5N0VnAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA>

How to visualize notebook file in the github?