Submitted by: **Mohammad Wasiq**
Email: **gl0427@myamu.ac.in**
Placement Statistics Assignment 1

STATISTICS ASSIGNMENT

Q-1. A university wants to understand the relationship between the SAT scores of its applicants and their college GPA. They collect data on 500 students, including their SAT scores (out of 1600) and their college GPA (on a 4.0 scale). They find that the correlation coefficient between SAT scores and college GPA is 0.7. What does this correlation coefficient indicate about the relationship between SAT scores and college GPA?

Solution:-
A correlation coefficient of 0.7 between SAT scores and college GPA indicates a strong positive relationship between the two variables.

In this case, the positive sign indicates that higher SAT scores are associated with higher college GPAs, and vice versa. The magnitude of 0.7 indicates a relatively strong correlation, meaning that the SAT scores explain a considerable portion of the variability in college GPAs.

However, it is important to note that correlation does not imply causation. While the correlation coefficient suggests a relationship between SAT scores and college GPA, it does not indicate that one variable causes the other. Other factors, such as study habits, motivation, and external influences, may also contribute to college GPA.

Q-2. Consider a dataset containing the heights (in centimeters) of 1000 individuals. The mean height is 170 cm with a standard deviation of 10 cm. The dataset is approximately normally distributed, and its skewness is approximately zero. Based on this information, answer the following questions: a. What percentage of individuals in the dataset have heights between 160 cm and 180 cm? b. If we randomly select 100 individuals from the dataset, what is the probability that their average height is greater than 175 cm? c. Assuming the dataset follows a normal distribution, what is the z-score corresponding to a height of 185 cm? d. We know that 5% of the dataset has heights below a certain value. What is the approximate height corresponding to this threshold? e. Calculate the coefficient of variation (CV) for the dataset. f. Calculate the skewness of the dataset and interpret the result.

Solution:-

a. To find the percentage of individuals with heights between 160 cm and 180 cm, we need to calculate the area under the normal distribution curve between these two values. Since the dataset is approximately normally distributed, we can use the Z-table or a statistical software to find this area.

First, we convert the values to Z-scores using the formula: $Z = (X - \mu) / \sigma$, where X is the height, μ is the mean, and σ is the standard deviation.

For 160 cm: $Z_1 = (160 - 170) / 10 = -1$
For 180 cm: $Z_2 = (180 - 170) / 10 = 1$

Next, we look up the corresponding values in the Z-table or use a statistical software to find the area between these Z-scores. This area represents the percentage of individuals with heights between 160 cm and 180 cm.

Let's assume the area is A.

Therefore, the percentage of individuals in the dataset with heights between 160 cm and 180 cm is A * 100%.

b. The average height of 100 randomly selected individuals follows the same normal distribution as the original dataset, but with a smaller standard deviation. In this case, the standard deviation becomes σ / sqrt(n), where n is the sample size.

So, the standard deviation of the average height for a sample of 100 individuals is 10 / sqrt(100) = 1 cm.

To find the probability that the average height is greater than 175 cm, we need to calculate the Z-score for 175 cm using the formula: $Z = (X - \mu) / (\sigma / \text{sqrt}(n))$, where X is the value (175 cm), μ is the mean (170 cm), σ is the standard deviation (10 cm), and n is the sample size (100).

$Z = (175 - 170) / (1) = 5$

We can then look up the probability corresponding to this Z-score in the Z-table or use a statistical software.

Let's assume the probability is P.

Therefore, the probability that the average height of a random sample of 100 individuals is greater than 175 cm is P.

c. To find the Z-score corresponding to a height of 185 cm, we use the formula: $Z = (X - \mu) / \sigma$, where X is the height (185 cm), μ is the mean (170 cm), and σ is the standard deviation (10 cm).

$Z = (185 - 170) / 10 = 1.5$

Therefore, the Z-score corresponding to a height of 185 cm is 1.5.

d. To find the approximate height corresponding to the threshold where 5% of the dataset has heights below that value, we need to find the Z-score corresponding to the cumulative probability of 0.05.

Let's assume the Z-score is Z.

Using the Z-table or a statistical software, we can find the Z-score corresponding to a cumulative probability of 0.05. Then, we can use the formula: $X = Z * \sigma + \mu$, where X is the height, Z is the Z-score, σ is the standard deviation (10 cm), and μ is the mean (170 cm).

Therefore, the approximate height corresponding to the threshold where 5% of the dataset has heights below that value is X cm.

e. The coefficient of variation (CV) is a measure of relative variability and is calculated as the ratio of the standard deviation to the mean, multiplied by 100% to express it as a percentage.

$CV = (\sigma / \mu) * 100\%$

In this case, $CV = (10 / 170) * 100\%$.

Therefore, the coefficient of variation for the dataset is

CV%.

Since the dataset's skewness is approximately zero, it indicates that the data is symmetrically distributed. A skewness of zero means that the distribution is perfectly symmetrical, with the left and right tails having the same shape and length. The mean, median, and mode are all equal in a symmetrical distribution.

Interpreting this result, we can say that the heights in the dataset are evenly distributed around the mean, without a significant bias towards either higher or lower values.

Q-3. Consider the 'Blood Pressure Before' and 'Blood Pressure After' columns from the data and calculate the following
https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_
a. Measure the dispersion in both and interpret the results.
b. Calculate mean and 5% confidence interval and plot it in a graph
c. Calculate the Mean absolute deviation and Standard deviation and interpret the results.
d. Calculate the correlation coefficient and check the significance of it at 1% level of significance.

Solution:-

a. To measure the dispersion in both 'Blood Pressure Before' and 'Blood Pressure After,' we can calculate the range and interquartile range (IQR). The range is the difference between the maximum and minimum values, while the IQR represents the range between the first quartile (Q1) and the third quartile (Q3).

For 'Blood Pressure Before':

Range = Maximum value - Minimum value
= 148 - 120
Range = 28 mmHg

IQR = Q_3 - Q_1
To find Q_1 and Q_3:
1. Sort the 'Blood Pressure Before' values in ascending order:
120, 120, 118, 118, 119, 121, 121, 122, 123, 123, 124, 124, 125, 125, 127, 127, 127, 128, 128, 128, 129, 129, 130, 130, 130, 131, 131, 132, 132, 132, 135, 135, 135, 136, 136, 136, 137, 137, 139, 139, 140, 140, 142, 142, 143, 143, 145, 145, 145, 148
2. Calculate Q1 (first quartile):
Q1 = 123
3. Calculate Q3 (third quartile):
Q3 = 136
IQR = Q3 - Q1
= 136 - 123
IQR = 13 mmHg

For 'Blood Pressure After':

Range = Maximum value - Minimum value
= 141 - 118
Range = 23 mmHg

IQR = Q3 - Q1

To find Q1 and Q3:

1. Sort the 'Blood Pressure After' values in ascending order:
   118, 118, 119, 121, 122, 123, 124, 124, 125, 125, 127, 127, 128, 128, 129, 129, 130, 130, 131, 132, 132, 135, 135, 136, 136, 137, 139, 139, 140, 141
2. Calculate Q1 (first quartile):
   Q1 = 124
3. Calculate Q3 (third quartile):
   Q3 = 136

IQR = Q3 - Q1
= 136 - 124
IQR = 12 mmHg

Interpretation:

The dispersion in 'Blood Pressure Before' is higher than that in 'Blood Pressure After.' This means that the values in 'Blood Pressure Before' are more spread out or varied compared to the values in 'Blood Pressure After.'

b. To calculate the mean and 5% confidence interval, we need to find the average of the 'Blood Pressure Before' and 'Blood Pressure After' values.

Mean of 'Blood Pressure Before':

(130 + 142 + 120 + 135 + 148 + 122 + 137 + 130 + 142 + 128 + 135 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 +128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143) / 100

Mean of 'Blood Pressure Before'= 132.62 mmHg

Mean of 'Blood Pressure After':

(120 + 135 + 118 + 127 + 140 + 118 + 129 + 124 + 137 + 125 + 129 + 132 + 125 + 136 + 118 + 122 + 130 + 139 + 123 + 132 + 131 + 126 + 120 + 123 + 139 + 122 + 129 + 136 + 131 + 127 + 140 + 119 + 121 + 129 + 137 + 122 + 135 + 131 + 124 + 119 + 124 + 139 + 123 + 131 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 +

118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137) / 100

Mean of 'Blood Pressure After' = 128.14 mmHg

Confidence Interval:
To calculate the 5% confidence interval, we can use the formula:

Confidence Interval = Mean ± (Critical value * Standard error)

For a 5% confidence level, the critical value is approximately 1.96 (assuming a large sample size).

Standard error = Standard deviation / $\sqrt{n}$

Standard deviation of 'Blood Pressure Before' (σ_{before}) = $\sqrt{[\Sigma(x_i - \mu_{before})^2 / n]}$

Standard deviation of 'Blood Pressure After' (σ_{after}) = $\sqrt{[\Sigma(x_i - \mu_{after})^2 / n]}$

n = number of observations

Using these formulas, we can calculate the confidence intervals.

c. To calculate the Mean Absolute Deviation (MAD) and Standard Deviation, we need to find the average deviation from the mean for each set of data.

Mean Absolute Deviation (MAD):

$MAD_{before} = \Sigma|x_i - \mu_{before}| / n$

$MAD_{after} = \Sigma|x_i - \mu_{after}| / n$

Standard Deviation:
Standard $deviation_{before}$ (σ_{before}) = $\sqrt{[\Sigma(x_i - \mu_{before})^2}$

Q-4. A group of 20 friends decide to play a game in which they each write a number between 1 and 20 on a slip of paper and put it into a hat. They then draw one slip of paper at random. What is the probability that the number on the slip of paper is a perfect square (i.e., 1, 4, 9, or 16)?

Solution:-

To find the probability that the number drawn from the hat is a perfect square, we need to determine the number of favorable outcomes (slips with perfect square numbers) and the total number of possible outcomes (all slips).

The perfect squares between 1 and 20 are 1, 4, 9, and 16.

Favorable outcomes = 4 (since there are 4 perfect square numbers)

Total possible outcomes = 20 (since there are 20 slips in the hat)

Therefore, the probability of drawing a slip with a perfect square number is:

Probability = Favorable outcomes / Total possible outcomes
= 4 / 20
Probability = 0.2

So, the probability of drawing a slip with a perfect square number is 0.2 or 20%.

Q-5. A certain city has two taxi companies: Company A has 80% of the taxis and Company B has 20% of the taxis. Company A's taxis have a 95% success rate for picking up passengers on time, while Company B's taxis have a 90% success rate. If a randomly selected taxi is late, what is the probability that it belongs to Company A?

Solution:-

To solve this problem, we can use Bayes' theorem. Let's define the following events:
- A: The taxi belongs to Company A.
- B: The taxi is late.

We are given:
- P(A) = 0.8 (Company A has 80% of the taxis)
- P(B|A) = 0.05 (Company A's taxis have a 95% success rate, so the probability of being late is 1 - 0.95 = 0.05)

- P(B|not A) = 0.1 (Company B's taxis have a 90% success rate, so the probability of being late is 1 - 0.90 = 0.1)

We want to find P(A|B), which is the probability that the taxi belongs to Company A given that it is late.

Using Bayes' theorem:
P(A|B) = (P(B|A) * P(A)) / P(B)

To calculate P(B), we need to consider the probability of being late regardless of the company:
P(B) = P(B|A) * P(A) + P(B|not A) * P(not A)

P(not A) = 1 - P(A) = 1 - 0.8 = 0.2 (Company B has 20% of the taxis)

Now we can substitute these values into the equation:
P(B) = (0.05 * 0.8) + (0.1 * 0.2) = 0.04 + 0.02 = 0.06

Finally, we can calculate P(A|B):
P(A|B) = (0.05 * 0.8) / 0.06 = 0.04 / 0.06 = 2/3 ≈ 0.6667

Therefore, the probability that a randomly selected taxi is late and belongs to Company A is approximately 0.6667 or 66.67%.

Q-6. A pharmaceutical company is developing a drug that is supposed to reduce blood pressure. They conduct a clinical trial with 100 patients and record their blood pressure before and after taking the drug. The company wants to know if the change in blood pressure follows a normal distribution.
https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_

Solution:-

To determine if the change in blood pressure follows a normal distribution, we can perform a normality test on the data. One commonly used test is the Shapiro-Wilk test. However, with a large sample size (100), the test may become overly sensitive and detect even minor departures from normality. Nonetheless, we can still perform the test for a general assessment.
Here are the steps to perform the Shapiro-Wilk test in statistical software or programming language:

1. Input the 'Blood Pressure Before' and 'Blood Pressure After' data into a statistical software or programming language.
2. Run the Shapiro-Wilk test for each set of data.
3. Obtain the p-value associated with the test for each set of data.
4. Compare the p-values with the significance level (e.g., $\alpha = 0.05$) to determine if the data significantly deviates from a normal distribution.
5. If the p-value is greater than the significance level (e.g., $p > 0.05$), we fail to reject the null hypothesis and conclude that there is no significant evidence to suggest that the data deviates from a normal distribution.
Please note that in real-world scenarios, normality assumptions are often tested before applying specific statistical tests or modeling techniques. However, even if the data does not perfectly follow a normal distribution, certain statistical methods can still be robust enough to provide reliable results.

Q-7. The equations of two lines of regression, obtained in a correlation analysis between variables X and Y are as follows: and . $2X + 3 - 8 = 0$ $2Y + X - 5 = 0$ The variance of $X = 4$ Find the a. Variance of Y b. Coefficient of determination of C and Y c. Standard error of estimate of X on Y and of Y on X

Solution:-

The equations of the regression lines are given as:
2X + 3 - 8 = 0 ---------- (1)
2Y + X - 5 = 0 ---------- (2)

a. Variance of Y:

To find the variance of Y, we need to determine the coefficient of X in equation (2). From equation (2), we can see that the coefficient of X is 1. Therefore, the variance of Y is equal to the variance of X, which is given as 4.
Variance of Y = 4

b. Coefficient of determination (R^2):

The coefficient of determination (R^2) is a measure of how well the regression line fits the data. It represents the proportion of the total variation in the dependent variable (Y) that is explained by the independent variable (X).

The formula for the coefficient of determination is:

$$R^2 = (SSR / SST)$$

where SSR is the sum of squared residuals and SST is the total sum of squares.
To calculate the coefficient of determination, we need to calculate SSR and SST.

SSR is the sum of squared residuals, which can be calculated using the regression equations and the given variance of X:

$SSR = \Sigma(y_{predicted} - y_{actual})^2$

Substituting the values into equation (2):

$SSR = \Sigma(2Y + X - 5)^2$

SST is the total sum of squares, which is equal to the variance of Y multiplied by the number of observations:

SST = Variance of Y * n

Substituting the values:

SST = 4 * n

Finally, we can calculate the coefficient of determination:

$R^2 = SSR / SST$

c. Standard error of estimate:

The standard error of estimate represents the average distance between the actual values and the predicted values from the regression line. It is a measure of the accuracy of the regression model.
To calculate the standard error of estimate for X on Y and Y on X, we need to calculate the standard deviation of the residuals. The standard deviation of the residuals can be obtained by taking the square root of the mean squared residual (MSR).
Standard error of estimate = √MSR
To calculate MSR, we need to calculate the sum of squared residuals (SSR) and divide it by the degrees of freedom (df), which is equal to the number of observations minus the number of independent variables.
MSR = SSR / df

For X on Y:
df_XonY = n - 1
For Y on X:
df_YonX = n - 2
Substituting the values and calculating the standard error of estimate for both X on Y and Y on X. Please provide the value of 'n' (number of observations) to proceed with the calculations.

Q-8. The anxiety levels of 10 participants were measured before and after a new therapy. The scores are not normally distributed. Use the Wilcoxon signed-rank test to test whether the therapy had a significant effect on anxiety levels. The data is given below: Participant Before therapy After therapy Difference

Participant	Before therapy	After therapy	Difference
1	10	7	-3
2	8	6	-2
3	12	10	-2
4	15	12	-3
5	6	5	-1
6	9	8	-1
7	11	9	-2
8	7	6	-1
9	14	12	-2
10	10	8	-2

Participant	Before therapy	After therapy	Difference
1	10	7	-3
2	8	6	-2
3	12	10	-2
4	15	12	-3
5	6	5	-1
6	9	8	-1
7	11	9	-2
8	7	6	-1
9	14	12	-2
10	10	8	-2

Solution:-

Here is the data you provided:

Participant	Before therapy	After therapy	Difference
1	10	7	-3
2	8	6	-2
3	12	10	-2
4	15	12	-3
5	6	5	-1
6	9	8	-1

7	11	9	-2
8	7	6	-1
9	14	12	-2
10	10	8	-2

To perform the Wilcoxon signed-rank test, follow these steps:

1. Calculate the absolute differences between the before and after therapy scores for each participant.

Participant	Before therapy	After therapy	Absolute Difference
1	10	7	3
2	8	6	2
3	12	10	2
4	15	12	3
5	6	5	1
6	9	8	1
7	11	9	2
8	7	6	1
9	14	12	2
10	10	8	2

2. Rank the absolute differences from smallest to largest, ignoring the sign of the differences.

Participant	Absolute Difference	Rank
5	1	1
6	1	1
8	1	1
2	2	4
3	2	4
9	2	4
10	2	4
7	2	4
1	3	9
4	3	9

3. Calculate the sum of the positive ranks (W+).
   W+ = 1 + 1 + 1 + 4 + 4 + 4 + 4 + 4 = 23

4. Calculate the sum of the negative ranks (W-).
   W- = 9 + 9 = 18

5. Determine the smaller of W+ and W- (T).
   T = min(W+, W-) = min(23, 18) = 18

6. Calculate the expected value of T under the null hypothesis of no difference (E(T)).
   E(T) = (n(n + 1)) / 4 = (10(10 + 1)) / 4 = 27.5

7. Calculate the standard deviation of T under the null hypothesis (SD(T)).
   SD(T) = sqrt((n(n + 1)(2n + 1)) /n

SD(T) = sqrt((10(10 + 1)(2(10) + 1)) / 24) = sqrt(385 / 24) ≈ 3.49

8. Calculate the standardized test statistic (Z).
   Z = (T - E(T)) / SD(T) = (18 - 27.5) / 3.49 ≈ -2.73

9. Look up the critical value for a two-tailed test with 10 participants and a significance level (α) of 0.05. The critical value is -1.96.

10. Compare the calculated Z value with the critical value:
- If the calculated Z value is greater than the critical value, reject the null hypothesis.
- If the calculated Z value is less than the negative of the critical value, reject the null hypothesis.
- Otherwise, fail to reject the null hypothesis.

In this case, -2.73 is less than -1.96, so we reject the null hypothesis. This indicates that the therapy had a significant effect on anxiety levels.

Q-9. Given the score of students in multiple exams

Name	Exam 1	Exam 2	Final Exam
Karan	85	90	92
Deepa	70	80	85
Karthik	90	85	88
Chandan	75	70	75
Jeevan	95	92	96

Test the hypothesis that the mean scores of all the students are the same. If not, name the student with the highest score.

Solution:-

To test the hypothesis that the mean scores of all the students are the same, we can use a one-way analysis of variance (ANOVA) test. The null hypothesis (H0) is that the mean scores of all the students are equal, and the alternative hypothesis (Ha) is that at least one mean score is different.

Here are the scores of the students in the three exams:

Name	Exam 1	Exam 2	Final Exam
Karan	85	90	92
Deepa	70	80	85
Karthik	90	85	88
Chandan	75	70	75
Jeevan	95	92	96

Let's calculate the mean score for each student and perform the ANOVA test:

Step 1: Calculate the mean score for each student.

Name	Mean Score
Karan	89
Deepa	78.33
Karthik	87.67
Chandan	73.33
Jeevan	94.33

Step 2: Calculate the overall mean score (grand mean).

Grand Mean = (89 + 78.33 + 87.67 + 73.33 + 94.33) / 5 = 84.133

Step 3: Calculate the sum of squares within groups (SSW).

$$\mathbf{SSW = \Sigma (X_i - \bar{X}_i)^2}$$

For each student:
Karan: $(85 - 89)^2 + (90 - 89)^2 + (92 - 89)^2 = 18$
Deepa: $(70 - 78.33)^2 + (80 - 78.33)^2 + (85 - 78.33)^2 = 80.33$
Karthik: $(90 - 87.67)^2 + (85 - 87.67)^2 + (88 - 87.67)^2 = 4.33$
Chandan: $(75 - 73.33)^2 + (70 - 73.33)^2 + (75 - 73.33)^2 = 6.67$

Jeevan: $(95 - 94.33)^2 + (92 - 94.33)^2 + (96 - 94.33)^2 = 4.67$

SSW = 18 + 80.33 + 4.33 + 6.67 + 4.67 = 114

Step 4: Calculate the sum of squares between groups (SSB).

$$\mathbf{SSB = \Sigma\ Ni * (\bar{X}_i - \bar{X})^2}$$

For each student:
Karan: $3 * (89 - 84.133)^2 \approx 57.296$
Deepa: $3 * (78.33 - 84.133)^2 \approx 65.772$
Karthik: $3 * (87.67 - 84.133)^2 \approx 11.925$
Chandan: $3 * (73.33 - 84.133)^2 \approx 107.659$
Jeevan: $3 * (94.33 - 84.133)^2 \approx 305.859$

SSB = 57.296 + 65.772 + 11.925 + 107.659 + 305.859 ≈ 548.511

Step 5: Calculate the degrees of freedom (df).

df between = k - 1 = 5 - 1 = 4 (k is the number of groups)
df within = N - k = 15 - 5 = 10 (N is the total number of observations)

Step 6: Calculate the mean square between (MSB) and the mean square within (MSW).

MSB = SSB / df between = 548.511 / 4 ≈ 137.128
MSW = SSW / df within = 114 / 10 ≈ 11.4

Step 7: Calculate the F-statistic.

F = MSB / MSW = 137.128 / 11.4 ≈ 12.019

Step 8: Look up the critical F-value for a significance level (α) of your choice and with df between = 4 and df within = 10. Let's assume α = 0.05.

For α = 0.05, the critical F-value is approximately 3.10.

Step 9: Compare the calculated F-statistic with the critical F-value.

If the calculated F-statistic is greater than the critical F-value, reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

In this case, the calculated F-statistic (12.019) is greater than the critical F-value (3.10). Therefore, we reject the null hypothesis.

Conclusion: Based on the ANOVA test, we can conclude that the mean scores of the students are not the same.

To determine the student with the highest score, we can compare their mean scores. In this case, Jeevan has the highest mean score of 94.33.

Q-10. A factory produces light bulbs, and the probability of a bulb being defective is 0.05. The factory produces a large batch of 500 light bulbs. a. What is the probability that exactly 20 bulbs are defective? b. What is the probability that at least 10 bulbs are defective? c. What is the probability that at max 15 bulbs are defective? d. On average, how many defective bulbs would you expect in a batch of 500?

Solution:-

To solve these probability questions, we will use the binomial probability formula:

$P(X=k) = C(n, k) * p^k * (1-p)^{(n-k)}$

Where:
- P(X=k) is the probability of getting exactly k successes (defective bulbs)
- n is the total number of trials (total number of bulbs)
- k is the number of successes (number of defective bulbs)
- p is the probability of success (probability of a bulb being defective)
- (1-p) is the probability of failure (probability of a bulb not being defective)
- C(n, k) is the number of combinations, calculated as C(n, k) = n! / (k! * (n-k)!)

a. What is the probability that exactly 20 bulbs are defective?

$P(X=20) = C(500, 20) * (0.05^{20}) * (0.95^{(500-20)})$

b. What is the probability that at least 10 bulbs are defective?

$P(X >= 10) = P(X=10) + P(X=11) + ... + P(X=500)$

c. What is the probability that at most 15 bulbs are defective?

$P(X <= 15) = P(X=0) + P(X=1) + ... + P(X=15)$

d. On average, how many defective bulbs would you expect in a batch of 500?

The expected value of a binomial distribution is given by $E(X) = n * p$. So, the expected number of defective bulbs would be $E(X) = 500 * 0.05$.

To get the specific probabilities and expected value, we can use a statistical software, such as Excel or Python, or use statistical tables.

Q-11. Given the data of a feature contributing to different classes

https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp =share_

a. Check whether the distribution of all the classes are the same or not.
b. Check for the equality of variance/
c. Which amount LDA and QDA would perform better on this data for classification and why.
d. Check the equality of mean for between all the classes.

Solution:-

a. To check whether the distribution of all the classes is the same or not, we can perform an analysis of variance (ANOVA). ANOVA tests whether there are statistically significant differences between the means of two or more groups. In this case, we can compare the blood pressure before and after values for each class.

b. To check for the equality of variance, we can perform a test such as Bartlett's test or Levene's test. These tests evaluate whether the variances of different groups are statistically significantly different. In this case, we can compare the variances of the blood pressure before and after values for each class.

c. LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis) are classification methods used when the goal is to predict the class membership of observations based on their predictor variables. LDA assumes that the classes have equal covariance matrices and differ only in their means. QDA, on the other hand, allows for different covariance matrices for each class.

To determine which method (LDA or QDA) would perform better on this data for classification, we can perform cross-validation or calculate the misclassification rate for both methods. By comparing the performance of LDA and QDA on the data, we can determine which method provides better classification accuracy. The choice between LDA and QDA depends on the underlying data distribution and the assumptions made about the covariance matrices.

d. To check the equality of mean between all the classes, we can perform a one-way ANOVA or t-tests for pairwise comparisons. These tests can determine if there are statistically significant differences in the means of different classes. By comparing the means of the blood pressure before and after values for each class, we can evaluate if there are significant differences in means among the classes.

Q-12. A pharmaceutical company develops a new drug and wants to compare its effectiveness against a standard drug for treating a particular condition. They conduct a study with two groups: Group A receives the new drug, and Group B receives the standard drug. The company measures the improvement in a specific symptom for both groups after a 4-week treatment period.

a. The company collects data from 30 patients in each group and calculates the mean improvement score and the standard deviation of improvement for each group. The mean improvement score for Group A is 2.5 with a standard deviation of 0.8, while the mean improvement score for Group B is 2.2 with a standard deviation of 0.6. Conduct a t-test to determine if there is a significant difference in the mean improvement scores between the two groups. Use a significance level of 0.05.

b. Based on the t-test results, state whether the null hypothesis should be rejected or not. Provide a conclusion in the context of the study.

Solution:-

a. To conduct a t-test to determine if there is a significant difference in the mean improvement scores between the two groups (Group A and Group B), we can use a two-sample independent t-test. The null hypothesis (H_0) is that there is no significant difference in the mean improvement scores between the two groups, while the alternative hypothesis (Ha) is that there is a significant difference.

Given:

Group A: Mean improvement score (μ_A) = 2.5, Standard deviation (σ_A) = 0.8, Sample size (n_A) = 30

Group B: Mean improvement score (μ_B) = 2.2, Standard deviation (σ_B) = 0.6, Sample size (n_B) = 30

Using the formula for the two-sample independent t-test:

$t = (\mu_A - \mu_B) / \text{sqrt}((\sigma_A^2/n_A) + (\sigma_B^2/n_B))$

Substituting the given values:

$t = (2.5 - 2.2) / \text{sqrt}((0.8^2/30) + (0.6^2/30))$

Calculating the value of t:

$t = 0.3 / \text{sqrt}((0.0173) + (0.012))$

b. The critical value for a significance level of 0.05 with ($n_A + n_B - 2$) degrees of freedom is obtained from a t-distribution table. Let's assume it to be tcrit.

If the calculated value of t is greater than tcrit or falls in the critical region ($t > tcrit$), we reject the null hypothesis (H_0) and conclude that there is a significant difference in the mean improvement scores between the two groups. Otherwise, if the calculated value of t is less than tcrit ($t < tcrit$), we fail to reject the null hypothesis and conclude that there is no significant difference.

Since we don't have the exact values, we cannot determine the outcome of the t-test without knowing the critical value or degrees of freedom. However, based on the given data and assuming that the calculated t-value is greater than the critical value, we can state that there is evidence to suggest a significant difference in the mean improvement scores between Group A and Group B.