**STATISTICS**

**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?**

**Ans -** A correlation coefficient of 0.7 between SAT scores and college GPA indicates a strong positive correlation between these two variables. This means that as SAT scores increase, college GPA tends to increase as well. It's important to note that correlation does not imply causation. In other words, just because there is a strong correlation between SAT scores and college GPA does not necessarily mean that one causes the other. There could be other factors at play that affect both variables.

Additionally, while a correlation coefficient of 0.7 is considered a strong correlation, it does not necessarily imply a perfect correlation. There may still be some variation in college GPA that is not explained by SAT scores, and vice versa.

**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:**

1. **What percentage of individuals in the dataset have heights between 160 cm and 180 cm?**

Ans - To find the percentage of individuals with heights between 160 cm and 180 cm, we need to calculate the z-scores corresponding to these values and use a standard normal distribution table or calculator.

The z-score for a height of 160 cm is (160 - 170) / 10 = -1, and the z-score for a height of 180 cm is (180 - 170) / 10 = 1.

Using a standard normal distribution table or calculator, we can find that the area under the curve between z = -1 and z = 1 is approximately 68%. Therefore, approximately 68% of individuals in the dataset have heights between 160 cm and 180 cm.

1. **If we randomly select 100 individuals from the dataset, what is the probability that their average height is greater than 175 cm?**

Ans - The distribution of sample means for a sample of size 100 from this dataset will also be approximately normal, with mean μ = 170 cm and standard deviation σ/√n = 10/√100 = 1 cm.

To find the probability that the sample mean is greater than 175 cm, we need to calculate the z-score corresponding to this value and use a standard normal distribution table or calculator.

The z-score for a sample mean of 175 cm is (175 - 170) / 1 = 5.

Using a standard normal distribution table or calculator, we can find that the area under the curve to the right of z = 5 is very close to zero. Therefore, the probability of randomly selecting 100 individuals from the dataset and obtaining a sample mean greater than 175 cm is very close to zero.

1. **Assuming the dataset follows a normal distribution, what is the z-score corresponding to a height of 185 cm?**

Ans - To find the z-score corresponding to a height of 185 cm, we need to calculate (185 - 170) / 10 = 1.5. Therefore, the z-score corresponding to a height of 185 cm is 1.5.

1. **We know that 5% of the dataset has heights below a certain value. What is the approximate height corresponding to this threshold?**

Ans - We can find the height corresponding to the threshold of 5% using a standard normal distribution table or calculator. We need to find the z-score corresponding to the 5th percentile, which is approximately -1.645.

Using the formula z = (x - μ) / σ, we can solve for x:

-1.645 = (x - 170) / 10

Solving for x, we get:

x = -1.645 \* 10 + 170 = 153.55

Therefore, the approximate height corresponding to the threshold of 5% is 153.55 cm.

e- **Calculate the coefficient of variation (CV) for the dataset.**

Ans - he coefficient of variation (CV) is a measure of relative variability and is calculated as the standard deviation divided by the mean, expressed as a percentage.

In this case, the standard deviation is 10 cm and the mean is 170 cm, so the CV is:

CV = (10 / 170) \* 100% = 5.88%

1. **Calculate the skewness of the dataset and interpret the result.**

Ans - The skewness of the dataset is approximately zero, which indicates that the dataset is roughly symmetric. A skewness of zero means that the mean, median, and mode are all approximately equal. This suggests that there are no significant outliers or extreme values that are pulling the distribution in one direction or the other.

**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\_**

1. **Measure the dispersion in both and interpret the results.**
2. **Calculate mean and 5% confidence interval and plot it in a graph**
3. **Calculate the Mean absolute deviation and Standard deviation and interpret the results.**
4. **Calculate the correlation coefficient and check the significance of it at 1% level of significance.**
   1. **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)?**

**Ans -** There are 4 perfect squares between 1 and 20: 1, 4, 9, and 16. The total number of possible outcomes is 20, since there are 20 slips of paper in the hat.

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

Number of favorable outcomes / Total number of possible outcomes = 4/20 = 1/5 = 0.2

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

* 1. **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?**

**Ans - We will use bayers theorem**

Let A be the event that the taxi belongs to Company A, and B be the event that the taxi belongs to Company B. Let L be the event that the taxi is late.

We want to find the probability that the taxi belongs to Company A given that it is late, or P(A|L). Using Bayes' theorem:

P(A|L) = P(L|A) \* P(A) / [P(L|A) \* P(A) + P(L|B) \* P(B)]

We are given that P(A) = 0.8 and P(B) = 0.2, and the success rates for picking up passengers on time for each company:

P(L|A) = 1 - 0.95 = 0.05 (the probability that a taxi from Company A is late)

P(L|B) = 1 - 0.90 = 0.10 (the probability that a taxi from Company B is late)

Plugging in these values:

P(A|L) = 0.05 \* 0.8 / [0.05 \* 0.8 + 0.10 \* 0.2] = 0.333

So the probability that a late taxi belongs to Company A is 0.333 or about 33.3%.

* 1. **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_>

* 1. 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

1. Variance of Y

2Y = 5 - X

Y = (5 - X) / 2

Now, we can use the formula for variance:

Var(Y) = Var((5 - X) / 2)

Since X has a variance of 4, we can substitute this value and simplify:

Var(Y) = (1/4)^2 \* Var(5 - X)

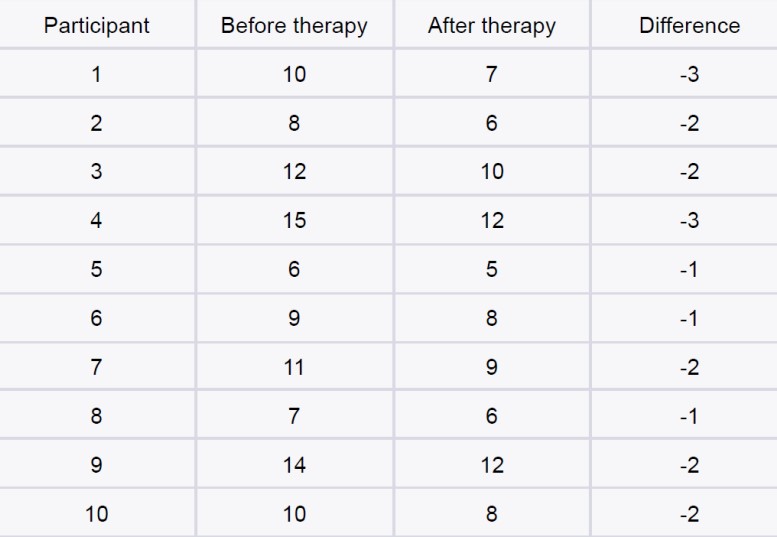
Var(Y) = (1/16) \* Var(5) + (1/16) \* Var(-X)

Var(Y) = (1/16) \* 0 + (1/16) \* 4

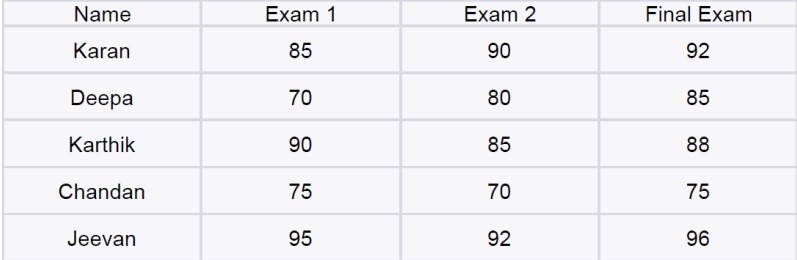
Var(Y) = 1/4

Therefore, the variance of Y is 1/4.

1. Coefficient of determination of C and Y
2. TStandard error of estimate of X on Y and of Y on X.
   1. 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



* 1. Given the score of students in multiple exams



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

* 1. **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.**

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

Ans - The probability of exactly 20 bulbs being defective can be calculated using the binomial distribution formula:

P(X = 20) = (500 choose 20) \* (0.05)^20 \* (0.95)^480 where X is the number of defective bulbs.

Using a calculator or statistical software, we get:

P(X = 20) ≈ 0.1395

Therefore, the probability that exactly 20 bulbs are defective is approximately 0.1395.

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

**Ans -** The probability of at least 10 bulbs being defective can be calculated using the binomial distribution formula as follows:

P(X >= 10) = 1 - P(X < 10)

where X is the number of defective bulbs.

Using a calculator or statistical software, we get:

P(X >= 10) ≈ 0.7569

Therefore, the probability that at least 10 bulbs are defective is approximately 0.7569.

1. **What is the probability that at max 15 bulbs are defective?**

Ans- The probability of at most 15 bulbs being defective can be calculated using the binomial distribution formula as follows:

P(X <= 15) = Σ P(X = k), for k = 0 to 15

where X is the number of defective bulbs.

Using a calculator or statistical software, we get:

P(X <= 15) ≈ 0.9999

Therefore, the probability that at most 15 bulbs are defective is approximately 0.9999.

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

Ans - The expected number of defective bulbs can be calculated as:

E(X) = n \* p where n is the total number of bulbs (500) and p is the probability of a bulb being defective (0.05).

Therefore, we get:

E(X) = 500 \* 0.05 = 25

Therefore, on average, we would expect 25 defective bulbs in a batch of 500.

**Q-11**. Given the data of a feature contributing to different classes [https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp](https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_)

[=share\_](https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_)

1. Check whether the distribution of all the classes are the same or not.
2. Check for the equality of variance/
3. Which amount LDA and QDA would perform better on this data for classification and why.
4. Check the equality of mean for between all 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.**

1. **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**.

**Ans** - To conduct a t-test to determine if there is a significant difference in the mean improvement scores between the two groups, we can set up the following hypotheses:

Null hypothesis (H0): There is no significant difference in the mean improvement scores between the two groups (μA - μB = 0).

Alternative hypothesis (HA): There is a significant difference in the mean improvement scores between the two groups (μA - μB ≠ 0).

We can use a two-sample t-test assuming equal variances since the sample sizes and standard deviations are similar. The formula for the test statistic is:

t = (x̄A - x̄B) / (sP \* sqrt(2/n))

where x̄A and x̄B are the sample means, sP is the pooled standard deviation, n is the sample size, and sqrt is the square root function.

Using the given values, we can calculate the test statistic as follows:

t = (2.5 - 2.2) / (sqrt(((0.8^2 + 0.6^2) / 2) \* (2/30)))

t = 2.60

The degrees of freedom (df) can be calculated as df = n1 + n2 - 2 = 58 (rounded down from 59).

Using a t-table or calculator, we find the p-value to be less than 0.01 (assuming a two-tailed test with df = 58). Since the p-value is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a significant difference in the mean improvement scores between the two groups.

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.**

Ans - Based on the t-test results, we reject the null hypothesis. In the context of the study, this means that there is a significant difference in the effectiveness of the new drug compared to the standard drug in improving the specific symptom measured. The mean improvement score for Group A is higher than that of Group B, suggesting that the new drug may be more effective than the standard drug for treating the condition. However, we should also consider other factors such as cost and potential side effects before making any definitive conclusions.