**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?

Explanation ➖

The utility of SAT scores is mainly focused on student outcomes for college admission decisions by the organization. The students with high SAT scores have an assumption to have higher grades in college, it also effects the overall college success.

Positive correlation coefficient will help identify students to score high grades in colleges.The students who scores higher on the SAT will automatically indicates that they have a higher GPA. Because by achieving higher GPA which indicates that their dedication towards school work will be much higher.

As mentioned in the above prediction that the correlation coefficient between SAT scores and college GPA is 0.7 indicates that there is 70% relationship where GPA is an Indepedent variable whileas SAT scores is dependent variable. Correlation coefficient varies from -1 to 1 whereas positive relationship is always better to understand the relationship between two variables.

Thus we can conclude that GPA does in fact influence high score of SAT.

**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?

Answer -

68% 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?

(lower value, upper value, *μ*,

σ

√

n

Probability is 1

c. Assuming the dataset follows a normal distribution, what is the z-score

corresponding to a height of 185 cm?

z-score =                              Value − Mean

                                     Standard deviation

(185 - 170)/10 = 1.5

d. We know that 5% of the dataset has heights below a certain value. What is

the approximate height corresponding to this threshold?

Two threshold value is 95% and remaining 5% of the dataset lies between a certain value which is 150 and below.

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

Answer -

* CV = standard deviation / sample mean x 100 =

10/170 \* 100 = 5.88 is the CV

f. Calculate the skewness of the dataset and interpret the result.

Answer Skewness is a measure of [symmetry](https://www.statisticshowto.com/symmetry-of-a-function/) in a distribution.-

 Skew = 3 \* (Mean – Median) / Standard Deviation.

= 0.001

Hence skewness is not there.

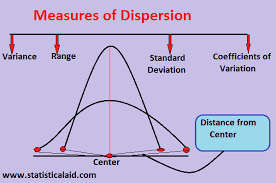
**Q-3**. Consider the ‘Blood Pressure Before’ and ‘Blood Pressure After’ columns from the

data and calculate the following

1. Measure the dispersion in both and interpret the results.

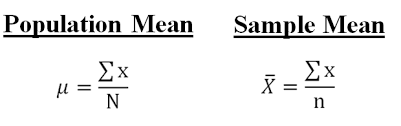
Explanation – Measure of Dispersion(Which talks about spread)

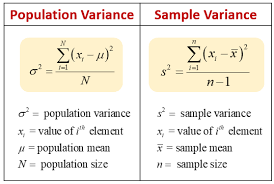
Standard Deviation and Variance



* Bell Curve is also called Normal Distribution.
* The central element can be either mean,median or mode(Measure of Central tendency)
* The spread of the curve mainly determines Variance and Standard Deviation

Population(N) Sample(n)





**Variance and Standard deviation**

* Dispersion is the quantification of deviation of each point from the mean value.
* Variance is average of squared distances of each point from the mean
* Variance is a fairly good measure of dispersion.
* σ2=∑ni=1(xi−x¯)2n
* **Variance Calculation**

Ipynb file path =

<https://github.com/Manasabhargi/Placement-Assignment_Manasa-SM/blob/main/Statistics/Statistics%20Q-3.ipynb>

**(b)** Calculate mean and 5% confidence interval and plot it in a graph

### 1. Find the sample mean

### 2. Calculate the standard deviation

### 3. Find the standard error

### 4. Find the margin of error

### 5. Use these results in the formula

### 6. Interpret your results

**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:- The perfect squared numbers are 1,4,9 or 16.

Total number = 4

Probability = 4/20 = 1/5

**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?

95% SUCCESS RATE

Company A

80% 5% Late taxi

City

Company B 90% SUCCESS Rate

20% 10% late taxi

Probability(Company A taxi coming late is)=

80 \* 0.05 = 4/6 = 2/3

80 \* 0.05 + 20 \* 0.10

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.

Ipynb file path =

<https://github.com/Manasabhargi/Placement-Assignment_Manasa-SM/blob/main/Statistics/Statistics%20Q-6.ipynb>

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 given equations of the lines of regression are:

2x + 3y – 8 = 0 ---🡪 1

2y + x – 5 = 0 ----🡪2

## Rewriting the equations (i) and (ii), we have From equation (i) 2x + 3y – 8 = 0

## 2x = 8 - 3y

## x = 8/2 – 3/2y

## x = 4 – 1.5y 🡪 equation(iii)

## (regression line of y on x) byx ​= rσx​////σy ​ ​ = -1.5 ....(iii)

## From eqution (ii), 2y + x – 5 = 0 x = 5 – 2y ( regression line of x on y) bxy ​= r σy/​σx​​ ∴r2 = byx ​× bxy ​= (-1.5) × (−2) = 3 ∴ r = sqrt(3)​ = +/- 1.732

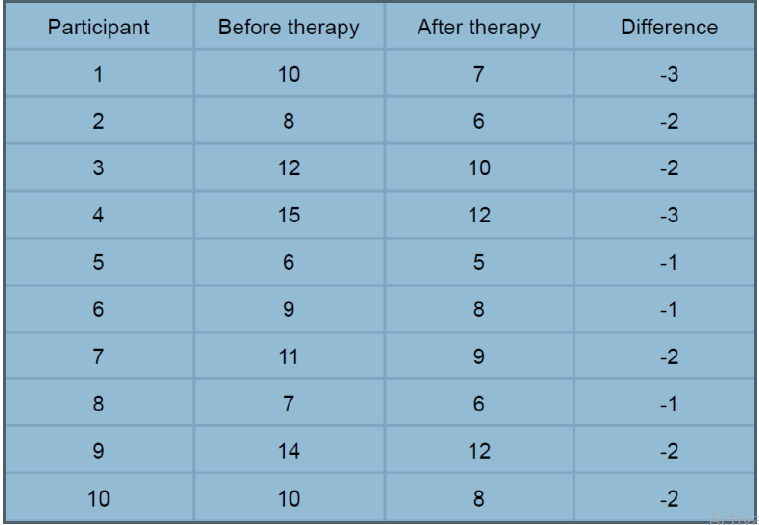
## But bxy​ and byx​ being both −ve therefore, r is also −ve.

## Correlation coefficient (r)=−1.732

## Varianceof x i.e., σx2​=4 ∴ σx​=4

## From equation (iii) rσx/​σy ​​= −1.5 −1.732 \* (​σy/√4)​​=−1.5 σy​=( -1.5 \* 2) / -1.732 ​​ = 1.732 ∴ Variance of y i.e., σy2​ = 1.732 \* 1.732 = 3

## 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



## The next step is to rank the difference scores. We first order the absolute values of the difference scores and assign rank from 1 through n to the smallest through largest absolute values of the difference scores, and assign the mean rank when there are ties in the absolute values of the difference scores.

|  |  |  |  |
| --- | --- | --- | --- |
| **Observed Differences** |  | **Ordered Absolute Values of Differences** | **Ranks** |
| -3 |  | -1 | -2 |
| -2 |  | -1 | -2 |
| -2 |  | -1 | -2 |
| -3 |  | -2 | -6 |
| -1 |  | -2 | -6 |
| -1 |  | -2 | -6 |
| -2 |  | -2 | -6 |
| -1 |  | -2 | -6 |
| -2 |  | -3 | -9.5 |
| -2 |  | -3 | -9.5 |
|  |  |  |  |

**Step 3: calculating the sum of positive and sum of negative ranks**

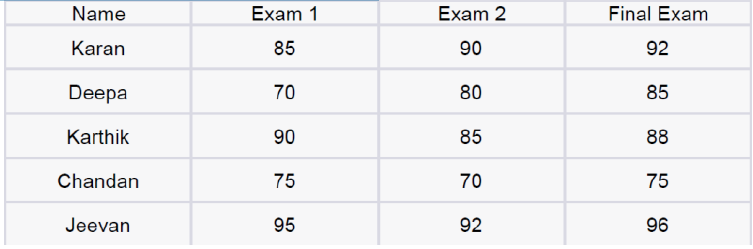
**All are negative ranks so,**

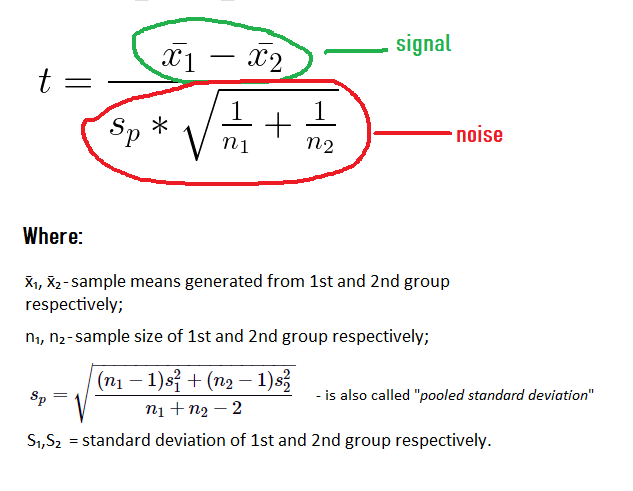
**W- = 55**

**Step 4: determining Wilcoxon test statistic W**

**We are rejecting the H0 hypothesis and accepting H1 hypothesis.**

## Q-9. 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.





## Jeevan is the student with the highest score.

## 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?