STATISTICS:

Q1:
ANS: The correlation coefficient of 0.7 between SAT scores and college GPA indicates a strong positive linear relationship between these two variables. In other words, as SAT scores increase, college GPA tends to increase as well, and vice versa.

Here's a breakdown of what different correlation coefficients typically indicate about the relationship between two variables:

1. Perfect Positive Correlation (r = 1.0): This would mean that as one variable increases, the other also increases in a perfectly linear fashion. In your case, a perfect positive correlation would imply that as SAT scores increase, college GPA also increases in a perfectly predictable manner.

2. Strong Positive Correlation (0.7 < r < 1.0): A correlation coefficient of 0.7 falls into the category of a strong positive correlation. It suggests that there is a strong tendency for students with higher SAT scores to have higher college GPAs, but it's not a perfect relationship. There can still be some variability in GPA among students with the same SAT scores.

3. Weak or No Correlation (0 < r < 0.3): If the correlation coefficient were closer to 0, it would indicate a weak or no linear relationship between SAT scores and college GPA. In other words, there would be little to no predictable connection between the two variables.

4. Perfect Negative Correlation (r = -1.0): A correlation coefficient of -1.0 would indicate a perfect negative linear relationship, meaning that as SAT scores increase, college GPA decreases perfectly predictably. However, this is not the case in your scenario.

In summary, a correlation coefficient of 0.7 suggests a strong positive linear relationship between SAT scores and college GPA, meaning that, on average, students with higher SAT scores tend to have higher college GPAs. However, it's important to note that correlation does not imply causation, and other factors could also be influencing college GPA.
—----------

Q 2:
ANS:
To find the percentage of individuals in the dataset with heights between 160 cm and 180 cm, you can use the properties of the normal distribution and the Z-score formula.

1. Calculate the Z-scores for both 160 cm and 180 cm using the formula:

$$Z = (X - \mu) / \sigma$$

Where:
- X is the value (height in this case)
- μ is the mean (170 cm)
- σ is the standard deviation (10 cm)

For 160 cm:
Z1 = (160 - 170) / 10 = -1
For 180 cm:
Z2 = (180 - 170) / 10 = 1

2. Look up the Z-scores in a standard normal distribution table or use a calculator to find the corresponding cumulative probabilities.
- For Z1 = -1, the cumulative probability is approximately 0.1587.
- For Z2 = 1, the cumulative probability is approximately 0.8413.

3. Calculate the percentage of individuals between these two Z-scores:
Percentage = (Cumulative probability at Z2) - (Cumulative probability at Z1)
Percentage = 0.8413 - 0.1587 = 0.6826

So, approximately 68.26% of individuals in the dataset have heights between 160 cm and 180 cm.

==========

B: ANS

To find the probability that the average height of a random sample of 100 individuals from the dataset is greater than 175 cm, you can use the properties of the sampling distribution of the sample mean. Since the dataset is approximately normally distributed, you can use the central limit theorem.

1. Calculate the standard error of the sample mean (standard deviation of the sample mean):

Standard Error (SE) = $\sigma / \sqrt{n}$

Where:
- σ is the population standard deviation (10 cm).
- n is the sample size (100 individuals).

SE = $10 / \sqrt{100} = 10 / 10 = 1$ cm

2. Calculate the Z-score for a sample mean of 175 cm:

$Z = (X - \mu) / SE$

Where:
- X is the value (175 cm).
- μ is the population mean (170 cm).
- SE is the standard error (1 cm).

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

3. Find the probability that a Z-score is greater than 5 using a standard normal distribution table or calculator. This represents the probability that the average height of the sample is greater than 175 cm.

P(Z > 5) is essentially zero because it is extremely unlikely to randomly select a sample of 100 individuals with an average height greater than 175 cm when the population mean is 170 cm and the population standard deviation is 10 cm.

So, the probability that the average height of a random sample of 100 individuals from the dataset is greater than 175 cm is practically zero.

==========

c:To find the Z-score corresponding to a height of 185 cm in a dataset with a normal distribution, you can use the following formula:

$Z = (X - \mu) / \sigma$

Where:

- X is the value you want to find the Z-score for (185 cm in this case).
- μ is the population mean (170 cm).
- σ is the population standard deviation (10 cm).

Now, plug in the values:

$Z = (185 - 170) / 10 = 15 / 10 = 1.5$

So, the Z-score corresponding to a height of 185 cm in this dataset is 1.5.

==========

D:

ANS :To find the approximate height corresponding to the threshold below which 5% of the dataset falls, we need to find the z-score that corresponds to the 5th percentile of the standard normal distribution. This is because the z-score represents the number of standard deviations a value is from the mean.

Using a standard normal distribution table or calculator, you can find the z-score that corresponds to the 5th percentile. This z-score is approximately -1.645.

Now, we can use this z-score to find the height:

$X=\mu+(z*\sigma)=170+(-1.645*10)=170-16.45=153.55X=\mu+(z*\sigma)=170+(-1.645*10)=170-16.45=153.55$

So, the approximate height corresponding to the threshold below which 5% of the dataset falls is approximately 153.55 cm.

==========

E :

ANS:The coefficient of variation (CV) is a measure of relative variability and is calculated as the ratio of the standard deviation (σ) to the mean (μ) expressed as a percentage:

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

In this case:

$CV=(10170)*100\approx5.88\%CV=(17010)*100\approx5.88\%$

So, the coefficient of variation for the dataset is approximately 5.88%. This indicates that the standard deviation is about 5.88% of the mean, providing a measure of relative variability in the dataset

f:ANS To calculate the skewness of a dataset, you can use the following formula for sample skewness:

$$\text{Skewness} = \frac{\frac{1}{n} \sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

Where:
- n is the number of data points in the dataset.
- X_i represents each individual data point.
- $\bar{X}$ is the sample mean.
- s is the sample standard deviation.

Given the information you provided:
- You have a dataset containing the heights of 1000 individuals.
- The mean height is 170 cm (μ = 170 cm).
- The standard deviation is 10 cm (σ = 10 cm).

First, calculate the skewness using the formula:

$$\text{Skewness} = \frac{\frac{1}{1000} \sum_{i=1}^{1000}(X_i - 170)^3}{(10)^3}$$

This formula will give you the skewness value for your dataset.

Interpreting the skewness result:
- If the skewness is approximately zero, it suggests that the dataset is approximately normally distributed or symmetric.
- If the skewness is negative, it indicates a left-skewed (negatively skewed) distribution, meaning the tail on the left side of the distribution is longer or fatter than the right tail.
- If the skewness is positive, it indicates a right-skewed (positively skewed) distribution, meaning the tail on the right side of the distribution is longer or fatter than the left tail.

So, calculate the skewness, and based on the result, you can interpret whether the dataset is symmetric, left-skewed, or right-skewed.
==========

Q 3:

	Patient ID	Blood Pressure Before (mmHg)	Blood Pressure After (mmHg)
1	130	120	
2	142	135	
3	120	118	
4	135	127	
5	148	140	
6	122	118	
7	137	129	
8	130	124	

9	142	137
10	128	125
11	135	129
12	140	132
13	132	125
14	145	136
15	124	118
16	128	122
17	136	130
18	143	139
19	127	123
20	139	132
21	135	131
22	131	126
23	127	120
24	130	123
25	142	139
26	128	122
27	136	129
28	140	136
29	132	127
30	145	140
31	124	119
32	128	121
33	136	129
34	143	137
35	127	122
36	139	135
37	135	129
38	131	124
39	127	119
40	130	124
41	142	139
42	128	123
43	136	131
44	140	135
45	132	127
46	145	141
47	124	118
48	128	121
49	136	129
50	143	137
51	127	123
52	139	135
53	135	130
54	131	125
55	127	121
56	130	124

57 142 139
58 128 123
59 136 131
60 140 136
61 132 127
62 145 141
63 124 118
64 128 121
65 136 129
66 143 137
67 127 123
68 139 135
69 135 130
70 131 124
71 127 121
72 130 124
73 142 139
74 128 123
75 136 131
76 140 136
77 132 127
78 145 141
79 124 118
80 128 121
81 136 129
82 143 137
83 127 123
84 139 135
85 135 130
86 131 125
87 127 121
88 130 124
89 128 122
90 136 129
91 140 135
92 132 127
93 145 141
94 124 118
95 128 121
96 136 129
97 143 137
98 127 123
99 139 135
100 135 130
Measure of dispersion are range,varince,and standard deviation :

Range for before columns 28
Range for after column: 23

Variance for before column 42.28778445684844
Variance for after column : 46.95069033530572

Std .deviation for before column and std deviation for after column :
6.502905847146216
6.85205737974411

==========

B: **Step 1: Calculate the Mean (Average)**

The mean (average) of a dataset is calculated by summing up all the values in the dataset and then dividing by the total number of values. Mathematically, the mean (μ) is calculated as:

μ=∑i=1nxinμ=n∑i=1nxi

Where:

- μμ is the mean.
- xixi represents each individual data point.
- nn is the total number of data point

Step 2 is caculating 5% confidence level which is

μ+-Z*(sigma/sq.rt n)

C:

Mean Absolute Deviation (MAD) for Blood Pressure Before: 5.66

Mean Absolute Deviation (MAD) for Blood Pressure After: 5.92

Standard Deviation (SD) for Blood Pressure Before: 6.48

Standard Deviation (SD) for Blood Pressure After: 6.85

==========

D: step1 :find pearson rank correlation r

Step 2: Perform the Significance Test:

To check the significance of the correlation coefficient at the 1% level of significance, you need to perform a hypothesis test. The null and alternative hypotheses are as follows:

- Null Hypothesis (H0): There is no significant correlation (correlation coefficient = 0).

- Alternative Hypothesis (H1): There is a significant correlation (correlation coefficient ≠ 0).

Step 3: Calculate the t-statistic:

The t-statistic is calculated using the following formula:

t=r1−r2n−2t=n−21−r2

r

Where:

- rr is the correlation coefficient calculated in Step 1.
- nn is the number of data points.

Step 4: Find the Critical Value:

To check the significance at the 1% level, you need to find the critical value from a t-distribution table. Since it's a two-tailed test, you'll look up the critical value for a 0.5% level

Output :Pearson Correlation Coefficient: 0.98

T-Statistic: 63.69

P-Value: 0.0000

Is the correlation significant at 1% level? Yes-

—----------

4:To find the probability that the number on the slip of paper drawn from the hat is a perfect square (i.e., 1, 4, 9, or 16), you need to calculate the ratio of the number of favorable outcomes (perfect squares) to the total number of possible outcomes.

Total number of possible outcomes:

- There are 20 friends, each of whom can write any number between 1 and 20.

- So, there are 20 choices for each friend.

Total possible outcomes = \(20 \times 20 = 400\)

Number of favorable outcomes (perfect squares):

- There are four perfect squares between 1 and 20, which are 1, 4, 9, and 16.

- Each of the 20 friends can choose one of these four numbers.

Number of favorable outcomes = \(20 \times 4 = 80\)

Now, you can calculate the probability:

Probability = Number of Favorable Outcomes/Total Number of Possible Outcomes = 80/400=⅕ =20%

—----------

5 :To find the probability that a randomly selected taxi that is late belongs to Company A, you can use conditional probability and Bayes' Theorem.

Let:

- \(A \) be the event that a taxi is from Company A.

- \(B \) be the event that a taxi is from Company B.

- \(L \) be the event that a taxi is late.

- \(P(A) \) be the probability that a taxi is from Company A (0.80, as 80% of the taxis are from Company A).

- \(P(B) \) be the probability that a taxi is from Company B (0.20, as 20% of the taxis are from Company B).

- \(P(L|A) \) be the probability that a taxi is late given that it belongs to Company A (0.05, as Company A's taxis have a 5% chance of being late).

- \(P(L|B) \) be the probability that a taxi is late given that it belongs to Company B (0.10, as Company B's taxis have a 10% chance of being late).

You want to find \(P(A|L) \), the probability that a taxi belongs to Company A given that it is late.

Using Bayes' Theorem:

\[P(A|L) = \frac{P(A) \cdot P(L|A)}{P(A) \cdot P(L|A) + P(B) \cdot P(L|B)} \]

Substitute the values:

$$ P(A|L) = \frac{0.80 \cdot 0.05}{0.80 \cdot 0.05 + 0.20 \cdot 0.10} $$

Calculate the numerator and denominator:

$$ P(A|L) = \frac{0.04}{0.04 + 0.02} $$

Simplify:

$$ P(A|L) = 0.04/0.06

Now, calculate the probability:

P(A|L) = {2}{3}

So, the probability that a randomly selected taxi that is late belongs to Company A is $\frac{2}{3}$ or approximately **66.67%**.

Q 6:

Null hypothesis H0=change in blood pressure follows normal distribution

Alternative hypothesis Ha=does not follow normal distribution

```
blood_presure_change=df[' Blood Pressure Before (mmHg)']-df[' Blood Pressure After (mmHg)']
```

```
#Perform Shapiro-Wilk test

statistic, p_value = stats.shapiro(blood_presure_change)

# Print the test statistic and p-value

print("Shapiro-Wilk Test Statistic:", statistic)

print("p-value:", p_value)
```

Shapiro-Wilk Test Statistic: 0.9538205862045288

p-value: 0.0014940275577828288

Since p value <0.05 we reject Null hypothesis and change in blood presure does not follow normal distribution

---

Q 7:

To calculate the variance of Y, the coefficient of determination (R^2), and the standard errors of the estimate of X on Y and Y on X, we can use the equations of the regression lines and some basic statistics formulas. Let's break it down step by step:

Given regression lines:

1. $2X + 3Y - 8 = 0$

2. $2Y + X - 5 = 0$

Also, it's given that the variance of X (Var(X)) is 4.

a. Variance of Y (Var(Y)):

The variance of Y can be calculated using the regression line formula. Since the coefficient of Y in the first equation is 3, we can calculate Var(Y) as follows:

Var(Y) = 1/3 * Var(X)

Var(Y) = 1/3 * 4

Var(Y) = 4/3 =

b. Coefficient of Determination (R^2) between X and Y:

The coefficient of determination (R^2) measures the proportion of the variance in Y that is predictable from X. It can be calculated as the square of the correlation coefficient (r) between X and Y.

The correlation coefficient (r) can be calculated using the coefficients of the regression lines. In this case, it's the coefficient of X in the second equation and the coefficient of Y in the first equation:

r = (-1) * (1/3) = -1/3

Now, calculate R^2:

R^2 = r^2

R^2 = (-1/3)^2

R^2 = 1/9

c. Standard Error of Estimate of X on Y (SE(X|Y)) and Y on X (SE(Y|X)):

The standard error of estimate measures the average distance between observed values and predicted values. For X on Y and Y on X, we can use the following formulas:

- Standard Error of Estimate of X on Y (SE(X|Y)):

SE(X|Y) = sqrt((1 - R^2) * Var(X))

SE(X|Y) = sqrt((1 - 1/9) * 4)

SE(X|Y) = sqrt(8/9)=0.94280

- Standard Error of Estimate of Y on X (SE(Y|X)):

SE(Y|X) = sqrt((1 - R^2) * Var(Y))

SE(Y|X) = sqrt((1 - 1/9) * 4/3)

SE(Y|X) = sqrt(32/27)=1.0886

These are the calculations for the variance of Y, the coefficient of determination (R^2), and the standard errors of the estimate for X on Y and Y on X based on the given regression lines and the variance of X.

Q 8: Wilcoxon signed-rank statistic: 0.0

P-value: 0.001953125

The therapy had a significant effect on anxiety levels.

Q9

```
import scipy.stats as stats



exam1=[85,70,90,75,95]

exam2=[90,80,85,70,92]

exam3=[92,85,88,75,96]

f_statistic,p_value=stats.f_oneway(exam1,exam2,exam3)

alpha=0.05

# Check if the p-value is less than alpha

if p_value < alpha:

    print("Reject the null hypothesis")
```

```
    print("There is enough evidence to conclude that at least one of
the means is different.")

else:

    print("Fail to reject the null hypothesis")

    print("There is not enough evidence to conclude that the means are
different.")



highest_avg_score=0

print("finding student name with highest average score")

names=['Karan','Deppa','Karthik','Chandan','Jeevan']

for name,m1,m2,m3 in zip(names,exam1,exam2,exam3):

  avg_score=(m1+m2+m3)/3

  print(f"Average score for {name} is {avg_score}")

  if avg_score > highest_avg_score:

    highest_avg_score=avg_score

    highest_avg_score_name=name

print(f"Student with highest avg score is {highest_avg_score_name}")
```

OUTPUT:

Fail to reject the null hypothesis

There is not enough evidence to conclude that the means are different.

finding student name with highest average score

Average score for Karan is 89.0

Average score for Deppa is 78.33333333333333

Average score for Karthik is 87.66666666666667

Average score for Chandan is 73.33333333333333

Average score for Jeevan is 94.33333333333333

Student with highest avg score is Jeevan

—----------

Q10 :USing binomial distribution

A: for exactly 20

P(X=20)={500 choose 20)(.05)^20 *(0.95)^480

Since formula is

P(X=k)=(n choose k)p^k *(1-p)^(n-k) wher n=500 ,k=20 and 0.0 so 1-p=0.95 so ans ~=0.0516 i.r5.16%

B:b. Probability that at least 10 bulbs are defective:

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

You can calculate P(X < 10) using the binomial probability formula for various values of x from 0 to 9 and then subtract it from 1.

Output: 99.98316463654902

C:c. Probability that at most 15 bulbs are defective:

P(X <= 15) = sum(P(X = x) for x in range(16))

You can sum the probabilities for x from 0 to 15.

ANS:0.019858377163006223

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

The expected number of defective bulbs (mean) in a binomial distribution is given by:

E(X) = n * p

Here, n = 500 and p = 0.05.

E(X) = 500 * 0.05 = 25

So, on average, you would expect 25 defective bulbs in a batch of 500.

—----------

Q11:

A: One-Way ANOVA p-value: 2.3565868442707578e-08

The distributio of classes are not same

B:

C:from scipy.stats import f_oneway

```
from scipy.stats import levene

import pandas as pd

df=pd.read_csv("data.csv")

df.head()




# Perform Levene's test for equality of variances

levene_statistic, p_value_levene = levene(df[' Blood Pressure Before
(mmHg)'],df[' Blood Pressure After (mmHg)'])



print("Levene's test p-value:", p_value_levene)



if p_value_levene >0.05:

  print("it suggests that the variances are approximately equal between
'Before BP' and 'After BP' data.")

else:

  print(" No equal variance")
```

Levene's test p-value: 0.6715080090945376

it suggests that the variances are approximately equal between 'Before BP' and 'After BP' data.

D:import scipy.stats as stats

```
# Perform one-way ANOVA

f_statistic, p_value = stats.f_oneway(df[' Blood Pressure Before
(mmHg)'],df[' Blood Pressure After (mmHg)'])

print("One-Way ANOVA p-value:", p_value)

if p_value <0.05:

  print("We reject null hypothesis  indicating that there are
significant Zifferences between groups.")

else:

  print("We failed to reject null hypothesis there is significant diff
between groups")
```

One-Way ANOVA p-value: 2.3565868442707578e-08

We reject null hypothesis indicating that there are significant differences between groups.

==========

Q12

n=30

Mean_improvement_score

Group A=2.5 Group B=2.2

Std deviation for Group A=0.8

Std deviation for Group B=0.6

alpha=0.05

Hull hypothesis: there is no significance difference in mean improvement score between group A and group B

Alternate hypothesis : There is significance difference in mean improvement score between group A and group B

We are using 2 sample independent t test

Using formula tcal=1.6431 and ttab from table (df=30+30-2=58 at alpha=0.05)=1.671

We failed to reject null hypothesis hence there is no significant dif in improvement score in group A and group B

```
import pandas as pd

from scipy.stats import ttest_ind

import scipy.stats as stats

# Given data for Group A

mean_A = 2.5

stddev_A = 0.8

n_A = 30

# Given data for Group B

mean_B = 2.2

stddev_B = 0.6

n_B = 30

# Significance level (alpha)

alpha = 0.05

# Calculate the t-statistic

t_statistic, p_value = stats.ttest_ind_from_stats(mean_A, stddev_A,
n_A, mean_B, stddev_B, n_B)

print(t_statistic)

# Degrees of freedom

df = n_A + n_B - 2

# Calculate the critical t-value for a two-tailed test

critical_t = stats.t.ppf(1 - alpha / 2, df)

# Compare the t-statistic with the critical t-value

if abs(t_statistic) > critical_t:
```

```
    print("Reject the null hypothesis: There is a significant
difference in mean improvement scores.")

else:

    print("Fail to reject the null hypothesis: There is no significant
difference in mean improvement scores.")

# Print the p-value

print("p-value:", p_value)
```

OUTPUT :

T_statistic :1.6431676725154976

Fail to reject the null hypothesis: There is no significant difference in mean improvement scores.

p-value: 0.10575916705583671