Case Study 1: Statistical Analysis of Weight Data

Objective: To analyze a dataset of participant weights and perform statistical analysis to draw conclusions about the population.
Scenario: A fitness trainer has collected weight data from participants in their weight loss program. The objective is to perform a statistical analysis on the dataset to gain insights, given that the average weight of participants in their weight loss program is 75 kg from historical data.
Dataset: You are provided with a dataset named "weight_data.csv. Download weight_data.csv." The dataset contains weight information for participants in the weight loss program.

As mentioned in the question, the average weight of participants in the past is 75kg.

Tasks:

1. **Explore the dataset and identify key statistical measures.**
   The key statistical measures are as follows.
   - Mean -> This will give the idea of central tendency of the dataset.
   - Median-> This will give the middle value of the dataset when it's sorted in sequence. It's useful for understanding the distribution, especially if there are outliers.
   - Mode-> To check the number frequently occurring weight. It also helps to identify dominant weight category.
   - Variance-> To measure how far each weight in the dataset is from the mean. A higher variance indicates greater variability.
   - Standard deviation-> It gives a more interpretable measure of the spread of the data.
   - Histogram-> To visualize the distribution of weights. This can help identify patterns and potential outliers.
   - Confidence interval-> Calculate the 95% confidence interval for the mean weight. This provides a range within which we can be reasonably confident the true population mean weight lies.

2. **Formulate hypotheses about the population based on the available information.**

   In this scenario, I formulate hypotheses to test whether the average weight of participants in the current weight loss program is significantly different from the historical average of 75 kg. The null hypothesis (H_0) states that there is no significant difference, and the alternative hypothesis (H_1) suggests the direction of the difference.

   1. Null Hypothesis (H_0): The average weight of participants in the current weight loss program is equal to the historical average (μ=75kg). $H_0 : \mu=75kg$
   2. Alternative Hypothesis (H_1): The average weight of participants in the current weight loss program is not equal to the historical average. H1: $\mu \neq 75Kg$

3. **Perform appropriate statistical tests to test the hypotheses.**

Computing the necessary data in the excel. The obtained data are presented below:
Sample mean (x) = 74.08
Variance (σ^2) = 2.238367
Standard deviation (σ) = 1.496117
Significance level (α) = 5% or 0.05
n = 50

Performing Z- Test as the sample size is 50. Calculations are shown in R file.

$$Z = \frac{x - \mu_0}{\sigma / n^{1/2}}$$

$$Z = \frac{74.08 - 75}{1.496117/50^{1/2}}$$

$$Z = -4.348195$$

So, the *z*-score is approximately -4.35

I am using a two-tailed test because I am interested in whether the average weight is different, not just whether it's greater or less than

P – Value = $2[\,1 - \varphi|\,(-4.35)\,|\,]$
= 1.373×10^{-5}

4. **Draw meaningful conclusions and recommendations for the fitness trainer.**

Conclusions:

Statistical Significance: The Z-test yielded a highly statistically significant result with a very low p-value (1.373×10^{-5}). This indicates strong case against the null hypothesis that the true mean weight of participants is equal to 75 kg. Hence, we can reject the null hypothesis.

Sample Mean: The observed mean weight of participants in the weight loss program is 74.08 kg. This mean is lower than the hypothesized mean of 75 kg.

Confidence Interval: The 95% confidence interval for the mean weight is (73.66531, 74.49469). This interval does not include the hypothesized mean of 75 kg, which concludes that the true mean is likely different.

Recommendations:

Target Common Weights: Continue to focus on participants with weights 75kg and 76 kg as 10 individuals lie in the category as we can see from the mode frequency table below and only one individual lie in 77 kg category. We can say that the one individual in 77 kg category as an outlier. Total 11 individuals are above 75kg and trainer can focus on them more.

Weights	Frequency
71	1
72	8
73	10
74	10
75	10
76	10
77	1

Individualized Plans: Understand that participants have different weights, including those outside the common range. Think about offering personalized exercise and nutrition plans to meet the specific needs of participants with various starting points.

Case Study 2: Analysis of Proportions in a Survey

Objective: To analyze a dataset from a customer satisfaction survey and perform statistical analysis to draw conclusions about the proportion of satisfied customers.
Scenario: A company conducted a customer satisfaction survey to assess the satisfaction level among its customers. The objective is to analyze the survey data to estimate the proportion of satisfied customers and draw meaningful conclusions. Your business has a historical average satisfaction rate of 75%.

Tasks:
Task 1: Explore the dataset Download dataset

The total sample size is 1000.

The sample proportion for the total satisfied customer is,
Number of satisfied customers = 775
P_0 = 75% = 0.75
P^ = Number of satisfied customers / Total sample size
= 775 / 1000
= 0.775

Assess the variability in the dataset.
P^ = 0.775
N = 1000

$$S.D, of\ P = \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$S.D, of\ P = \sqrt{\frac{0.75(1-0.75)}{1000}}$$

$$S.D, of\ P = 0.0136$$

Task 2: Formulate hypotheses

- Null Hypothesis (H_0): The proportion of satisfied customers = 75%.
  $H_0 : \mu=75\%$
- Alternative Hypothesis (H_1): The proportion of satisfied customers is different from the specified value and is not = 75%
  $H_1: \mu \neq 75\%$

Task 3: Perform statistical tests

- Choosing Z-test for proportions since we are comparing a sample proportion to a specified value.
- **Calculate Test Statistic:** for proportions.

$$z = \frac{p^\wedge - p_0}{S.D.}$$

$$z = \frac{0.775 - 0.75}{0.0136}$$

$$z = 1.838$$

P-value = 2*[1 – φ (1.838)]
= 2*0.0322
= 0.0678

Task 4: Draw conclusions

- Here, p-value is > 0.05. So, Null hypothesis is not rejected here.
- The p-value is greater than 0.05, there is not enough evidence to suggest a significant difference in the proportion of satisfied customers.

Additional Insights:

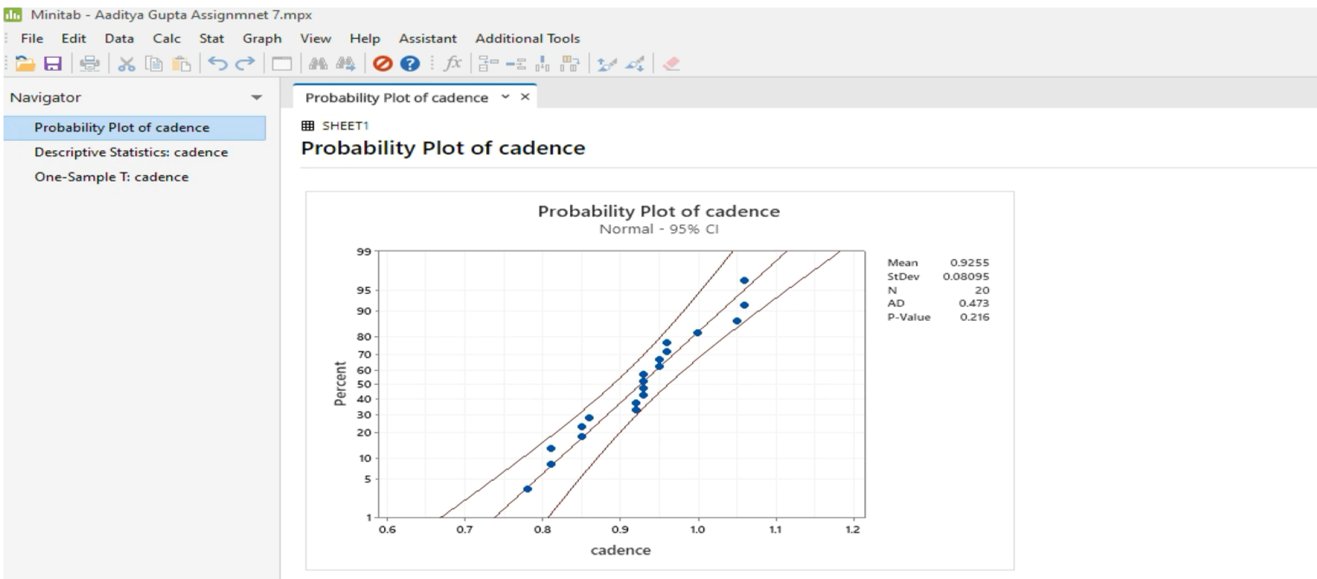
- Analyze customer feedback to identify specific areas that contribute to satisfaction or dissatisfaction.
- Look for patterns or common themes in customer comments to gain qualitative insights.

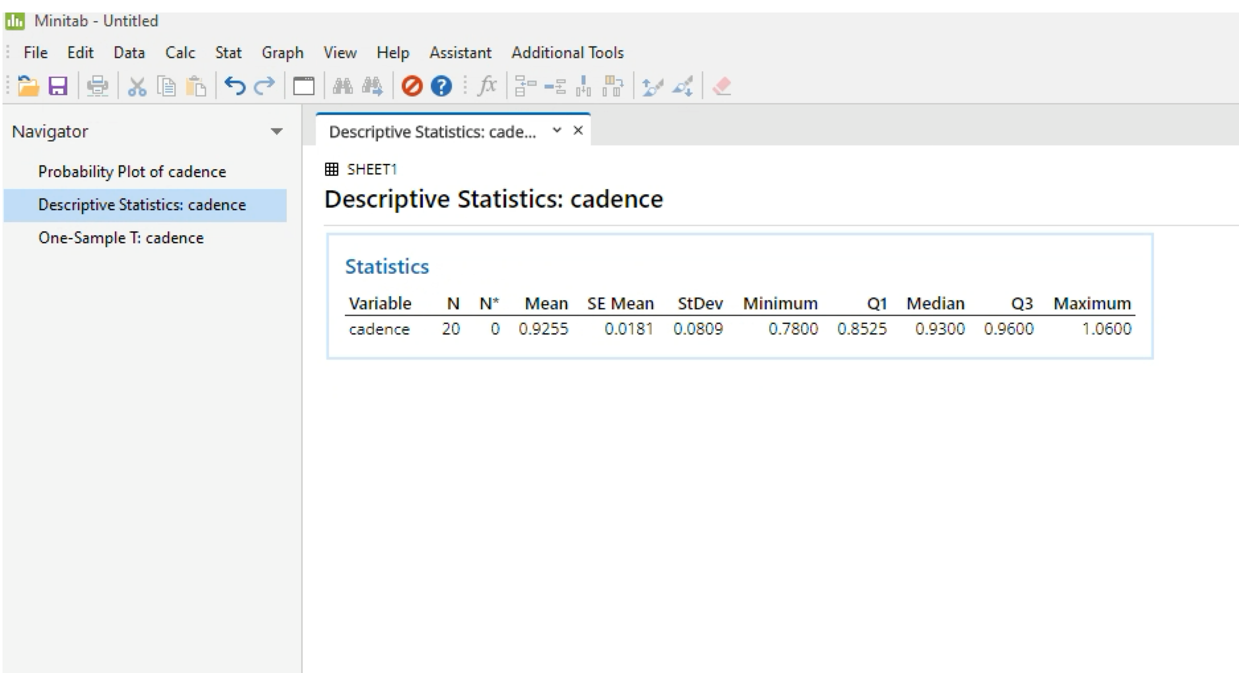
Recommendations:

- Establish a continuous monitoring system for customer satisfaction.
- Regularly collect and analyze the feedback for changing customer trend.

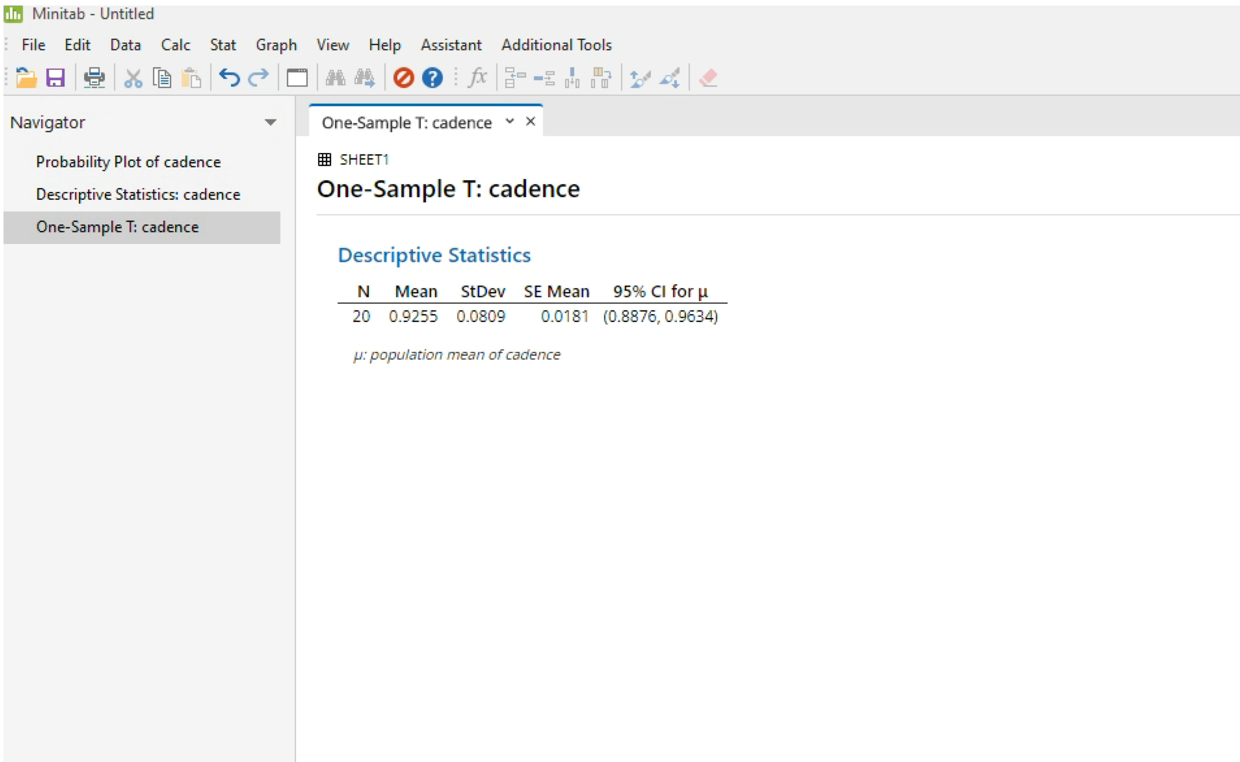
Question 3: For this [articleLinks to an external site.]() , the attached dataset is [here Download here]().

1. Check the normality of the data (normal probability plot) and obtain a descriptive summary of the data using Minitab.





2. Calculate and interpret a 95% confidence interval for population mean cadence. (Do by hand calculations and by Minitab)



95% confidence interval for population mean cadence is,

$$= \bar{x} \pm t\left(\frac{s}{n^{\frac{1}{2}}}\right)$$

x̄ = Sample mean = 0.9255
S = Standard deviation = 0.0809
n = Sample size = 20
α = 0.5/2 = 0.025
t = for α/2. t-value from the t-distribution with n-1 degrees of freedom = 2.093

$$= 0.9255 \pm 2.093\left(\frac{0.0809}{20^{\frac{1}{2}}}\right)$$

CI = (0.8846, 0.9603)

3. Calculate and interpret a 95% prediction interval for the cadence of a single individual randomly selected from this population.

95% prediction interval for the cadence of a single individual,

$$= \bar{x} \pm t * (n-1) * s\sqrt{1 + 1/n}$$

$\bar{x}$ = Sample mean = 0.9255
S = Standard deviation = 0.0809
n = Sample size = 20
α = 0.5/2 = 0.025
t = for α/2. t-value from the t-distribution with n-1 degrees of freedom = 2.093

$$= 0.9255 \pm 2.093 * (20-1) * 0.0809\sqrt{1 + 1/20}$$

$$\text{PI} = (0.7518, 1.0991)$$

We are 95% confident that the cadence of the next individual will fall in the range PI shown above.

4. Calculate an interval that includes at least 99% of the cadences in the population distribution using a confidence level of 95%.

Assuming two-sided interval, we use tolerance interval concept.

$\boldsymbol{TI} = \bar{x} \pm (\text{tolerance critical value}) * s$

Tolerance Critical value = 3.615

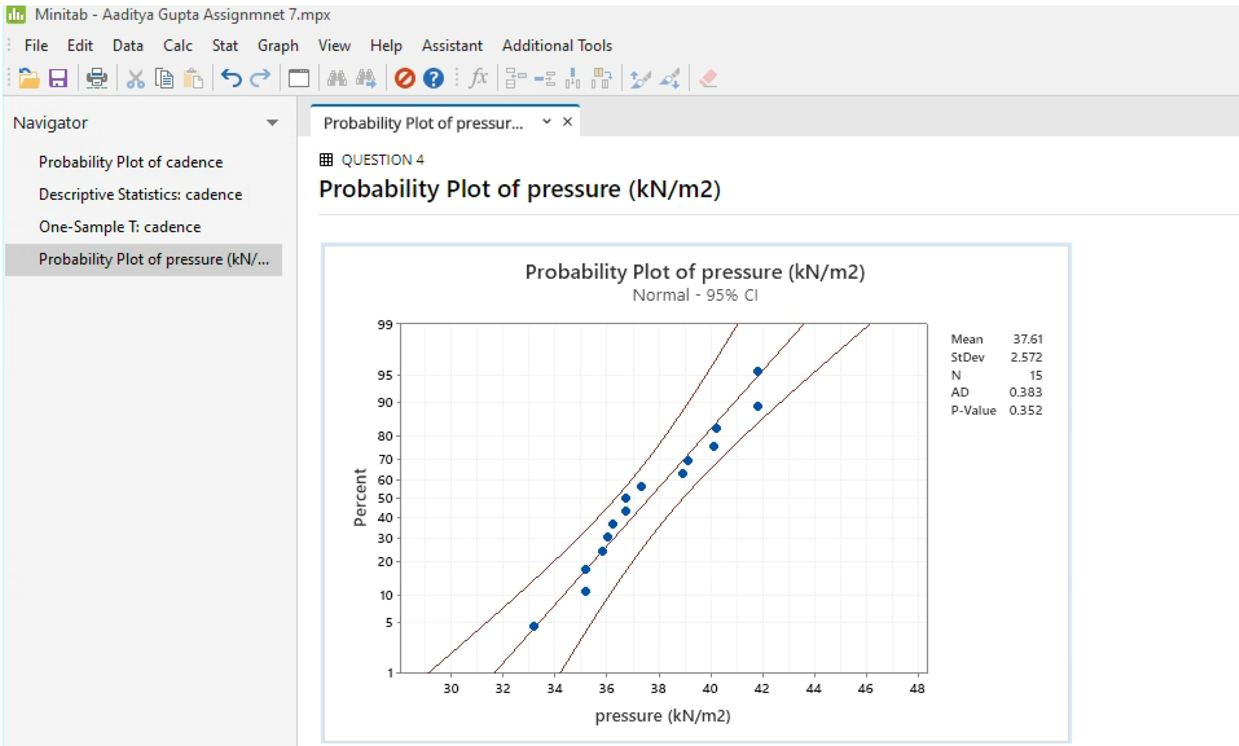
$$\boldsymbol{TI} = 0.9\bar{2}55 \pm (3.615) * 0.0809$$

$$\boldsymbol{TI} = (0.6328, 1.2181)$$

Upper TI = 1.2181, Lower TI = 0.6328

Question 4: From this [articleLinks to an external site.], we got this [data Download data].

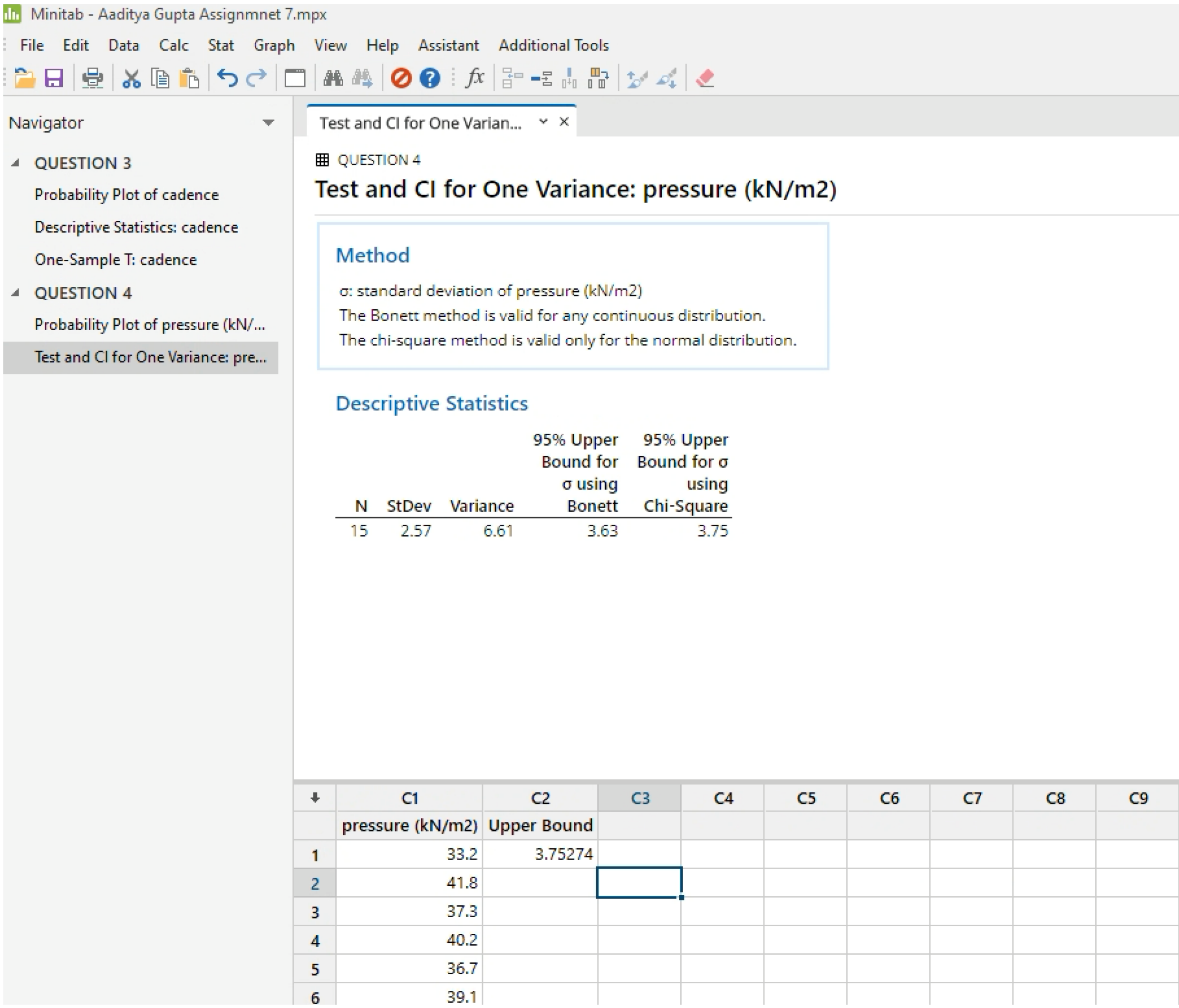
1. Is it plausible that this sample was selected from a normal population distribution? (Check normal probability plot using Minitab)



In above result, the null hypothesis states that the data follow a normal distribution. Because the p-value is 0.352, which is greater than the significance level of 0.05, the decision is to fail to reject the null hypothesis. I cannot conclude that the data do not follow a normal population distribution.

Also, we can see the graph, most of the points are close to the straight line and its roughly linear. So, I think, its plausible to say that the sample was selected from a normal population distribution.

2. Calculate an upper confidence bound with a confidence level 95% for the population standard deviation of maximum pressure. (Do by hand calculations and by Minitab)



Upper Confidence Bound Calculations:

$$UCB = \sqrt{\frac{(n-1)s^2}{Critical\ value}}$$

Where, n = Sample size = 15

s^2= Variance = 6.61

Degree of freedom = n-1 = 15-1 = 14

Level of significance = 1- α = 1-0.05 = 0.95

Critical value for upper bound from the table = 6.571

$$UCB = \sqrt{\frac{(15-1)6.61}{6.571}}$$

Therefore, the UCB is found to be= 3.752