|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Integer |
| Results of rolling a dice | Discrete Integer |
| Weight of a person | Continuous |
| Weight of Gold | Continuous |
| Distance between two places | Continuous |
| Length of a leaf | Continuous |
| Dog's weight | Continuous |
| Blue Color | Categorical |
| Number of kids | Integer |
| Number of tickets in Indian railways | Integer |
| Number of times married | Integer |
| Gender (Male or Female) | Categorical |

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ratio |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

A3)Total outcomes of tossing a coin = 2

Total outcomes of tossing 3 Coins = 8 , { 2\*2\*2 }

Outcomes of 2 Heads and 1 Tail when we toss 3 coins = 3 , { HHT , HTH , THH }

Probability of 2 Heads and 1 Tail when we toss 3 coins = 3/8

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

A4) Number of outcomes by rolling 2 dice = 36

a) Probability of sum being equal to 1 = 0 { because the lowest sum of two outcome after rolling two dice is 2 which is (1,1) }

b) Probability of sum less than or equal to 4 = 6/36 { (1,1) , (1,2) , (1,3) , (2,1) , (2,2) , (3,1) }

c) Probability of sum is divisible by 2 and 3 = 6/36 { (1,5) , (2,4) , (3,3) , (4,2) , (5,1) , (6,6) }

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

A5) Probability of getting a blue ball is = 2/7 { because total no of balls are 7 and there are 2 blue balls }

Probability of not getting a blue ball will be = 1 – 2/7 = 5/7

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

A6) Expected Number of Candies for a Randomly Selected Child = 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120 = 3.135

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**A7)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.558394 | 3.070189 | 17.76311 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Standard Deviation | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.17 | 3.911 | 8.4 |
|  |  |  |  |
|  |  |  |  |
| \* For Points and Score, the mean, median, and mode provide insights into the skewness of the distributions. | | | |
| \* Weigh appears to have a more symmetric distribution based on mean, median, and mode. | | | |
| \* The variability in Weigh, as indicated by variance and standard deviation, is higher compared to Points but lower than Score. | | | |

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

A8) Sum of Given Weights = 108+110+123+134+135+145+167+187+199 = 1308

Number of Given Weights = 9

Expected Value = 1308/9 = 145.33

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

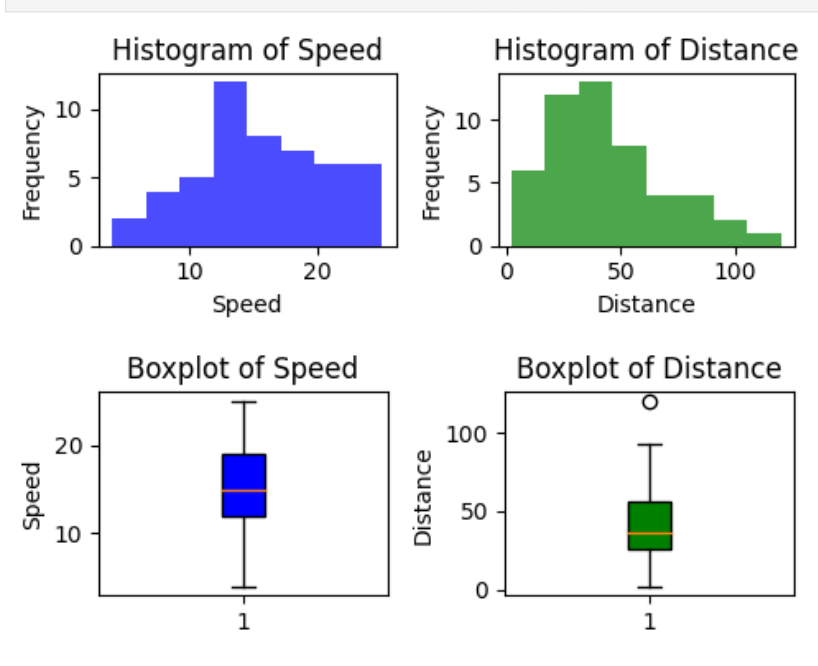
**Use Q9\_a.csv**

**A9 a) Results**

* **Speed:**
  + **Skewness: −0.114-0.114−0.114 (slightly negative, nearly symmetric distribution).**
  + **Kurtosis: −0.577-0.577−0.577 (platykurtic, flatter than a normal distribution).**
* **Distance:**
  + **Skewness: 0.7820.7820.782 (positive skew, tail extends to the right).**
  + **Kurtosis: 0.2480.2480.248 (close to normal, slightly sharper than a normal distribution).**

**Conclusion**

* **The speed data shows a balanced distribution, implying consistent measurements.**
* **The distance data has a moderate right skew, indicating occasional higher stopping distances, likely due to outliers or unique scenarios.**

****

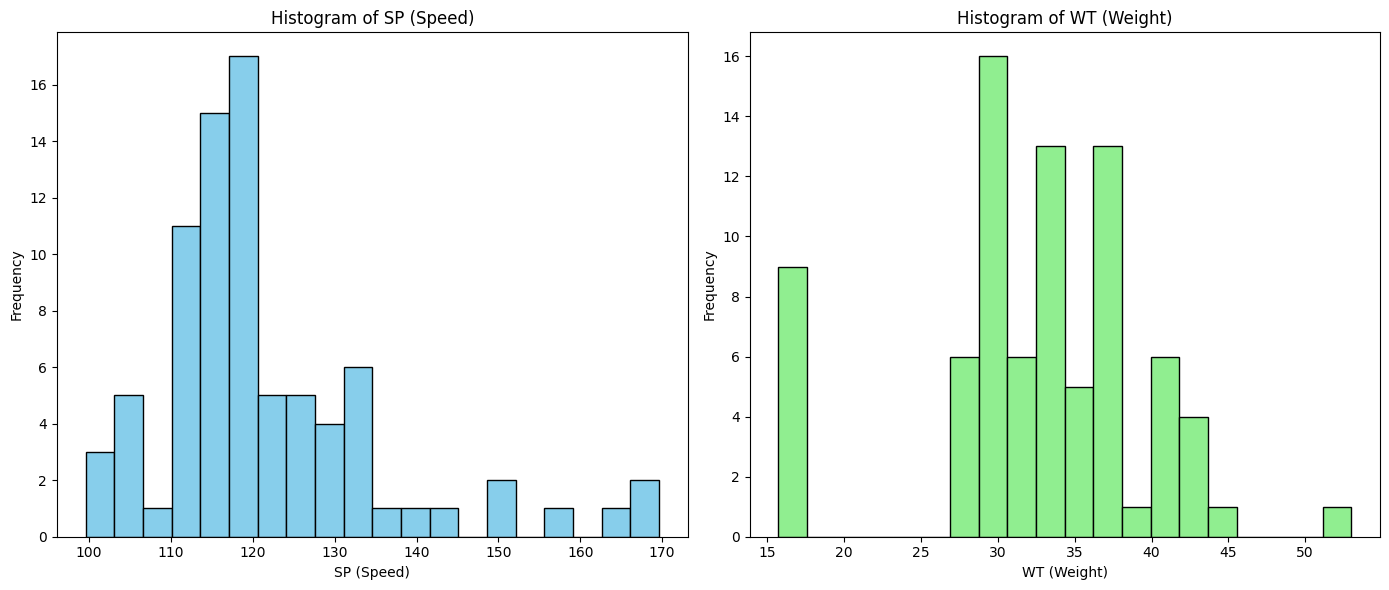
**SP and Weight(WT)**

**Use Q9\_b.csv**

**A9 b) Skewness of SP: 1.5814536794423764**

**Kurtosis of SP: 2.7235214865269244**

**Skewness of WT: -0.6033099322115126**

**Kurtosis of WT: 0.8194658792266849**

**Speed (SP):**

* **The right-skewed distribution with a kurtosis less than 3 suggests that most of the vehicles are traveling at lower speeds, but there are occasional instances of vehicles traveling at very high speeds. However, extreme high-speed values are relatively rare.**

**Weight (WT):**

* **The left-skewed distribution with a kurtosis less than 3 suggests that most entities are heavier, with fewer light-weight entities. The distribution doesn't exhibit significant outliers, indicating that extreme weights are uncommon.**

**In summary,**

**Speed (SP) is characterized by a right-skewed distribution with relatively few extreme high-speed observations, and Weight (WT) is left-skewed with most data clustered around the heavier weights and fewer instances of extreme light weights.**

**Q10) Draw inferences about the following boxplot & histogram**



A 10) **Histogram Inferences:**

1. **Shape of Distribution:**
   * The histogram appears to have a **right-skewed distribution** since the majority of the values are concentrated on the left side (around 0 to 100) and there is a longer tail extending towards the right side.
   * This suggests that most observations (likely weights) are clustered around the lower end, with a few observations having higher values, but the higher values are sparse.
2. **Peak:**
   * The **highest bars** are concentrated between 50 and 100. This indicates that the most frequent values are within this range, suggesting that the majority of the chicks weigh between 50 and 100 units.
3. **Spread:**
   * The distribution is fairly **dispersed**, with fewer observations at the higher weight ranges (above 100), particularly between 200 and 400.
4. **Outliers:**
   * The histogram does not suggest many **extreme outliers** at higher weights. It shows that the number of chicks with weight values above 200 is **relatively low**.

**Boxplot Inferences:**

1. **Central Tendency (Median):**
   * The median (middle line of the box) seems to be around **100**, indicating that the typical chick's weight is in the range of 50 to 100.
2. **Spread (Interquartile Range - IQR):**
   * The **box** spans from roughly 50 to 150. This shows that 50% of the chicks' weights are between 50 and 150 units.
   * The width of the box indicates the **spread of weights** in the central half of the dataset. Since the box is relatively wide, it shows that there is considerable variability in the chick weights within this range.
3. **Whiskers:**
   * The **lower whisker** extends from the box down to about 10-20, indicating that there are some chicks with weights as low as this.
   * The **upper whisker** extends up to around 200, but there are no significant points beyond this, suggesting that there are very few chicks with very high weights.
4. **Outliers:**
   * There are no clear **outliers** in the boxplot, as no data points are plotted beyond the whiskers, which typically indicate outliers.

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

A11) To calculate the confidence intervals for the average weight of adult males in Mexico, we use the following formula for a confidence interval (CI) for the population mean when the sample size is large and the population standard deviation is unknown:

CI = x̄ ± z \* (s / √n)

Where:

* x̄ = sample mean = 200 pounds
* s = sample standard deviation = 30 pounds
* n = sample size = 2,000
* z = z-score corresponding to the desired confidence level
* (s / √n) = standard error of the mean

**Step 1: Find the Z-scores for the given confidence levels**

* For 94% confidence level, the Z-score is approximately 1.88.
* For 98% confidence level, the Z-score is approximately 2.33.
* For 96% confidence level, the Z-score is approximately 2.05.

**Step 2: Calculate the Standard Error (SE)**

The standard error of the mean is given by:

SE = s / √n = 30 / √2000 ≈ 30 / 44.72 ≈ 0.671

**Step 3: Calculate the Confidence Intervals**

**94% Confidence Interval:**

CI₉₄% = 200 ± 1.88 \* 0.671 CI₉₄% = 200 ± 1.26 CI₉₄% = (198.74, 201.26)

**98% Confidence Interval:**

CI₉₈% = 200 ± 2.33 \* 0.671 CI₉₈% = 200 ± 1.56 CI₉₈% = (198.44, 201.56)

**96% Confidence Interval:**

CI₉₆% = 200 ± 2.05 \* 0.671 CI₉₆% = 200 ± 1.37 CI₉₆% = (198.63, 201.37)

**Final Results:**

* 94% Confidence Interval: (198.74, 201.26)
* 98% Confidence Interval: (198.44, 201.56)
* 96% Confidence Interval: (198.63, 201.37)

These intervals provide a range of plausible values for the average weight of adult males in Mexico at different confidence levels.

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

A12) **Given Scores:**

34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

**1. Mean:**

The mean of the scores is **42.5**.

**2. Median:**

The median of the scores is **40.5**.

**3. Variance:**

The variance of the scores is **16.69**.

**4. Standard Deviation:**

The standard deviation of the scores is **4.08**.

**2. Inferences about the Student's Marks:**

* **Mean (42.5):** The student's average score is 42.5, indicating that, on average, the student is performing reasonably well.
* **Median (40.5):** The median is close to the mean, suggesting that the distribution of scores is fairly symmetrical.
* **Variance (16.69) and Standard Deviation (4.08):** The variance and standard deviation indicate that the scores are moderately spread out around the mean, with a standard deviation of 4.08. This shows that while most scores are near the average, there is some variability in the student's performance.
* **Range of Scores (34 to 56):** The scores range from 34 to 56, suggesting that there are a few lower and higher scores, but most scores lie within the middle range of 36 to 45.

In summary, the student is performing well on average, with some fluctuations in the scores. Most of the scores are around the middle range, and the student has a few lower and higher marks, indicating that there is room for improvement but also areas where the student excels.

Q13) What is the nature of skewness when mean, median of data are equal?

A13) When the **mean** and **median** of a data set are equal, the distribution is said to be **symmetrical**. In this case, the **skewness** of the data is **zero**.

This indicates that the data is evenly distributed around the central point, and there is no tendency for the data to lean more towards the left (negative skew) or right (positive skew).

To summarize:

* **Equal Mean and Median**: Symmetrical distribution
* **Skewness**: Zero (no skew)

Q14) What is the nature of skewness when mean > median ?

A14) When the **mean** is greater than the **median**, the distribution is said to be **positively skewed** or **right-skewed**.

In this case, the tail of the distribution is stretched towards the right, meaning there are relatively few larger values that pull the mean to the right of the median. The skewness value for this type of distribution will be **positive**.

To summarize:

* **Mean > Median**: Right-skewed (positive skew)
* **Skewness**: Positive (right tail longer)

Q15) What is the nature of skewness when median > mean?

A15) When the **median** is greater than the **mean**, the distribution is said to be **negatively skewed** or **left-skewed**.

In this case, the tail of the distribution is stretched towards the left, meaning there are relatively few smaller values that pull the mean to the left of the median. The skewness value for this type of distribution will be **negative**.

To summarize:

* **Median > Mean**: Left-skewed (negative skew)
* **Skewness**: Negative (left tail longer)

Q16) What does positive kurtosis value indicates for a data ?

A16 ) A **positive kurtosis value** indicates that the data has a **leptokurtic distribution**, which means the distribution has **heavier tails** and a **sharper peak** than a normal distribution. In simpler terms:

* The data has more frequent extreme values (outliers) compared to a normal distribution.
* The data is more concentrated around the mean, leading to a higher peak in the center of the distribution.

In statistical terms, if the kurtosis is greater than 3 (the kurtosis of a normal distribution), it is considered **positive kurtosis** or **leptokurtic**.

To summarize:

* **Positive kurtosis**: Leptokurtic (sharper peak, heavier tails)
* Indicates a higher probability of extreme values or outliers in the dataset.

Q17) What does negative kurtosis value indicates for a data?

A17 ) A **negative kurtosis value** indicates that the data has a **platykurtic distribution**, which means the distribution has **lighter tails** and a **flatter peak** compared to a normal distribution. In simpler terms:

* The data has fewer extreme values (outliers) compared to a normal distribution.
* The data is more evenly spread out, leading to a lower and broader peak in the center of the distribution.

In statistical terms, if the kurtosis is less than 3 (the kurtosis of a normal distribution), it is considered **negative kurtosis** or **platykurtic**.

To summarize:

* **Negative kurtosis**: Platykurtic (flatter peak, lighter tails)
* Indicates fewer extreme values or outliers in the dataset.

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

What is nature of skewness of the data?

What will be the IQR of the data (approximately)?   
  
A18)

1. **Distribution of the Data:**
   * The data appears to be fairly symmetric, as the median is approximately in the center of the box, and the whiskers on both sides are of similar lengths.
2. **Nature of Skewness:**
   * The data does not show significant skewness. It is likely **approximately symmetric**.
3. **Interquartile Range (IQR):**
   * IQR = Q3 - Q1.
   * Based on the boxplot, the approximate values are:
     + Q1 (25th percentile): ~12.
     + Q3 (75th percentile): ~16.
   * IQR ≈ 16 - 12 = **4**.

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

A19) **Observations from the Boxplots:**

1. **Boxplot 1 :**
   * The range of the data (difference between max and min) is small, indicating lower variability.
   * The data is tightly clustered around the median.
   * The interquartile range (IQR) is relatively narrow compared to Boxplot 2, suggesting less spread among the middle 50% of the data.
2. **Boxplot 2 :**
   * The range is significantly larger than Boxplot 1, indicating higher variability.
   * The IQR is wider, showing that the middle 50% of the data points are more dispersed.
   * The whiskers extend farther, representing a wider spread of data.

**Inference:**

* **Comparison of Variability:** The data in Boxplot 2 exhibits more variability and spread than the data in Boxplot 1.
* **Clustering:** Boxplot 1 has tightly clustered data, suggesting consistency or uniformity, while Boxplot 2's data is more spread out, suggesting greater diversity or heterogeneity.
* **Possible Implication:** If the boxplots represent two different groups or scenarios, the first group (Boxplot 1) is more consistent, while the second group (Boxplot 2) shows greater diversity in observations.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

A20) The calculated probabilities are:

* P(MPG > 38) = 33/81 = 0.407
* P(MPG < 40) = 61/81 = 0.753
* P(20 < MPG < 50) = 69/81 = 0.851

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

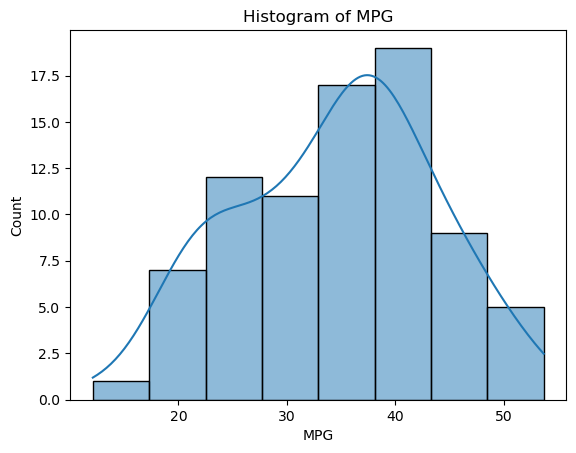
Dataset: Cars.csv  
A21)

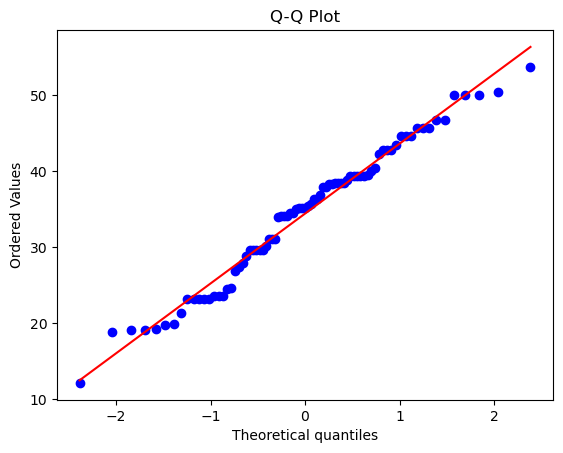
Shapiro-Wilk Test:

Statistic=0.9779686331748962, p-value=0.17639249563217163

Kolmogorov-Smirnov Test:

Statistic=0.09885051329529326, p-value=0.3822048513500623





**Conclusion:**

1. **Shapiro-Wilk Test**:
   * Test Statistic = 0.9779, p-value = 0.1764.
   * Since the p-value **(0.1764)** is greater than 0.05, we **fail to reject the null hypothesis**. This indicates that the **MPG** data does not significantly deviate from a normal distribution.
2. **Kolmogorov-Smirnov Test**:
   * Test Statistic = 0.0989, p-value = 0.3822.
   * Similarly, the p-value **(0.3822)** is greater than 0.05, so we **fail to reject the null hypothesis**. This suggests that the data follows a normal distribution.

**Final Conclusion:**

Based on the results of both the Shapiro-Wilk test and the Kolmogorov-Smirnov test, we conclude that the **MPG of cars follows a normal distribution**.

1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

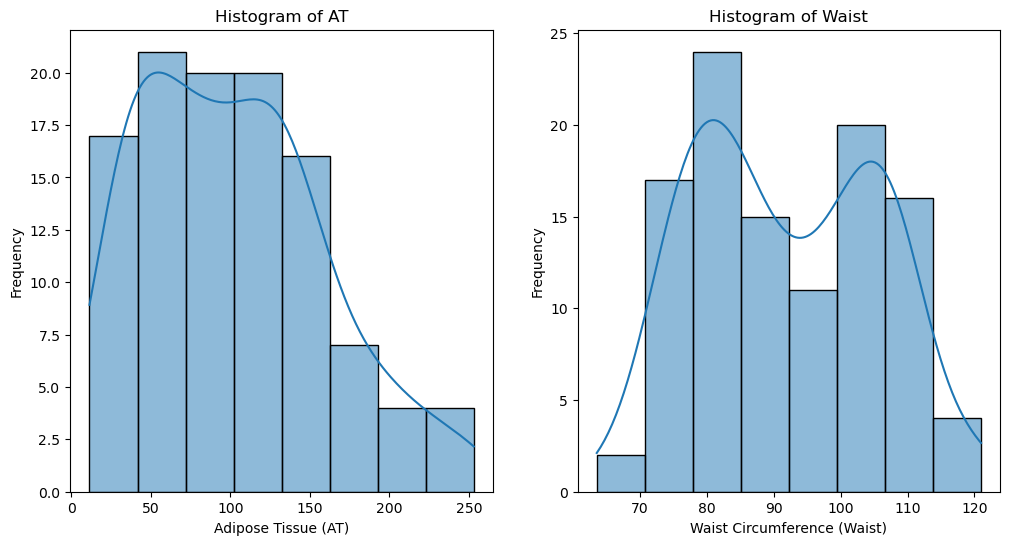
A21 b)

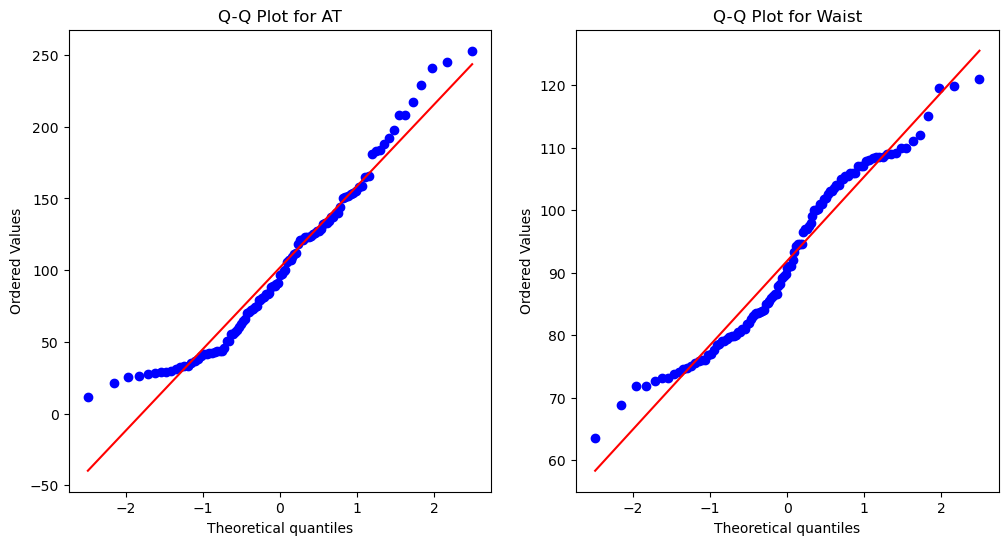
Shapiro-Wilk Test for AT: Statistic=0.9523, p-value=6.5400e-04

Shapiro-Wilk Test for Waist: Statistic=0.9559, p-value=1.1705e-03

Kolmogorov-Smirnov Test for AT: Statistic=0.0746, p-value=5.5390e-01

Kolmogorov-Smirnov Test for Waist: Statistic=0.1065, p-value=1.5661e-01





**Conclusion:**

1. **Shapiro-Wilk Test**:
   * For **AT (Adipose Tissue)**:
     + Test Statistic = 0.9523, p-value = 0.000654.
     + Since the p-value **(0.000654)** is less than 0.05, we **reject the null hypothesis**. This indicates that the **AT data does not follow a normal distribution**.
   * For **Waist Circumference**:
     + Test Statistic = 0.9559, p-value = 0.0011705.
     + Since the p-value **(0.0011705)** is less than 0.05, we **reject the null hypothesis**. This indicates that the **Waist data does not follow a normal distribution**.
2. **Kolmogorov-Smirnov Test**:
   * For **AT (Adipose Tissue)**:
     + Test Statistic = 0.0746, p-value = 0.5539.
     + Since the p-value **(0.5539)** is greater than 0.05, we **fail to reject the null hypothesis**. This suggests that the **AT data follows a normal distribution** under this test.
   * For **Waist Circumference**:
     + Test Statistic = 0.1065, p-value = 0.1566.
     + Since the p-value **(0.1566)** is greater than 0.05, we **fail to reject the null hypothesis**. This suggests that the **Waist data follows a normal distribution** under this test.

**Final Conclusion:**

* The **Shapiro-Wilk test** indicates that neither **AT** nor **Waist** follows a normal distribution.
* The **Kolmogorov-Smirnov test** suggests that both **AT** and **Waist** follow a normal distribution.
* Since the Shapiro-Wilk test is generally considered more robust for small sample sizes, we conclude that **AT and Waist data do not follow a normal distribution**.

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

A22) For a **confidence interval**, the Z-score is based on the area under the standard normal distribution curve. The formula to calculate the Z-score is:

Z = invNorm(1 - α/2)

where:

* α is the significance level, which is 1 - confidence level.
* For example, for a 90 percent confidence interval, α = 1 - 0.90 = 0.10.

Now, let's calculate the Z-scores for each confidence level:

1. **90 percent Confidence Interval**

* α = 1 - 0.90 = 0.10
* The Z-score is at the point corresponding to 1 - 0.10 / 2 = 0.95.
* Z-score = 1.645

1. **94 percent Confidence Interval**

* α = 1 - 0.94 = 0.06
* The Z-score is at the point corresponding to 1 - 0.06 / 2 = 0.97.
* Z-score = 1.881

1. **60 percent Confidence Interval**

* α = 1 - 0.60 = 0.40
* The Z-score is at the point corresponding to 1 - 0.40 / 2 = 0.80.
* Z-score = 1.282

So the Z-scores are:

* **90 percent Confidence Interval: 1.645**
* **94 percent Confidence Interval: 1.881**
* **60 percent Confidence Interval: 1.282**

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

A23) To calculate the t-scores (also known as critical values of the t-distribution) for different confidence intervals, you can use the formula:

**t = invT(1 - α/2, degrees of freedom)**

where:

* α is the significance level, which is 1 - confidence level.
* Degrees of freedom = sample size - 1.

For a sample size of 25, the degrees of freedom will be:

**Degrees of freedom = 25 - 1 = 24.**

Now, let's calculate the t-scores for each confidence level:

**1. 95% Confidence Interval:**

* α = 1 - 0.95 = 0.05
* Degrees of freedom = 24
* The t-score is at the point corresponding to 1 - 0.05 / 2 = 0.975.
* Using a t-table or statistical calculator, the **t-score** for a 95% confidence interval and 24 degrees of freedom is **2.064**.

**2. 96% Confidence Interval:**

* α = 1 - 0.96 = 0.04
* Degrees of freedom = 24
* The t-score is at the point corresponding to 1 - 0.04 / 2 = 0.98.
* Using a t-table or statistical calculator, the **t-score** for a 96% confidence interval and 24 degrees of freedom is **2.171**.

**3. 99% Confidence Interval:**

* α = 1 - 0.99 = 0.01
* Degrees of freedom = 24
* The t-score is at the point corresponding to 1 - 0.01 / 2 = 0.995.
* Using a t-table or statistical calculator, the **t-score** for a 99% confidence interval and 24 degrees of freedom is **2.797**.

**Summary:**

* **95% Confidence Interval t-score: 2.064**
* **96% Confidence Interval t-score: 2.171**
* **99% Confidence Interval t-score: 2.797**

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

A24) **Step 1: Set up the parameters**

* Population mean (claimed by the CEO), μ = 270 days
* Sample mean, x̄ = 260 days
* Sample standard deviation, s = 90 days
* Sample size, n = 18
* Degrees of freedom, df = n - 1 = 18 - 1 = 17

**Step 2: Calculate the t-score**

The formula for the t-score is:

t = (x̄ - μ) / (s / √n)

Substitute the known values:

t = (260 - 270) / (90 / √18) = -10 / (90 / 4.2426) = -10 / 21.2132 = -0.471

**Step 3: Find the probability using the t-distribution**

Now, to find the probability that the sample mean is less than or equal to 260 days, we use the cumulative distribution function (CDF) for the t-distribution:

P(X ≤ 260) = pt(t, df)

The result is approximately:

P(X ≤ 260) ≈ 0.318

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days, assuming the CEO's claim is true, is approximately **0.318** or **31.8%**.