Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| 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 | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ (Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

Sol: Total events are: HHH THH

HHT THT

HTH TTH

HTT TTT

Two Heads and One Tail occurs 3 out of 8 events.

So, Probability is 3/8. i.e. 37.5%

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

Sol: a. Least number on die is 1. When two dice rolled, sum is 2. So, the probability of sum equal to 1 is “Zero”.

b. Total events are: 6\*6 = 36

(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)

Probability is 6/36 = 1/6. i.e. 16.66%

c. To divisible by 2 and 3, the sum must be 6 and 12.

Possibilities are (1,5), (2,4), (3,3), (4,2), (5,1), (6,6).

Probability is 6/36 = 1/6. i.e. 16.66%

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?

Sol: Total balls are 7 and

possible ways of drawing two balls: 7C2 = (7!) / (5! \* 2!) = 7\*6 / 2\*1 = 21

None of the balls drawn is BLUE. So, two balls need to be drawn from 2 RED and 3 GREEN.

Total 5 balls and 2 balls can be drawn in 5C2 ways. (5!) / (3! \* 2!) = 5\*4 / 2\*1 = 10.

Probability that none of the balls drawn is blue: 10/21.

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

Sol: Expected number of candies for a randomly selected child is

= (1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.001) + (2\*0.12)

= 3.09

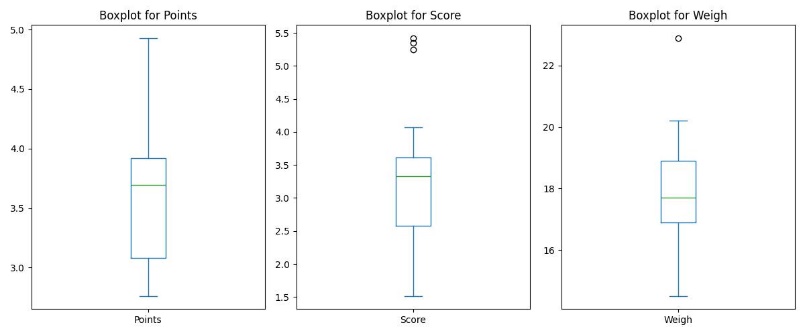
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

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

Sol:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.276948 | 0.927461 | 3.09338 |
| SD | 0.526258 | 0.963048 | 1.758801 |
| Max | 4.93 | 5.424 | 22.9 |
| Min | 2.76 | 1.513 | 14.5 |
| Range | 2.17 | 3.911 | 8.4 |



Points: Mean and Median are relatively close, suggesting a symmetric distribution.

Low variance and low SD indicate that the values are clustered around the mean

No outliers present in the data

Score: The range is wider compared to Points and Weigh, indicating a greater spread of values.

The variance and standard deviation are higher compared to the other columns, signifying more dispersion. 3 outliers present in the data.

Weight: measures indicate a moderate spread, with the mean and medium very close, suggesting

a balanced distribution.

The variance and standard deviation are moderate, indicating moderate variability in the data.

In summary, the "Score" column has the widest range and highest variability among the three columns, while the "Points" and "Weigh" columns show relatively lower variability and narrower ranges.

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?

Sol: Expected Value formula for Discrete Random Variable is:

E(X) = Σ[x\*P(x)]

Total weights are: 9

Probability of each random variable is 1/9

EV = (108\*1/9) + (110\*1/9) + (123\*1/9) + (134\*1/9) + (135\*1/9) + (145\*1/9) + (167\*1/9) + (187\*1/9) + (199\*1/9)

EV = 145.33

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

Car’s speed and distance

Use Q9\_a.csv

Sol: Skewness: Speed -0.117510

Dist 0.806895

Speed:

The skewness is close to zero but slightly negative, indicating a very slight left-skewed distribution.

The data might have a slightly longer tail on the left side of the distribution compared to a perfectly symmetric distribution.

Distance:

The skewness is moderately positive, indicating a moderately right-skewed distribution.

The data might have a longer tail on the right side, suggesting that there might be some outliers or higher values compared to a symmetric distribution.

Kurtosis: Speed -0.508994

Dist 0.405053

Speed:

The kurtosis is negative, indicating a platykurtic distribution.

This suggests that the distribution has lighter tails and is flatter compared to a normal distribution. The data might have fewer outliers or extreme values compared to a normal distribution.

Distance:

The kurtosis is positive, indicating a mesokurtic or slightly leptokurtic distribution.

This suggests that the distribution has a shape closer to a normal distribution but with slightly heavier tails than a normal distribution.

SP and Weight (WT)

Use Q9\_b.csv

Sol: Skewness SP 1.611450

WT -0.614753

SP: Skewness is moderately high positive, indicates a right-skewed distribution. The tail of the distribution is longer on the right side, implying the presence of higher values or outliers in the data pulling the mean towards the higher end.

WT: Skewness is moderately high negative, indicates a left-skewed distribution. The tail of the distribution is longer on the left side, implying the presence of lower values or outliers in the data pulling the mean towards the lower end.

Kurtosis: SP 2.977329

WT 0.950291

SP: The Speed variable has a high positive kurtosis (2.977).

This indicates a leptokurtic distribution, meaning it has heavier tails and is more peaked compared to a normal distribution.

The data has more extreme values (outliers) or higher probabilities in the tails compared to a normal distribution.

WT: The Weight variable has a moderate positive kurtosis (0.950).

This indicates a mesokurtic or slightly leptokurtic distribution.

The data has a shape closer to a normal distribution, but with slightly heavier tails compared to a normal distribution.

Q10) Draw inferences about the following boxplot & histogram



Sol:

most chicks have weights clustered around 100 units, as indicated by the highest bar. The chick weight dataset is not normally distributed, It appears that the dataset is skewed to the right, has a long tail to the right and is slightly flatter than a normal distribution.



Boxplot has outliers on the maximum side.

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?

Sol: sample mean (x’) = 200 (point estimate)

Sample Standard Deviation (s) = 30

Sample size (n) = 2000

Population size (N) = 3000000

Degrees of freedom = sample size -1 = 2000-1 =1999

Standard error = s/sqrt(n)

Marginal Error = t\*Standard Eror = t\* s/sqrt(n)

Confidence interval = Point Estimate (+/-) Marginal Error

= **[x’ - (t \* s/sqrt(n))]** to **[x’ + (t \* s/sqrt(n))]**

Critical Value of t at 94% is 1.881

Critical Value of t at 96% is 2.054

Critical Value of t at 98% is 2.326

For 94% confidence interval Range is [ 198.74 – 201.26]

For 96% confidence interval range is [198.62 – 201.38]

For 98% confidence interval range is [198.44 – 201.56]

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.

Sol: Mean =41, Median = 40.5, Mode =41, Variance = 25.53, SD = 5.05

1. What can we say about the student marks?

Sol:

Observations:

There is a slight positive skew in the data, meaning the distribution is slightly "tailed" towards higher scores. This is indicated by the mean being slightly higher than the median.

Scores range from 34 to 56, showing a spread of 22 points.

There are two potential outliers: 49 and 56. These scores are significantly higher than the rest of the data and might be worth investigating

Interpretation:

The student's overall performance is above average. The mean score of 41.00 suggests consistent performance above the typical range.

There is some variability in the scores, as indicated by the standard deviation of 5.05. However, the scores are mostly clustered around the mean, suggesting relatively consistent performance throughout the tests.

The presence of potential outliers requires further exploration to understand their context and impact on the overall analysis.

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

Ans: when mean and median of dataset are equal, it generally indicates that the distribution is symmetrical.

Zero skewness: In a perfectly symmetrical distribution, the mean and median are indeed equal, and the skewness value will be zero. This confirms that the data is evenly distributed on both sides of the central point.

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

Ans: Skewness is positive and it is said right skewed. Higher values present on the right side. So, the tail extends. Most of the data points are concentrated on the left side of the distribution than the right.

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

Ans: Skewness is negative and it is said left skewed. Higher values present on the left side. So, the tail extends. Most of the data points are concentrated on the right side of the distribution than the left.

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

Ans: A positive kurtosis value indicates that a data distribution has heavier tails and a more peaked center compared to a normal distribution.

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

Ans: A negative kurtosis value indicates that a data distribution has lighter tails and a flatter peak compared to a normal distribution.

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



What can we say about the distribution of the data?

Ans: It is not normally distributed. It is skewed.

What is nature of skewness of the data?

Ans: The data is left skewed or negative skewness.

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

Ans: IQR = Q3-Q1 18-10 = 8

Q19) Comment on the below Boxplot visualizations?



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

Ans: **Boxplot 1** shows a more compact distribution of data with a smaller interquartile range compared to **Boxplot 2.**

**Boxplot 2** shows a more spread-out distribution of data with a larger interquartile range compared to **Boxplot 1**.

There are no outliers. Both boxplots share same median that is around 263 and they are normally distributed with zero skewness.

Q20) 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)

Ans: Prob\_MPG\_greater\_than\_38 = 1 - stats.norm.cdf(38, loc= cars.MPG.mean(), scale= cars.MPG.std())

print('P(MPG>38) =', Prob\_MPG\_greater\_than\_38)

**P(MPG>38) = 0.34759392515827137**

1. P(MPG<40)

Ans: prob\_MPG\_lessthan\_40 = stats.norm.cdf(40, loc= cars.MPG.mean(), scale= cars.MPG.std())

print('P(MPG<40) =', prob\_MPG\_lessthan\_40)

**P(MPG<40) = 0.7293498762151609**

1. P (20<MPG<50)

Ans: Prob\_MPG\_greater\_than\_20 = stats.norm.cdf(20, loc= cars.MPG.mean(), scale= cars.MPG.std())

print('P(MPG>20) =', Prob\_MPG\_greater\_than\_20)

P(MPG>20) = 0.05712377632115912

Prob\_MPG\_less\_than\_50 = stats.norm.cdf(50, loc= cars.MPG.mean(), scale= cars.MPG.std())

print('P(MPG<50) =', Prob\_MPG\_less\_than\_50)

P(MPG<50) = 0.9559926932893639

prob\_20\_lessthan\_MPG\_lessthan\_50 = Prob\_MPG\_less\_than\_50 - Prob\_MPG\_greater\_than\_20

print('P(20<MPG<50) =', prob\_20\_lessthan\_MPG\_lessthan\_50)

**P(20<MPG<50) = 0.8988689169682047**

Q 21) Check whether the data follows normal distribution

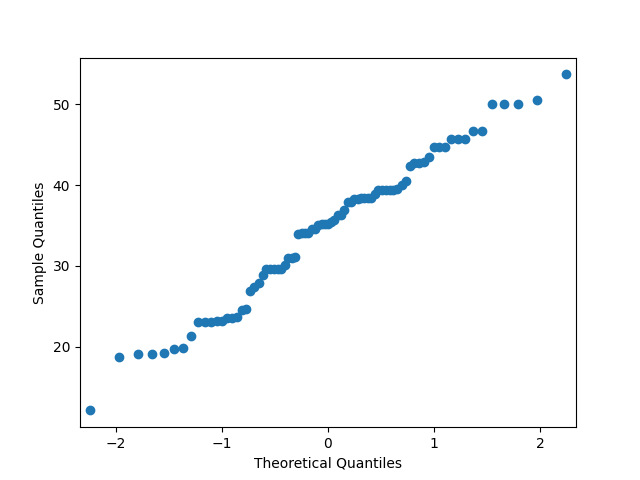
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:

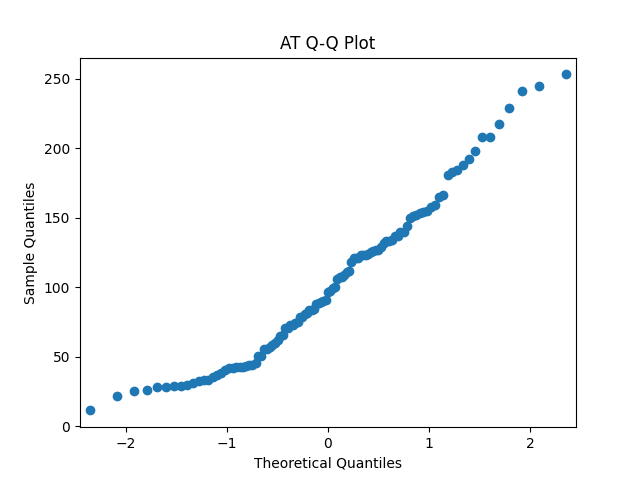
Mean = 34.422, Median = 35.152, Skewness = -0.177, Kurtosis = -0.611

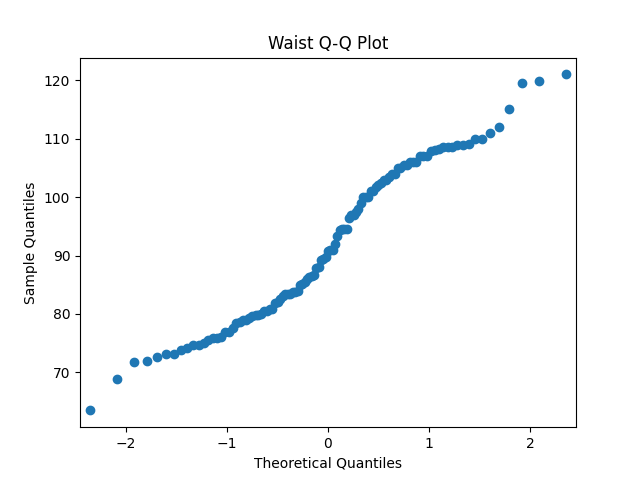
Mean and Median are approx. same and Skewness and Kurtosis are near to zero indicates that MPG of Cars follows approx. normal distribution.



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

Dataset: wc-at.csv





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

Ans:

> qnorm(0.95)

[1] 1.644854

> qnorm(0.97)

[1] 1.880794

> qnorm(0.80)

[1] 0.8416212

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

Ans:

> # Sample size

> n <- 25

> # Degrees of freedom

> df <- n - 1

> # T score for 95% confidence interval

> t\_95 = qt(0.975, df)

T score for 95% confidence interval: 2.063899

> # T score for 96% confidence interval

> t\_96 = qt(0.98, df)

T score for 96% confidence interval: 2.171545

> # T score for 99% confidence interval

> t\_99 = qt(0.995, df)

T score for 99% confidence interval: 2.79694

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

Ans: Population Mean = 270

Sample = 18

Sample mean = 260

Standard deviation = 90

Degree of Freedom = 18-1 = 17

Tscore = = = -0.47

> pt(-0.47, 17)

[1] 0.3221639

Required Probability = 0.3221639 = 32.21%