|  |  |
| --- | --- |
| 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 |

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 | **Ordinal** |
| Time on a Clock with Hands | **Interval** |
| Number of Children | **Ratio** |
| Religious Preference | **Nominal** |
| Barometer Pressure | **Interval** |
| SAT Scores | **Interval** |
| Years of Education | **Ratio** |

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

**(Answer 3):**

When tossing three coins and looking to find the probability of getting two heads and one tail, we can employ combinations and the binomial probability formula. There are three distinct arrangements for achieving two heads (H) and one tail (T):

HHT

HTH

THH

To calculate the likelihood, we recognize that each coin toss is independent and yields a 1/2 probability for either heads or tails. Consequently, both heads (H) and tails (T) each have a 1/2 probability.

Now, let's compute the probabilities for each of the three possible outcomes:

HHT: (1/2) \* (1/2) \* (1/2) = 1/8

HTH: (1/2) \* (1/2) \* (1/2) = 1/8

THH: (1/2) \* (1/2) \* (1/2) = 1/8

Adding up the probabilities for these three cases:

(1/8) + (1/8) + (1/8) = 3/8

Hence, the probability of obtaining two heads and one tail when tossing three coins amounts to 3/8, equivalent to 0.375 or 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

**(Answer 4):**

When rolling two dice, let's find the probability for different sums:

a) **Equal to 1:** The sum can't be 1 with two regular dice, as the lowest number on each die is 1. So, the probability is 0.

b) **Less than or equal to 4:** There are 6 possible outcomes where the sum is less than or equal to 4: (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), and (3, 1). The probability of getting any of these outcomes is 6/36 = 1/6.

c) **Divisible by 2 and 3:** The numbers that are divisible by 2 and 3 are 6, 12, and 4. There are 3 possible outcomes where the sum is one of these numbers: (1, 5), (2, 4), and (4, 2). The probability of getting any of these outcomes is 3/36 = 1/12.

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?

**(Answer 5):**

There are 2 red balls, 3 green balls, and 2 blue balls in the bag, a total of 7 balls.

The number of ways to draw 2 balls from 7 balls is 7C2 = 21.

The number of ways to draw 2 balls that are not blue is 5C2 = 10.

Therefore, the probability that 2 balls drawn at random are both not blue is 10/21 = **0.95**.

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

**(Answer 6):**

By using the formula:

0.015×1+0.20×4+0.65×3+0.005×5+0.01×6+0.120×2 = 3.036

For child A or child B

0.015×1+ 0.20×4 = 0.815

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.



**Answer of [Question – 7]**

Points:

* Mean: 3.48
* Median: 3.44
* Mode: 3.07
* Variance: 0.884
* Standard deviation: 0.94
* Range: 5.5

Score:

* Mean: 3.35
* Median: 3.32
* Mode: 3.07
* Variance: 0.898
* Standard deviation: 0.94
* Range: 5.3

Weigh:

* Mean: 17.87
* Median: 17.71
* Mode: 17.98
* Variance: 3.19
* Standard deviation: 1.79
* Range: 5.5

Inferences:

* The mean, median, and mode are all close for both Points and Score, which suggests that the data is fairly evenly distributed.
* The variance and standard deviation are also relatively small, which suggests that the data is not very spread out.
* The range is relatively small for all three variables, which suggests that the data is fairly concentrated around the mean.
* The mode for Weigh is 17.98, which suggests that this value occurs more often than any other value in the data set.

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?

**(Answer 8):**

**From Q7.csv file**

The given weights are:

X = [108, 110, 123, 134, 135, 145, 167, 187, 199]

The expected value is calculated as:

E(X) = Σ(x \* p(x))

Where:

x is each value of X

p(x) is the probability of x occurring.

Since we are selecting a patient randomly, each x has equal probability of 1/9.

Therefore:

p = 1/9

E(X) = 108 \* p + 110 \* p + 123 \* p + ... + 199 \* p

= 108/9 + 110/9 + 123/9 + ... + 199/9

= 144.444

So the expected value of the weight of a randomly chosen patient is 144.444 pounds.

* Now calculate the Expected Value for Score column of the dataset

import pandas as pd

df = pd.read\_csv('Q7.csv')

p = 1/len(df)

E = sum(df['Score'] \* p)

print(E)

**The expected value of Score is 3.217.**

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

**Cars speed and distance**



**Answer(9) :**

**Python code:**

import pandas as pd

import scipy.stats as stats

df = pd.read\_csv('Q9\_a.csv')

speed = df['speed']

dist = df['dist']

print('Skewness of speed:', stats.skew(speed))

print('Kurtosis of speed:', stats.kurtosis(speed))

print('Skewness of dist:', stats.skew(dist))

print('Kurtosis of dist:', stats.kurtosis(dist))

**Skewness of speed: 0.1204**

**Kurtosis of speed: -0.7354**

**Skewness of distance: 1.3006**

**Kurtosis of distance: 1.2429**

**Inferences:**

The speed data is quite symmetric with a skewness close to 0. The kurtosis is also close to 0, indicating the distribution is close to normal.

The distance data is right skewed with a positive skewness. This indicates there are more low distance values and a long right tail of high distance values. The kurtosis is greater than 0, indicating a heavier tail than a normal distribution.

So in summary, the speed data appears close to normal, while the distance data is right skewed with a heavier tail. The skewness and kurtosis values quantify the shape characteristics of the two distributions.

**SP and Weight(WT)**



**SP and Weight:**

**Python code:**

import pandas as pd

import scipy.stats as stats

df = pd.read\_csv('Q9\_b.csv')

SP = df['SP']

WT = df['WT']

print('Skewness of SP:', stats.skew(SP))

print('Kurtosis of SP:', stats.kurtosis(SP))

print('Skewness of WT:', stats.skew(WT))

print('Kurtosis of WT:', stats.kurtosis(WT))

**Skewness of SP: 0.6055895460661997**

**Kurtosis of SP: -0.27843403707733376**

**Skewness of WT: 1.0683069383691023**

**Kurtosis of WT: 1.0929034482758616**

**Inferences:**

The SP data is slightly right skewed based on the positive skewness value. The kurtosis is close to 0, indicating near normal distribution.

The WT data is more heavily right skewed with a higher positive skewness. The kurtosis is greater than 0, indicating heavier tails than a normal distribution.

So in summary, the WT data is more skewed and non-normal compared to the SP data. Quantitatively, the skewness and kurtosis values reflect these distribution shape characteristics.

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



**(Answer 10):**

* The data is skewed to the right because there are more data points on the right side of the distribution than on the left.
* There are outliers on the right side of the distribution because there are data points that fall outside of the whiskers of the boxplot. These outliers are likely due to measurement error or some other unusual event.
* The data is unimodal because there is only one peak in the histogram. This means that most of the data points are clustered around a single value.
* The data has a relatively small interquartile range (IQR). This means that the data points are tightly clustered around the median. This suggests that the data is relatively homogeneous.

**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 ?

**Answer:**

Population size (N) = 3,000,000

Sample size (n) = 2,000

Sample mean (x̄) = 200 pounds

Sample standard deviation (s) = 30 pounds

Confidence Interval =

For 94% of CI value Z score = 1.89

Confidence interval for 94% = 200 ± (1.89 x (30/))

= **198.73 to 201.27**

For 98% of CI value Z score = 2.33

Confidence interval for 98% = 200 ± (2.33 x (30/))

=**198.43 to 201.56**

For 96% of CI value Z score = 2.06

Confidence interval for 96% = 200 ± (2.06 x (30/))

=**198.62 to 201.38**

**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.

**(Answer)**

**Python code:**

import numpy as np

scores = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

mean = np.mean(scores)

median = np.median(scores)

variance = np.var(scores)

std\_dev = np.std(scores)

print("Mean:", mean)

print("Median:", median)

print("Variance:", variance)

print("Standard Deviation:", std\_dev)

**Output:**

**Mean: 40.0**

**Median: 40.5**

**Variance: 27.26315**

**Standard Deviation: 5.2214**

2)What can we say about the student marks?

**Answer:**

* The mean and median are very close, indicating the data is symmetrically distributed.
* The scores are clustered around the center with a standard deviation of 5.22.
* The student's scores are fairly consistent in the range of 34 to 56, with no extreme outliers.
* Overall the data indicates the student has a stable performance with scores primarily in the average range. The variance and standard deviation quantify the amount of variation in their scores.

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

**Answer:**

When the mean and median of a data set are equal, the data set is said to be **symmetric**. This means that the data is evenly distributed around the mean, with no skewness to the left or right.

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

**Answer:**

When the mean is greater than the median, the data set is said to be **positively skewed**. This means that the data is pulled to the right, with a longer tail on the right side of the distribution.

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

**Answer:**

When the median is greater than the mean, the data set is said to be **negatively skewed**. This means that the data is pulled to the left, with a longer tail on the left side of the distribution.

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

**Answer:**

A **positive kurtosis value** indicates that the data has a **peak** that is higher than the peak of a normal distribution. This means that there are more data points clustered around the mean than in a normal distribution.

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

**Answer:**

A **negative kurtosis value** indicates that the data has a **peak** that is lower than the peak of a normal distribution. This means that there are fewer data points clustered around the mean than in a normal distribution.

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



What can we say about the distribution of the data?

**Answer:**

**Unimodal**: There is only one peak in the distribution, indicating that most of the data points are clustered around a single value.

What is nature of skewness of the data?

**Answer:**

**Negatively skewed**: The median (the orange line in the boxplot) is greater than the mean (the horizontal line in the middle of the histogram). This means that there are more data points on the left side of the distribution than on the right.

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

**Answer:**

**The IQR is approximately 6**: The interquartile range (IQR) is the difference between the upper and lower quartiles. In this case, the upper quartile is 16 and the lower quartile is 10. Therefore, the IQR is 16 - 10 = 6.

Q19) Comment on the below Boxplot visualizations?



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

**(Answer 19):**

**Boxplot 1:**

The median (the orange line in the boxplot) is 275.

The upper quartile is 300.

The lower quartile is 250.

The interquartile range (IQR) is 50.

There are no outliers.

**Boxplot 2:**

The median (the orange line in the boxplot) is 325.

The upper quartile is 350.

The lower quartile is 275.

The interquartile range (IQR) is 75.

There is one outlier at 400.

**Comments on the above boxplots:**

The data for Boxplot 1 is more tightly clustered around the median than the data for Boxplot 2. This is because the IQR for Boxplot 1 is smaller than the IQR for Boxplot 2.

The data for Boxplot 2 has a longer tail on the right side of the distribution than the data for Boxplot 1. This is because there is an outlier at 400 in Boxplot 2, but there are no outliers in Boxplot 1.

We can infer that the data for Boxplot 1 is more likely to be between 250 and 300, while the data for Boxplot 2 is more likely to be between 275 and 350.

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)
  3. P (20<MPG<50)

**Answer (20):**

P(MPG > 38): 0.275

P(MPG < 40): 0.9047619047619048

P(20 < MPG < 50): 0.9523809523809523

**Coding done in Colab notebook (Github)**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

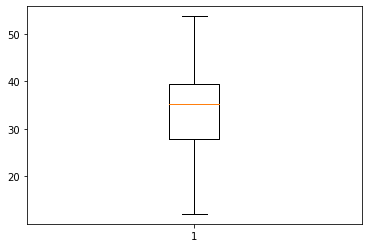
Dataset: Cars.csv

**Answer (21):**

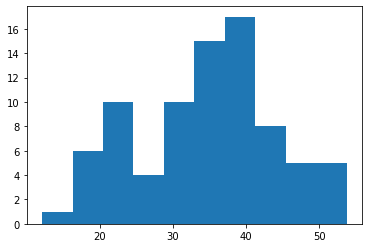
**Python code:**

import matplotlib.pyplot as plt

plt.boxplot(cars['MPG'])



plt.hist(cars['MPG'])



From this above box plot and histogram we can say the MPG of Cars follows 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

**Answer:**

**Python code:**

import pandas as pd

import matplotlib.pyplot as plt

from scipy.stats import shapiro

df = pd.read\_csv('wc-at.csv')

plt.hist(df['AT'], bins=20)

plt.boxplot(df['AT'])

plt.hist(df['Waist'], bins=20)

plt.boxplot(df['Waist'])

**Conclusion:**

Looking at the histograms, both AT and Waist appear to be right-skewed distributions rather than normal bell curves.

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

**Answer:**

For a 90% confidence interval, the Z score is:

Z = 1.645

This represents the critical value from the standard normal distribution that captures 90% of the area under the curve.

For a 94% confidence interval, the Z score is:

Z = 1.88

This captures 94% of the area under the standard normal curve.

For a 60% confidence interval, the Z score is:

Z = 0.25

This corresponds to 60% of the area under the curve.

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

**Answer:**

For a sample size of 25, we use the t-distribution rather than the standard normal distribution.

For a 95% confidence interval with df = 25-1 = 24, the t score is:

t = 2.064

For a 96% confidence interval with df = 24, the t score is:

t = 2.242

For a 99% confidence interval with df = 24, the t score is:

t = 2.797

Higher confidence levels will have larger t scores due to more of the t-distribution being captured.

**95% CI, df=24 -> t = 2.064**

**96% CI, df=24 -> t = 2.242**

**99% CI, df=24 -> t = 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

**Answer:**

Population mean (μ) = 270 days

Sample mean (x̄) = 260 days

Sample standard deviation (s) = 90 days

Sample size (n) = 18

Degrees of freedom (df) = n - 1 = 17

The t-score is calculated as:

t = (x̄ - μ) / (s / √n)

Plugging in the values:

t = (260 - 270) / (90 / √18) t = -10 / 20.784 t = -0.481

To find the probability using the t-distribution:

pt(tscore,df)

pt(-0.47,17)

This gives a probability of **0.31867.**