|  |  |
| --- | --- |
| 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 | 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?**

P = (Number of Favorable Outcomes)/ (Total Possible Outcomes)

= 3/8

Answer:

0.375 = 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

P = (Number of Favorable Outcomes)/ (Total Possible Outcomes)

Total possible outcome = 6\*6 =36

1. Sum of two dice equal to 1 is not possible so P = 0
2. Sum of two dice less than or equal to 4, P = = 16.66%
3. Sum of two dice is divisible by 2 and 3, P = = 16.66%

Answer:

1. 0

b) 16.66%

c) 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?**

Total balls are = 7

Total no.of outcomes for 2 balls are drawn randomly from bag = = 21

No.of outcomes of none of the balls drawn is blue = = 10

None of the balls drawn is blue P =

Answer:

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

P = 0.015+0.8+1.95+0.025+0.06+0.24 = 3.14

Answer:

3.14 candies

**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**

**Answers:** For points,

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weight |
| MEan | 3.59 | 3.21 | 17.84 |
| Median | 3.695 | 3.325 | 17.71 |
| Variance | 0.285 | 0.957 | 3.19 |
| Mode | 0 3.07 | 0 3.44 | 0 17.02 |
| Standard Deviation | 0.534 | 0.978 | 1.786 |
| Range | 2.17 | 3.911 | 8.399 |

These statistics provide insights into the distribution and spread of the data. For example, the mean, median, and mode values give an idea of the central tendency of the data. The variance and standard deviation indicate the spread or dispersion of the data points. The range represents the difference between the minimum and maximum values in each category.

A picture containing text, screenshot, diagram, plot

Description automatically generated

**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?

Expected value = Sum (X \* Probability of X)

= + 110 + + + + + +

Answer:

145.4 pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Answers:**

import pandas as pd

import numpy as np

from scipy.stats import skew,kurtosis

df1 = pd.read\_csv(""C:\Users\FAYAZ AHAMED\OneDrive\Desktop\Data Science\Basic Statistics level1\Q9\_a.csv"")

df1

df1["speed"].hist()

df1["speed"].skew()

df1["speed"].kurtosis()

df1["dist"].hist()

df1["dist"].skew()

df1["dist"].kurtosis()

skewness = skew(df1,bias= True)

skewness

kurtosis = kurtosis(df1,bias= True)

kurtosis

Skewness (Speed): -0.1175

Skewness (Dist): 0.8069

Kurtosis (Speed): -0.509

Kurtosis (Dist): 0.4051

**SP and Weight(WT)**

**Use Q9\_b.csv**

df2 = pd.read\_csv(""C:\Users\FAYAZ AHAMED\OneDrive\Desktop\Data Science\Basic Statistics level1\Q9\_b.csv"")

df2

df2["SP"].hist()

df2["SP"].skew()

df2["SP"].kurtosis()

df2["WT"].hist()

df2["WT"].skew()

df2["WT"].kurtosis()

skewness = skew(df2,bias= True)

skewness

kurtosis = kurtosis(df2,bias= True)

kurtosis

Skewness (SP): 1.581

Skewness (WT): -0.6033

Kurtosis (SP): 2.7235

Kurtosis (WT): 0.8194

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



**Answer:** This is a histogram with a right-skewed distribution, where the majority of the data is concentrated on the left side, and there are outliers on the right side, leading to a situation where the mean is greater than the median due to the skewness of the data.



* These circles on the plot are considered outliers because they fall outside the typical range of the data.

**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:

Mean = 200

standard\_deviation = 30

Sample\_size = 2000

from scipy.stats import norm

# 94% Confidence interval

CI = norm.interval(0.94,Mean,standard\_deviation)

CI

# 98% Confidence interval

CI = norm.interval(0.98,Mean,standard\_deviation)

CI

# 96% Confidence interval

CI = norm.interval(0.96,Mean,standard\_deviation)

CI

94.0% Confidence Interval: (198.738, 201.26)

98.0% Confidence Interval: (198.44, 201.56)

96.0% Confidence Interval: (198.622, 201.37)

**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.**

Mean: 41.0

Median: 40.5

Variance: 24.11111111111111

Standard Deviation: 4.910306620885412

1. **What can we say about the student marks?**

**Answer**: There are no outliers, and the plot is skewed towards the right because the mean > median.

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

**Answer**: When the mean and median of a dataset are equal, it indicates that the data is symmetrically distributed and has no skewness. The distribution can be considered as perfectly symmetrical.

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

**Answer**: When the mean of a dataset is greater than the median, it suggests that the distribution is positively skewed. The tail of the distribution extends towards the higher values, indicating the presence of outliers or a longer right tail.

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

**Answer**: When the median of a dataset is greater than the mean, it suggests that the distribution is negatively skewed. The tail of the distribution extends towards the lower values, indicating the presence of outliers or a longer left tail.

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

**Answer**: A positive kurtosis value indicates that the distribution has heavy tails and a sharper peak compared to the normal distribution. It suggests that the data has more extreme values (outliers) and is leptokurtic, meaning it has more data points in the tails than the normal distribution.

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

**Answer**: A negative kurtosis value indicates that the distribution has lighter tails and a flatter peak compared to the normal distribution. It suggests that the data has fewer extreme values (outliers) and is platykurtic, meaning it has fewer data points in the tails than the normal distribution.

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



**What can we say about the distribution of the data?**

**Answer**: From the box plot we can say the at the median is having higher value. Which means the plot does not have a normal distribution

**What is nature of skewness of the data?**

**Answer**: The plot has negative skewness.

**What will be the IQR of the data (approximately)?   
   
 Answer**: IQR = Q3 -Q1, = 18-10

IQR= 8

**Q19) Comment on the below Boxplot visualizations?**



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

The plot is normally distributed. Both the boxplots have same median which is in between 250 - 275 and also have no outliers.

**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

**Answer:** import pandas as pd

import numpy as np

df = pd.read\_csv(""C:\Users\FAYAZ AHAMED\OneDrive\Desktop\Data Science\Basic Statistics level1\Cars.csv"")

df

MPG\_Value = df["MPG"]

MPG\_Value

## a.P(MPG>38)

len(MPG\_Value[MPG\_Value>38])/len(MPG\_Value)

## b.P(MPG<40)

len(MPG\_Value[MPG\_Value<40])/len(MPG\_Value)

## c.P(20<MPG<50

len(MPG\_Value[(MPG\_Value>20)&(MPG\_Value<50)]) / len(MPG\_Value)

* 1. P(MPG>38)

**Ans:** 0.407407

* 1. P(MPG<40)

**Ans:** 0.7530

* 1. P (20<MPG<50)

**Ans:** 0.851851

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

**Ans:**  import pandas as pd

import numpy as np

df1 = pd.read\_csv("/content/Cars.csv")

df1

MPG\_Value = df1["MPG"]

MPG\_Value

from scipy.stats import shapiro

Statistics, P\_Value = shapiro(MPG\_Value)

Statistics

P\_Value

The P\_value is 0.176 which is greater than 0.05 so, the MPG of Cars are Nominal Distribution.

A graph with a red line and blue dots

Description automatically generated

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

**Dataset: wc-at.csv**

df2 = pd.read\_csv("/content/wc-at.csv")

df2

AT = df2["AT"]

AT

Waist = df2["Waist"]

Waist

from scipy.stats import shapiro

statistic\_AT, P\_Value\_AT = shapiro(AT)

statistic\_AT

P\_Value\_AT

statistic\_Waist, P\_Value\_Waist = shapiro(Waist)

statistic\_Waist

P\_Value\_Waist

if P\_Value\_AT < 0.05:

    print("H0 is rejected and H1 is accepted")

else:

    print("H0 is accepted and H1 is rejected")

if P\_Value\_Waist < 0.05:

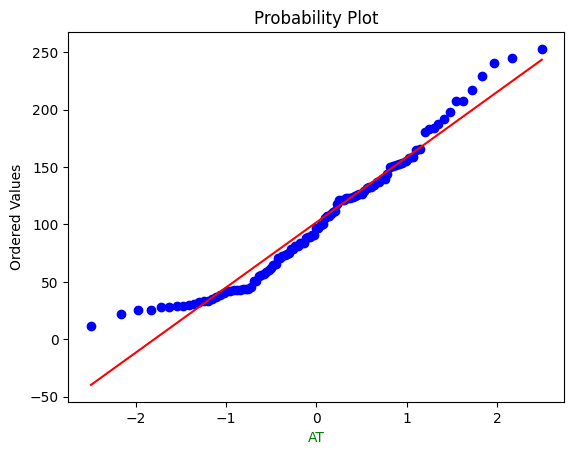
    print("H0 is rejected and H1 is accepted")

else:

    print("H0 is accepted and H1 is rejected")

A graph with a red line

Description automatically generated



The P\_values of Adipose Tissue and Waist Circumference(Waist) are 0.000653 and 0.00117 respectively which are less than 0.05 so, both Aditive tissue and Waist Circumference does not have nominal distribution.

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

**Ans:**

import scipy.stats as stats

Z\_Score = stats.norm.ppf((1+0.90)/2)

Z\_Score

Z\_Score = stats.norm.ppf((1+0.94)/2)

Z\_Score

Z\_Score = stats.norm.ppf((1+0.60)/2)

Z\_Score

Z score of 90% confidence interval - 1.644

Z score of 94% confidence interval - 1.880

Z score of 60% confidence interval - 1.841

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

**Ans:**  Sample\_size = 25

import scipy.stats as stats

Degree\_of\_freedom = Sample\_size - 1

T\_Score = stats.t.ppf((1+0.95)/2,Degree\_of\_freedom)

T\_Score

T\_Score = stats.t.ppf((1+0.96)/2,Degree\_of\_freedom)

T\_Score

T\_Score = stats.t.ppf((1+0.99)/2,Degree\_of\_freedom)

T\_Score

T score of 90% confidence interval - 2.093

T score of 94% confidence interval - 2.204

T score of 60% confidence interval - 2.860

**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

# Given Data

Population\_mean\_Days = 270

Sample\_size = 18

Sample\_mean\_Days = 260

Standard\_deviation\_Days = 90

    ## P of n <= 260 Days

## A:-

T\_Score = (Sample\_mean\_Days-Population\_mean\_Days)/(Standard\_deviation\_Days/(Sample\_size\*\*0.5))

T\_Score

Degree\_of\_freedom = Sample\_size - 1

Degree\_of\_freedom

import scipy.stats as stats

P\_Value = stats.t.cdf(T\_Score,Degree\_of\_freedom)

P\_Value

print("The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is ", (P\_Value\*100).round(1),"%")

**Ans**: The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32.2 %