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

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

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

Ans: When three coins are tossed then the outcome will be any one of the combinations of TTT, THT, TTH, THH, HTT, HHT, HTH, HHH.

So, the total number of outcomes is 8.

Now, for two heads and one tail favorable outcomes is THH,HHT,HTH.

So, the total favorable outcomes is 3.

We know,

Probability = Number of favorable outcomes/Total number of outcomes

Probability = 3/8 = 0.375 = 37.5 %

So, the probability of getting two heads and one tail is 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

Ans:

(a) Upon rolling two dice, the minimum sum on their faces would be 1 + 1 = 2.

So, the sum is never 1.

i.e., it's an Impossible event. The probability is 0.

(b) Let ‘x’ be the number on the first dice.

‘y’ be the number on second dice.

We have to find the probability that the sum is 4 or lesser

i.e. (x + y) <= 4

Total number of possible results from two dice is 6 × 6 = 36

The possible results that the sum is lower or equal than 4 is:

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

The probability that the sum is lower or equal than 4 is

6/36 = 1/6 = 0.16 = 16.66 %.

(c) Total number of possible results from two dice is 6 × 6 = 36

we want the sum to be divisible by both 2 and 3. This means we are looking for sums that are multiples of 6 since any number divisible by both 2 and 3 is also divisible by 6.

The sums that are multiples of 6 are: 6, 12.

For a sum of 6, the combinations are (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) - a total of 5 combinations.

For a sum of 12, the only combination is (6, 6).

So, there are 6 favorable outcomes.

Probability = Number of Favorable Outcomes / Total Number of Outcomes

Probability = 6 / 36 = 1 / 6 = 0.16 = 16.66 %.

So, the probability that the sum of two dice rolls is divisible by both 2 and 3 is 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?

Ans:

Total number of balls = (2 + 3 + 2) = 7  
Then, Let “x” be the number of ways of drawing 2 balls out of 7 is

7C2 = (7×6)/(2×1) = 21

let “y” be the event of drawing 2 balls, none of which is blue.

Number of ways of drawing 2 balls out of (2 + 3) balls is 5C2 = (5×4)/(2×1) = 10  
P (None of the balls drawn is blue) = x/y= 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

Ans: The expected number of candies for a randomly selected child is 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**

Ans.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.59 | 3.21 | 17.84 |
| Median | 3.69 | 3.32 | 17.71 |
| Mode | 3.07 | 3.44 | 17.02 |
| Variance | 0.28 | 0.95 | 3.19 |
| Standard Deviation | 0.53 | 0.97 | 1.78 |
| Range | 2.17 | 3.911 | 8.4 |





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?

Ans:

The expected weight of the patient that is selected randomly is Mean.

Mean = Sum of all weight/Total number of patients

Mean =1308/9

Expected weight is similarly to 145.33 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans: Using Q9\_a.csv**

**Skewness of speed is (-0.12) its indicates that it is a left-skewed distribution.**

**Kurtosis of speed is (-0.51) it is less than “3” it indicates a platykurtic distribution.**

**Skewness of distance is (0.81) its indicates that it is a right-skewed distribution.**

**Kurtosis of distance is (0.41) it is less than “3” it indicates a platykurtic distribution.**

**Using Q9\_b.csv**

**Skewness of SP is (1.61) its indicates that it is a right-skewed distribution.**

**Kurtosis of SP is (2.98) it is less than “3” it indicates a normal distribution.**

**Skewness of WT is (-0.61) its indicates that it is a left-skewed distribution.**

**Kurtosis of WT is (0.95) it is less than “3” it indicates a platykurtic distribution.**

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



Ans: The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side.



Ans: The 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?

Ans :

94.0% Confidence Interval: (198.74,201.26) pounds

96.0% Confidence Interval: (198.62,201.38) pounds

98.0% Confidence Interval: (198.44,201.56) pounds

**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?

**Ans:**

1) mean = 41.0

Median = 40.5

Variance = 25.53

Standard Deviation = 5.05

1. we don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

Ans: The mean is equal to the median as well as the mode, hence the skewness is zero.

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

Ans: The mean is greater than the median, the distribution is positively skewed tail is towards Right.

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

Ans: The median is greater than the mean, the distribution is negatively skewed tail is towards Left.

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

Ans: Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.

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

Ans: Negative values of kurtosis indicate that distribution is flatter and broader.

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



What can we say about the distribution of the data?

Ans: The above Boxplot is not normally distributed the median is towards the higher value

What is nature of skewness of the data?

Ans: The data is a skewed towards left. The whisker range of minimum value is greater than maximum

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

Ans: The Inter Quantile Range IQR = Q3 Upper quartile – Q1 Lower Quartile = 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: First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range

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)

Ans : (a) Prob\_MPG\_greater\_than\_38 = 1 - stats.norm.cdf((38 - Cars.MPG.mean()) / Cars.MPG.std())

print('P(MPG>38)=',round(Prob\_MPG\_greater\_than\_38,3))

P(MPG>38) = 0.348

(b) prob\_MPG\_less\_than\_40 = stats.norm.cdf((40 - Cars.MPG.mean()) / Cars.MPG.std())

print('P(MPG<40)=',round(prob\_MPG\_less\_than\_40,3))

P(MPG<40) = 0.729

(c) prob\_MPG\_greater\_than\_20 = 1 - stats.norm.cdf((20 - Cars.MPG.mean()) / Cars.MPG.std())

print('P(MPG>20)=',round(prob\_MPG\_greater\_than\_20,3))

p(MPG>20) = 0.943

prob\_MPG\_less\_than\_50 = stats.norm.cdf((50 - Cars.MPG.mean()) / Cars.MPG.std())

print('P(MPG<50)=',round(prob\_MPG\_less\_than\_50,3))

P(MPG<50) = 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20)

print('P(20<MPG<50)=',round(prob\_MPG\_greaterthan20\_and\_lessthan50,3))

P(20<MPG<50) = 0.013

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

Ans :

1. MPG of cars follows normal distribution.



1. Adipose Tissue (AT) and Waist does not follow Normal Distribution





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

Ans :

confidence\_intervals = [0.60,0.90,0.94]

z\_scores = [norm.ppf((1 + ci) / 2) for ci in confidence\_intervals]

for ci, z\_score in zip(confidence\_intervals, z\_scores):

print(f"{ci \* 100}% Confidence Interval Z-score: {z\_score:.2f}")

90.0% Confidence Interval Z-score: 1.64

94.0% Confidence Interval Z-score: 1.88

60.0% Confidence Interval Z-score: 0.84

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

confidence\_intervals = [0.95,0.96,0.99]

degrees\_of\_freedom = sample\_size - 1

t\_scores = [stats.t.ppf((1 + ci) / 2, df=degrees\_of\_freedom) for ci in confidence\_intervals]

for ci, t\_score in zip(confidence\_intervals, t\_scores):

print(f"{ci \* 100}% Confidence Interval t-score: {t\_score:.2f}")

95.0% Confidence Interval t-score: 2.06

96.0% Confidence Interval t-score: 2.17

99.0% Confidence Interval t-score: 2.80

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 :

from scipy.stats import t

sample\_mean = 260

population\_mean = 270

sample\_std = 90

sample\_size = 18

t\_score = (sample\_mean - population\_mean) / (sample\_std / (sample\_size\*\*0.5))

round(t\_score,3) = -0.471

df = sample\_size – 1

df = 17

probability = t.cdf(t\_score, df)

round(probability, 3) = 322

print(f"The probability that 18 randomly selected bulbs have an average life of no more than 260 days is approximately: {probability:.2f}")

The probability that 18 randomly selected bulbs have an average life of no more than 260 days is approximately 0.32 = 32%