|  |  |
| --- | --- |
| 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 | **Ratio** |
| Time on a Clock with Hands | **Ratio** |
| Number of Children | **Nominal** |
| 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?

**Ans:** When three coins are tossed, the possible outcomes can be:

HHH (Head, Head, Head)

HHT (Head, Head, Tail)

HTH (Head, Tail, Head)

THH (Tail, Head, Head)

TTH (Tail, Tail, Head)

THT (Tail, Head, Tail)

HTH (Head, Tail, Tail)

TTT (Tail, Tail, Tail)

Out of these eight outcomes, three of them have two heads and one tail: HHT, THH, and HTH.

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

3/8 or 37.5%.

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

Equal to 1

**Ans:**

a) P (sum is Equal to 1) = ‘0’ (zero, null, nada, none).

b) P (Sum is less than or equal to 4) = N (Event (Sum is less than or equal to 4)) / N (Event (Two dice rolled)) = 6 / 36 = 1/6 = 0.166 = 16.66%.

c) P (Sum is divisible by 2 and 3) = N (Event (Sum is divisible by 2 and 3)) / N(Event (Two dice rolled)) = 6 / 36 = 1/6 = 0.166 = 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 = 7 balls

Event (2 balls are drawn randomly from the bag):

N(Event) = 7! / (2! \* 5!) = (7654321) / (21 \* 54321) = 21

Event (None of the balls drawn is blue):

N(Event) = 5! / (2! \* 3!) = 54 / (2 \* 1) = 10

Probability (None of the balls drawn is blue):

P(Event) = N(Event) / N(Event (2 balls are drawn randomly from the bag)) = 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:**

To calculate the expected number of candies for a randomly selected child, you can use the following formula:

E(X) = ∑(Xi × P(Xi))

Where E(X) is the expected value,Xi is the candies count, and P(Xi) is the probability of that candies count.

Let’s calculate it:

E(X) = (1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120)

E(X) = 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24

E(X) = 3.09

So, 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.

**ANS:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MEAN | MEDIAN | MODE | Variance | Std. Dev | Range |
| Points | 3.59 | 3.70 | 3.92 | 0.29 | 0.53 | 2.17 |
| Score | 3.22 | 3.33 | 3.44 | 0.96 | 0.98 | 3.91 |
| Weigh | 17.85 | 17.71 | 17.02 | 3.19 | 1.79 | 8.40 |

No case has the variable Mean = Median = Mode

Thus as seen in the graph dataset “score” and “Weigh” has outliers

**Use Q7.csv file**

Q8) Calculate Expected Value for the problem below

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:**

Expected value

= Sum (X \* Probability of X)

=(1/9)(108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+ (1/9)(145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199)

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANS**:

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| speed | -0.11 | 2.42 |
| dist | 0.76 | 3.25 |

“dist” is positively skewed where as “speed” is negatively skewed

Thus dist has distribution of data concentrated on the left whereas speed has distribution on the right. As seen in the graph

Both dist and speed has positive Kurtosis

**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANS:**

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| SP | 1.55 | 5.72 |
| WT | -0.59 | 3.82 |

“SP” is positively skewed where as “WT” is negatively skewed

Thus SP has distribution of data concentrated on the left whereas WT has distribution on the right. As seen in the graph

Both WT and SP has positive Kurtosis

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



**ANS :**

Majority of the Chicks has weight in range 50 – 100, followed by 100 -150 and 150 – 200

The data is positively Skewed

3 Categories can be define (Under weight >50; Avg weight 51 – 150; Over weight <150)



**ANS:**

Data has outliers

Data is positively skewed

**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% | 98% | 96% |
| Upper | 201.04 | 201.38 | 201.17 |
| Lower | 198.96 | 198.62 | 198.83 |

**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**

Find mean, median, variance, standard deviation.

**ANS:**

|  |  |
| --- | --- |
| Mean | 41 |
| Median | 40.50 |
| Variance | 25.53 |
| Std Deviation | 5.05 |

What can we say about the student marks?

**ANS:**

Not normally distributed

Data has outlier

Majority of the students scored between 35 – 45 Marks

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

**ANS:** Skewness = 0. Perfectly symmetric bell shaped curve

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

**ANS:** Skewness = Positive. Data is distributed more on left

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

**ANS:** Skewness = Negative. Data is distributed more on right

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

**ANS :** Positive kurtosis means the curve is more peaked and it is Leptokurtic

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

**ANS:** Negative Kurtosis means the curve will be 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 = 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)

**ANS:**Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3)

print('P(MPG>38)=',

Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

1. P(MPG<40)

**ANS:** prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

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

P(MPG<40)= 0.729

1. P (20<MPG<50)

**ANS:** prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20))

p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

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

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)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))

P(20<MPG<50)= 0.013000000000000012

Q 21) Check whether the data follows normal distribution

Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:** MPG of cars follows normal distribution



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

Dataset: wc-at.csv

**Ans:** 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:**

# z value for 90% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.05),4))

Z score for 60% Conifidence Intervla = -1.6449

# z value for 94% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.03),4))

Z score for 60% Conifidence Intervla = -1.8808

# z value for 60% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.2),4))

Z score for 60% Conifidence Intervla = -0.8416

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

**Ans:**

# t score for 95% confidence interval

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))

T score for 95% Confidence Interval = -2.0639

# t value for 94% confidence interval

print('T score for 94% Confidence Inteval =',np.round(stats.t.ppf(0.03,df=24),4))

T score for 94% Confidence Inteval = -1.974

# t value for 99% Confidence Interval

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))

T score for 95% Confidence Interval = -2.7969

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:** import numpy as np

Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

(260-270)/90/np.sqrt(18))

t\_score = -0.471

stats.t.cdf(t\_score, df = 17)

0.32 = 32%