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

**Ans:** Three coins are tossed the probability is ,

P={HHH,HHT,HTT,HTH,THH,TTH,THT,TTT}

Probability of getting two heads and one tail is = {HHT,HTH,THH}

i.e

**The probability that two heads and one tail are obtained P=3/8**

**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:** Two dice are rolled the probability is .

P={ (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)

(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6)

(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)

(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)

(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6)

(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

1. **Equal to 1 :**

P=0/36 i.e 0

**the probability of getting a sum of 1 is 0.**

1. **Less than or equal to 4 :**

Possible outcomes are ={(1, 1), (1, 2), (1, 3)}

i.e P=3/36 or 1/12

**the probability of getting sum less than or equal to 4 is 1/12**

1. **Sum is divisible by 2 and 3:**

Possible outcomes are = {(1, 1), (1, 2), (1, 3), (1, 5), (2, 1), (2, 2), (2, 4), (2, 6), (3, 1), (3, 3), (3, 5), (3, 6), (4, 2), (4, 4), (4, 5), (4, 6) ,

(5, 1), (5, 3), (5, 4), (5, 5), (6, 2), (6, 3), (6, 4), (6, 6)}

i.e P=24/36 or 2/3

**the Probability of getting sum is divisible by 2 and 3 is 2/3**

**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 ball is 2+3+2=7 ball

The Probability of drowning two balls from the 7 ball is:

= 7!/(7-2)!\*2!

= 21

The probability of drowning two balls from 5 balls is (not blue) is 5C2

= 5!/(5-2)!\*2!

=10

**The probability that none of the balls drawn is blue P=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:** Expected number of candies = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) +

(5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

= 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24

= 3.085

**The expected number of candies for a randomly selected child is 3.085**

**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**

**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:** Calculating the mean of the weight of the patient

Mean = (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199) / 9

= 1,308 / 9

= 145.33

145.33 Pound is the Expected Value of the Weight of that Patient

**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**

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



**Inference About histogram:**

1. The Histogram is rightly skewed i.e. means there is outlier is present on the right side of the histogram
2. Mean of the data is greater than the median value since the data point is present at the right side
3. The right-skewed histogram may indicate a non-normal distribution of the data



**Inference About Boxplot:**

1. There are some data point is present above the upper extreme that means the data is rightly skewed.
2. The outliers may be caused by measurement errors or data entry errors, and it may be necessary to examine the data closely to identify and correct any errors.

**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:**

Given: n=2000

x̄=200

s=30

Calculating 94 % confidence level :

Applying t-test

= x̄ ± Z1- α , n-1 \* s/**√n**

**=** 200 ± Z0.94,1999 \* 30/ **√2000**

Upper Range ( 94% ) = 200 + Z0.94,1999 \* 30/ **√2000**

**=** 201.03 Pound

Lower Range ( 94% )= 200 - Z0.94,1999 \* 30/ **√2000**

**=** 198.97 Pound

Calculating 96 % confidence level :

Applying t-test

= x̄ ± Z1- α , n-1 \* s/**√n**

**=** 200 ± Z0.96,1999 \* 30/ **√2000**

Upper Range ( 96% ) = 200 + Z0.96,1999 \* 30/ **√2000**

**=** 201.17 Pound

Lower Range ( 96% )= 200 - Z0.96,1999 \* 30/ **√2000**

**=** 198.82 Pound

Calculating 98 % confidence level :

Applying t-test

= x̄ ± Z1- α , n-1 \* s/**√n**

**=** 200 ± Z0.98,1999 \* 30/ **√2000**

Upper Range ( 98% ) = 200 + Z0.98,1999 \* 30/ **√2000**

**=** 202.94 Pound

Lower Range ( 98% )= 200 - Z0.98,1999 \* 30/ **√2000**

**=** 197.06 Pound

**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**.

**Ans:** Mean: 41.0

Median: 40.5

Variance: 25.52

Standard Deviation: 5.05

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

**Ans:** The Range of the student marks lies between 34 – 59 ,

The marks is not disperse widely that means its follows the normal distribution pattern.

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

**Ans:** The Data is Symmetric in nature that means the skewness is 0 , we can also say that there is no outlier is present in that data.

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

**Ans:** If the mean > median that means the data is positively skewed the value

Of the skewness is positive , high values are present in the right side of the

data.

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

**Ans:** If the median>mean that means the data is negatively skewed the value

Of the skewness is negative, high values are present in the left side of the

Data

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

**Ans:** Positive Kurtosis indicates that there are higher peaks and more longer

tail than the normal distribution

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

**Ans:** Negative Kurtosis indicates that there are lower peaks than the normal

Distribution

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



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

**Ans:** The lower whisker is much greater than the right whisker that means the data is highly skewed at left side there are so many data points are present at the left side of the data . Also we can say that the data is not normally distributed.

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

**Ans:** The data is skewed at left side majority of the data point is present at the left side of the data.

**What will be the IQR of the data (approximately)?   
Ans:** approximately the IQR of the data is 8 %

Q19) Comment on the below Boxplot visualizations?



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

**Ans:** The range of the data of the first box plot is less compared with the second

Boxplot . Both boxplot having same length of whisker right and left respectively that means the data is following normal distribution.

There are no outlier is present either in the first boxplot or in the second box plot

**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) == 0.3474 or 34 %
  2. P(MPG<40) == 0.72 or 72 %

c. P (20<MPG<50) == 0.89 or 89 %

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:** Yes the data is normally distributed

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

**Dataset: wc-at.csv**

**Ans:** AT column is not following the normal distribution as it right skewed

Waist column is following the normal distribution

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

**Ans:** for 90 % confidence level = 0.90 + 0.05 =0.95 or 95 % i.e 1.644

for 94 % confidence level = 0.94 + 0.03 =0.97 or 95 % i.e 1.880s

for 60 % confidence level = 0.60 + 0.20 =0.80 or 80 % i.e 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-value for 95% confidence level with 24 degree of freedom is 2.6

t-value for 96% confidence level with 24 degree of freedom is 2.17

t-value for 96% confidence level with 24 degree of freedom is 2.79

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:**

Given:

Population Mean=270

Sample Size=18

Sample Average=260

Sample Standard Deviation=90

Degree of freedom is (n-1)=17

H0 : The true average life of a light bulb is 270 days

HA  : The true average life of a light bulb is less than 270 days

So the popular parameter is unknown so we use t test

t = (x̄ - μ) / (s / √n)

t =(260-270)/(90/√18)

t=-0.47

From t value we calculate the probability that is p value

p = 2 \* stats.t.cdf(-0.47, 17)

so the p value = 0.646

now, P value Is greater than the alpha value 0.05 (p> α) we failed to reject the null hypothesis that means the average light bulb is last for 270 days