|  |  |
| --- | --- |
| 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) | Interval |
| 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 | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

**Answer:**

The total possible outcome 23 =8

HHH,HHT,HTT,THT,TTH,HTH,THH,TTT

Number of favourable outcomes = 3

P(two head and one tail) = 3/8 = 0.375

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

1. Equal to 1

**Answer:**

Total number of outcomes = 36

Number of outcomes sum equal to 1=0

p(equal to 1) = 0/36

1. Less than or equal to 4

**Answer:**

Total number of outcomes =36

Number of outcomes less than or equal to 4 = 6

p(less than or equal to 4) = 6/36 =1/6

1. Sum is divisible by 2 and 3

**Answer:**

Total number of outcomes =36

Number of outcomes sum is divisible by 2 and 3 =6

p(sum is divisible by 2 and 3)=1/6

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:**

Number of ways of drawing 2 balls out of 7 n(S)= 7C2 = 21

Number of ways of drawing 2 balls out of 5 n(S) =5C2 = 10

Probability = 10/21

Total number of balls  
= (2 + 3 + 2)  
= 7  
Let S be the sample space  
Then, n(S) = Number of ways of drawing 2 balls out of 7

n(S)=7C2

n(S)=(7×6) / (2×1) = 42/ 2

n(S)= 21

Let E = Event of 2 balls, none of which is blue  
n(E) = Number of ways of drawing 2 balls out of 5 balls

n(E)=5C2

n(E)=(5×4) / (2×1)

n(E)=10

P(E) = n(E) / n(S)

P(E) = 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

**Answer:**

Expected number of candies for randomly selected child

= 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 + 0.240

= 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**

**Answer:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| Mean | 3.596563 | 3.21725 | 17.84875 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Standard Deviation | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.17 | 3.911 | 8.4 |

**Score and Weigh both have outliers**

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:**

Expected Value  =  ∑ ( probability  \* Value )

 ∑ P(x).E(x)

There are 9 patients

Probability of selecting each patient = 1/9

Expected Value  =  (1/9)\*(108) + (1/9)\*110  + (1/9)\*123 + (1/9)\*134 + (1/9)\*135 + (1/9)\*145 + (1/9)\*167 + (1/9)\*187 + (1/9)\*199

= (1/9) \*( 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9)\* ( 1308)

= 145.33

Expected Value of the Weight of that patient = **145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
|  | **speed** | **distance** |
| Skewness | -0.11395477 | 0.78248352 |
| Kurtosis | -0.57714742 | 0.24801866 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| Skewness | 1.58145368 | -0.60330993 |
| Kurtosis | 2.72352149 | 0.81946588 |

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



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



**Answer:** The median is closer to the lower quartile meansdistribution is "Positively Skewed". mean > median.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?

**Answer:**

The **information given** is:

* Sample **mean** x̅ =200
* Sample **standard deviation**  ᵟ=30
* Sample **size** n=2000

The **interval** is:



* In which **t** is the critical value for the two-tailed confidence interval.

Considering a **94%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 1.8916**, hence:

 x̅ - t \* ᵟ/√n = 200-(1.8916\*30/√2000) = 198.73

 x̅ + t \* ᵟ/√n = 200 + (1.8916\*30/√2000) = 201.27

The **94%** confidence interval is **(198.73, 201.27)**

Considering a **96%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.0673**, hence:

x̅ - t \* ᵟ/√n = 200-(2.0673 \*30/√2000) = 198.61

 x̅ + t \* ᵟ/√n = 200 + (2.0673 \*30/√2000) = 201.39

The **96%** confidence interval is **(198.61, 201.39)**

Considering a **98%** confidence level, using a calculator, with 200 - 1 = **199 df**, the critical value is **t = 2.3452**, hence:

x̅ - t \* ᵟ/√n = 200-(2.3452\*30/√2000) = 198.43

x̅ + t \* ᵟ/√n = 200 + (2.3452\*30/√2000) = 201.57

The **98%** confidence interval is **(198.43, 201.57)**

**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**

**Median = 40.5**

**Variance = 25.529412**

**Standard deviation = 5.052664**

1. What can we say about the student marks?

We can say that mean of marks of student is 41 which is slightly greater than median.

Most of the students got marks in between 34-45, there are two outlier 49,56.

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

**Answer:** When the mean is equal to the median then distribution is symmetric and has zero skewness.

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

**Answer:** When mean > median, Skewness is Positive and tail is towards Right

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

**Answer:** When median > mean, Skewness is Negative and tail is towards left

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

**Answer:** A distribution with a positive kurtosis value indicates that the distribution has heavier tails and a sharper peak than the normal distribution.

An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.

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

**Answer:** A distribution with a negative kurtosis value indicates that the distribution has lighter tails and a flatter peak 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:** The above Boxplot is not normally distributed the median is towards the higher value

What is nature of skewness of the data?

**Answer:** The data is a skewed towards left i.e Negative Skewness. The whisker range of minimum value is greater than maximum

What will be the IQR of the data (approximately)?   
**Answer:**   
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.

**Answer:**

There are no outliers. 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 : 0.3475939251582705**
  2. P(MPG<40) **Ans : 0.7293498762151616**

c. P (20<MPG<50) **Ans : 0.8988689169682046**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**MPG of cars does not follow 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

**Adipose Tissue (AT) and Waist follow Normal Distribution**

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

**Answer**:

**For 90% confidence interval**:

We have the significance level at 5 % ( as it is a two tailed test)

that is:

α = 5 % = 0.05

z at α = 0.05 from the z table will be:

**z = 1.645**

**For 94 % confidence interval**:

We have the significance level at 3 % ( as it is a two tailed test)

that is:

α = 3 % = 0.03

z at α = 0.03 from the z table will be:

**z = 1.555**

**For 60 % confidence interval**:

We have the significance level at 20 % ( as it is a two tailed test)

that is:

α =20 % = 0.2

z at α = 0.2 from the z table will be:

**z = 0.253**

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

**Answer:**

t scores of 95% confidence interval = **2.0638985616280205**

t scores of 96% confidence interval = **2.1715446760080677**

t scores of 99% confidence interval = **2.796939504772804**

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

df 🡪 degrees of freedom

**Given:**

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.

**To find:**

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

**Solution:**

t - statistics for the data is given as follows:



x = mean of the sample of bulbs =  260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18









t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.