|  |  |
| --- | --- |
| **­**  **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 |

**Topic: Basic Statistics**

**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 | Ordinal |
| 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**:

* Probability = Number of observations/Total number of observations
* Possible outcomes are {HHH, TTT, HTT, THT, TTH, THH, HTH, HHT}
* Number of observations = 3
* Total number of observations = 8
* Probability = **3/8** or **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**:

* Total outcomes when one die is rolled = {1,2,3,4,5,6}
* Total possible outcomes when two die are rolled = 6 x 6 = 36, i.e.

(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)

(**a**).

* Total outcomes when sum is equal to 1 = 0 , as the minimum outcome possible is (1,1).

(**b**).

* Total outcomes when sum is less than or equal to 4 = (1,1), (2,1), (1,2), (1, 3), (2, 2), and (3,1) = 6
* Total number of observations = 36.
* Probability = 6/36 = **1/6**

(**c**).

* Total outcomes when sum is divisible by 2 and 3 = (1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6) = 5
* Total number of observations = 36.
* Probability = 5/36 or **13.88%.**

**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
* Let S be the sample space.
* Then, n(S) = Number of ways of drawing 2 balls out of 7  
  =7C2​  
  =(2×1)(7×6)​  
  =21
* Let E = Event of drawing 2 balls, none of which is blue.
* n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.  
  =5C2​  
  =(2×1)(5×4)​  
  =10

Hence, P(E)=10/21 ​or **47.61%**

**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**:

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count | Probability | Expected |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.2 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.12 | 0.24 |
| Total expected |  |  | 3.09 |

Total number of expected candies for 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:**

**Table

Description automatically generated**

Note:50% value represents the median in the above table.

**Mode:**

**Table

Description automatically generated**

**Skewness:**

**Word

Description automatically generated with low confidence**

Above results show the Mean, Median, Mode, Variance, Standard Deviation, and Range of the given dataset. We can say that data is bimodal in points and weigh data. Looking at the skewness values , we can say that the data is positively skewed.

**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:**

There are 9 patients in total, therefore

Probability of selecting one patient at random = 1/9

|  |  |  |
| --- | --- | --- |
| Weights of patient (W) | Probability P(X) | X\*P(X) |
| 108 | 1/9 | 12.00 |
| 110 | 1/9 | 12.22 |
| 123 | 1/9 | 13.67 |
| 134 | 1/9 | 14.89 |
| 135 | 1/9 | 15.00 |
| 145 | 1/9 | 16.11 |
| 167 | 1/9 | 18.56 |
| 187 | 1/9 | 20.78 |
| 199 | 1/9 | 22.11 |
|  |  | 145.33 |

Hence, expected value of the weights of patients is 145.33.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:**

**Background pattern

Description automatically generated**

Ignoring the index column, we can see that speed data is slightly negatively skewed whereas the distance data is highly positively skewed.

For the kurtosis, Speed

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans:**

**Application

Description automatically generated with low confidence**

As we can see , SP values are highly positively skewed whereas the WT values are negatively skewed. Also, the kurtosis values for both, SP and WT are positive indicating high peakedness.

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

Shape, rectangle

Description automatically generated



**Ans:** Looking at the histogram we can say that it is negatively skewed indicating Mean > Median > Mode.

From the box plot we can say that Outliers exist on the higher side of the data as the distance between max and 75% is unusually large.

**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

N = 3000000

S = 30

x̄ = 200

Interval estimate = Point estimate ± Margin of error

x̄ ± Z1 – α σ / √n

Graphical user interface, text, application, chat or text message

Description automatically generated

**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:**

**Graphical user interface

Description automatically generated with low confidence**

**Graphical user interface, application

Description automatically generated**

From above plot and information, we can say that mean of marks of student is 41 which is slightly greater than median 40.5.

There are two outliers i.e., 49 and 56.

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

**Ans:** Symmetric in nature.

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

**Ans:** Positively skewed in nature.

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

**Ans:** Negatively skewed in nature.

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

**Ans:** Positive kurtosis value implies that the peak is pointer and tail is thicker.

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

**Ans:** Negative kurtosis value implies that the peak is wider and tail is thinner.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

**Ans:**

* The above box plot show that the median is closer to the upper quartile and the distribution of the data is Skewed.
* We can also say that the data Negatively skewed in nature.
* Q1=10 , Q3=18

IQR of the data = Q3-Q1=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:** As we can see, median of both the plots lies at around 262.5. Though, the median of the plots is same data points of plot 1 are saturated between 245 to 290, whereas, for the second plot, data points are widely spread in the range of 200 to 345. IQR of plot 2 will be greater than plot 1.

**Q 20) Calculate probability from the given dataset for the below cases**

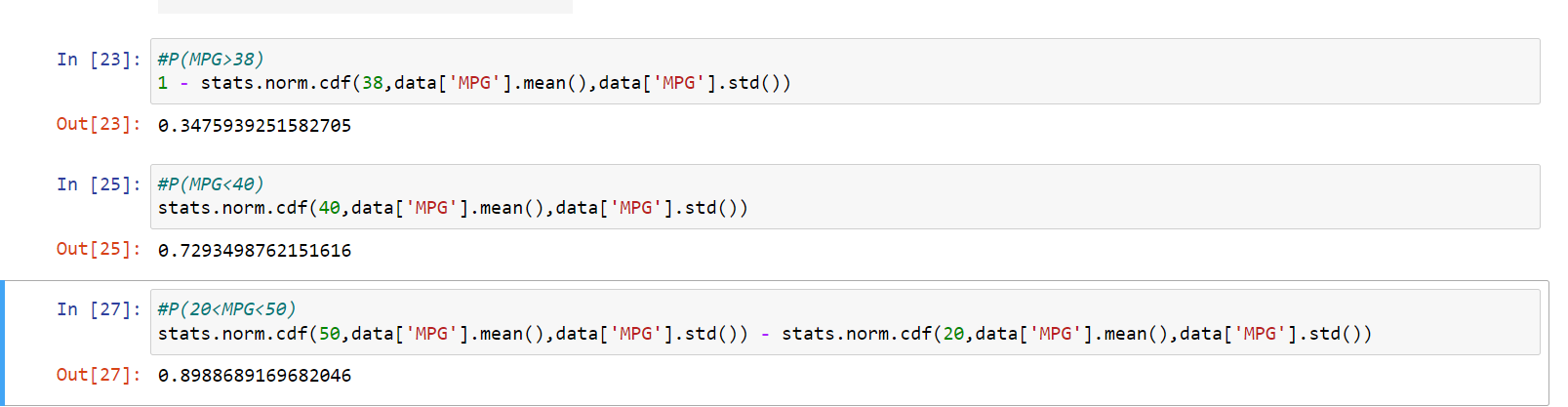
**Data \_set: Cars.csv**

**Calculate the probability of MPG of Cars for the below cases.**

**MPG <- Cars$MPG**

* 1. **P(MPG>38)**
  2. **P(MPG<40)**
  3. **P (20<MPG<50)**

**Ans:**

****

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

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

**Dataset: Cars.csv**

**Ans:**

**Chart

Description automatically generated with medium confidence**

**A picture containing background pattern

Description automatically generated**

As we can see median, mode and mean are not same the data is slightly negatively skewed evident from the above plot.

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

**Dataset: wc-at.csv**

**Ans:** Graphical user interface, application, Word

Description automatically generated

As we can see though the mean and median are not equal , the data does not follow normal distribution.

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

**Ans:**

**Graphical user interface, text, application, email

Description automatically generated**

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

**Ans:**

**Graphical user interface, text, application

Description automatically generated**

**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

Mean of the sample of bulbs = 260

Population mean = 270

Standard deviation of the sample = 90

Number of items in the sample = 18 (square root or 18 = 4.24)

t = (260 – 270) / (90/4.24) = - 0.471

df = 18 - 1 = 17

Table

Description automatically generated with medium confidence

Therefore, 0.322 or 32% is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days.