Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| 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 | **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:**

Answer: When coin is tossed three times **– (2)3 = 8**

The outcomes with two heads and one tail occurs in **– HHT,HTH,THH**

Therefore probability = number of favorable outcomes / total number of possible outcomes

**= 3/8**

Therefore, the probability of getting two head and one tail when three coin are tossed **is 3/8.**

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

1. Equal to 1

**Ans:** There is no outcomes which corresponds sum is equal to one **i.e. 0/36 = Probability is 0**

1. Less than or equal to 4

**Ans:** (1,3) (2,2) (3,1) => 3 outcomes,

3/36 = **1/6**

1. Sum is divisible by 2 and 3

**Ans:** The sums divisible by both **2 and 3 – 6,12**

∴ Sum of 6 : (1,5),(2,4),(3,3),(4,2),(5,1)

Sum of 12 : (6,6)

There are total 6 possible outcomes, and we rolling two dice so

6\*6=36 total outcomes

The probability(sum divisible by 2 and 3) **= 6/36**

= 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?

**Ans:** : **Total number of ways to draw 2 balls out of 7 balls is -7C2**

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

=7\*6\*5\*4\*3\*2\*1 / 2! (5!)

=7\*6 / 2\*1 **= 21**

Now,

Number of ways to draw 2 balls without getting any blue balls.

There are 2 red and 3 green balls which in total is 5 balls.

So, 5C2

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

=5\*4\*3\*2\*1 / 2! (3)!

=5\*4 / 2\*1

**=10**

Finally, the probability(none is blue) = number of ways without blue / total number of ways **= 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 for randomly selected child = 1\*0.015+ 4\*0.20+ 3\*0.65+ 5\*0.005+ 6\*0.01 +2\*0.120 **= 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:** As the Standard Deviation spreads are very low and we can conclude that the values are close to the mean in the Data sets.

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

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:**

Probability of selecting each patient = 1/9

Expected Value = ∑ ( Probability \* Value )

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**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:**

Skewness for speed= -0.117509861.

Skewness value is negative so it is left skewed. Since magnitude is slightly greater than 0 it is slightly left skewed.

And for distance= 0.80689496.

Right skewed (Positive) slight magnitude to right.

**Kurtosis for Speed= -0.50899442.**

**Is negative(flatter than normal distribution)**

**Distance=0.405052582**

**Is positive(peaked than normal distribution)**

**SP and Weight(WT)**

**Use Q9\_b.csv**

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



**Ans:**

The most of the data points are concerted in the range 50-100 with frequency 200.

And least range of weight is 400 somewhere around 0-10.

So the expected value the above distribution is 75.

Skewness- we can notice a long tail towards right so it is heavily right skewed.



**Ans:**

**Median is less than mean right skewed and we have outlier on the upper side of box plot and there is less data points between Q1 and bottom point.**

**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% confidence interval = stats.norm.interval(0.97, loc = 200, scale = 30) [134.897 265.103]**

**98% confidence interval = stats.norm.interval(0.99, loc = 200, scale = 30) [122.725 277.275]**

**96% confidence interval = stats.norm.interval(0.98, loc = 200, scale = 30) [130.21 269.79]**

**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:** **Mean= 41, Median= 40, variance= 24.111, Standard deviation= 4.910**

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

**Ans :**  When the mean and median of a dataset are equal, the skewness of the data is symmetrical or approximately symmetrical. In a perfectly symmetrical distribution, **the mean and median coincide, and the skewness is zero.**

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

**Ans:** When the mean is greater than the median, the distribution is said to be right-skewed or positively skewed. This indicates that there are some relatively high values pulling the mean to the right. The tail on the right side of the distribution is longer compared to the left side.

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

**Ans:** When the median is greater than the mean, the distribution is said to be left-skewed or negatively skewed. This indicates that there are some relatively low values pulling the median to the left. The tail on the left side of the distribution is longer compared to the right side.

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

**Ans:** The data is normally distributed and kurtosis value is **0.**

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

**Ans:** The distribution of the data has lighter tails and a flatter 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?

Let’s assume above box plot is about ages of the students in a school.

50% of the people are above 10 years old and remaining are less.

And students whose age is above 15 are approx. 40%.

What is nature of skewness of the data?

**Ans:** Left skewed, median is greater than mean.

What will be the IQR of the data (approximately)?   
Approximately= **-8**

Q19) Comment on the below Boxplot visualizations?



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

**Ans:**

By observing both the plots whisker’s level is high in boxplot 2, mean and median are equal hence distribution is symmetrical.

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:**

**1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std()) = 0.3475**

* 1. **P(MPG<40)**

**Ans:**

**stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std()) = 0.7293**

* 1. **P (20<MPG<50)**

**Ans:**

**stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std()) = 1.2430**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:**

**Mean= 34.4220757280247 and Median= 35.15272697**

MPG of Cars does follow normal distribution approximately (as mean and median are approx. same)

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

Dataset: wc-at.csv

**Ans:** Waist: mean> median, both the whisker are of same length, median is slightly shifted towards left. Data is fairly symmetrically distributed.

**AT: mean> median, right whisker is larger than left whisker, data is positively skewed.**

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

**Ans: Z score of 90% confidence interval is 1.65**

**Z score of 94% confidence interval is 1.55**

**Z score of 60% confidence interval is 0.85**

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

**Ans: t scores at 95% confidence interval is 2.06**

**t scores at 96% confidence interval is 2.17**

**t scores at 99% confidence interval is 2.8**

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:**

critical value = -0.47

Probability for average life of no more than 260 days is 0.32

t - statistics for the data is given as follows:

t=\dfrac{x-\mu}{\frac{s}{\sqrt n}}

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 = ( 260 – 270 ) / ( 90/sqrt(18) )

t = ( -10 ) / ( 90/3\*sqrt(2) )

t = ( -10 ) / ( 30/sqrt(2) )

t = ( -1 \* sqrt(2) ) / 3

**t = - 0.471**