|  |  |
| --- | --- |
| 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 | Continuous |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Interval |

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

Ans: Given, 3 coins are tossed then the number of all possible outcomes = = 8 outcomes. Let’s ‘S’ be the sample space for all possible outcomes.

P(S) = {HHH, HHT, HTT, TTT, THH, HTH.THT, TTH}

But, for all event that two heads and one tail occurs P(E) = {HHT, THH, HTH}

No. Of outcomes of event (E) = 3

P = =

P

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: Given, two dice are rolled then the no of all possible outcomes = =36 outcomes. Let’s S be sample space for all these possible outcomes.

P(S) = {(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) The sum is equal to 1**

There is nothing possible outcomes that we get, sum is equal to 1.

So, n(E) = 0 (sum of each outcomes exceed to 1)

=

**P(E) = 0**

**b) Sum is less than or equal to 4**

n(E) = {(1,3), (2,2), (3,1)}

n(E) = 3

P =

**P =**

**c) Sum is divisible by 2 and 3**

n(E) = {(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)}

**n(E) =**

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?

A bag contains total number of balls = 2+ 3+ 2 = 7

No. Of ways of drawing 2 balls out of 7 = = = = 21

Number of balls other than blue = 5

No. Of ways of drawing 2 balls out of 5 = = = =10

Probability that name of the balls drawn is blue,

**P(E) = =**

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.**

Let's X be no. Of candies expected number of candies E(X)

Expected candies for child A = n \* p

= 1\*0.015

= 0.015

**Child B – probability of having 4 candies = 0.20**

Expected candies for child B = n\*p

= 4\*0.20

= 0.8

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**

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: To find expected value(Ev)

Expected value = =

There are 9 patients',

Probability of selecting each patient's =1/9

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| E(x) | 108 | 110 | 123 | 134 | 135 | 145 | 167 | 187 | 199 |
| P(x) | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 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)

145.3333

Expected value of the weight of the patient's = 145.3333

**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**



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

And least range of weight is 400 somewere 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 and there are less data points between histogram and bottom points.

**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: Using the t-distribution, it is found that:The 94% confidence interval is (198.73, 201.27).

The 96% confidence interval is (198.61, 201.39).The 98% confidence interval is (198.43, 201.57).  
  
We are given the standard deviation for the sample, which is why the t-distribution is used to solve this question.  
  
The information given is: Sample mean of. Sample standard deviation of . Sample size of .  
  
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:  
  
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:  
  
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:  
  
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.
2. What can we say about the student marks?

Ans: For Mean = X=

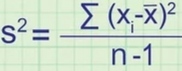
= = 41

For Median Firstly, arrange the data in ascending order,

34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

= = = = 40.5

For sample Variance,

 mean (X) =41, n=18

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 34 | -7 | 49 |
| 36 | -5 | 25 |
| 36 | -5 | 25 |
| 38 | -3 | 9 |
| 38 | -3 | 9 |
| 39 | -2 | 4 |
| 39 | -2 | 4 |
| 40 | -1 | 1 |
| 40 | -1 | 1 |
| 41 | 0 | 0 |
| 41 | 0 | 0 |
| 41 | 0 | 0 |
| 41 | 0 | 0 |
| 42 | 1 | 1 |
| 42 | 1 | 1 |
| 45 | 4 | 16 |
| 49 | 8 | 64 |
| 56 | 15 | 225 |
| sum | 0 | 434 |

For Standard deviation,

The std. Deviation of a set of observation is the positive square root of the variance of the set.

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

Ans: Data is normlized and there is no skewness.

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

Ans: Negative Skewness implies mass of the Distribution concentrated on right side.

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

Ans: Positive skewness implies mass of the Distribution concentrated on left side.

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

Ans: Positive kurtosis value indicates that thinner peak and wider tails.

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

Ans: Negative kurtosis indicates that wider peak and thinner tails.

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



What can we say about the distribution of the data?

Ans: Lets assume above box plot is about age’s of the students in a school. 50% of the people are above 10 yrs old and remaining are less and students who’s 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)?

Ans: 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)
  2. P(MPG<40)

c. P (20<MPG<50)

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

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

Ans: For 90% of the confidence interval:

The significance level at 5% (as it is two tailed test)

α= 5% = 0.05

z at α = 0.05 from the z table will be: z = 1.645

For 94% confidence interval, we get:

The significance level at 3% (as it is a two tailed test )

α= 3% = 0.03

z at α = 0.03 from the z table will be: z = 1.555

For 60% confidence interval, we get:

The significance level at 20% (as it is a two tailed test)

α = 20% = 0.2

z at α = 0.2 from the z table will be: z = 0.253

**Therefore, we get that the z score at 90% confidence interval is 1.645, at** **94% confidence interval is 1.555 and at 60% confidence interval 0.253**

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

Ans: Well, explain in coding file:

|  |  |
| --- | --- |
| Confidence interval | T - score |
| 95% | 2.063899 |
| 96% | 2.171545 |
| 99% | 2.79694 |

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 à pt(tscore,df)

df à degrees of freedom