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 |

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?

Soln:

Total Events:{HHH,HHT,HTT,THH,TTH,TTT,HTH,THT}=8

Interested Events: {HHT, THH, HTH} =3

**P (2head&1tail) =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

Soln:

Total Events :6\*6=36

Total outcomes for 2 dice rolled

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 11 | 21 | 31 | 41 | 51 | 61 |
| 12 | 22 | 32 | 42 | 52 | 62 |
| 13 | 23 | 33 | 43 | 53 | 63 |
| 14 | 24 | 34 | 44 | 54 | 64 |
| 15 | 25 | 35 | 45 | 55 | 65 |
| 16 | 26 | 36 | 46 | 56 | 66 |

Sum of each outcomes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2 | 3 | 4 | 5 | 6 | 7 |
| 3 | 4 | 5 | 6 | 7 | 8 |
| 4 | 5 | 6 | 7 | 8 | 9 |
| 5 | 6 | 7 | 8 | 9 | 10 |
| 6 | 7 | 8 | 9 | 10 | 11 |
| 7 | 8 | 9 | 10 | 11 | 12 |

X:Sum of value of 2 dice rolled

1. **P(X|X=1)=0**
2. **P(X|X<=4)=6/36=1/6= 0.16666667**
3. **P(X|X%2 & X%3)= 1/6= 0.16666667**

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?

Soln: BAG

X: two balls are drawn randomly.

P (X|X are not blue balls)

2R , 2B, 3G

Total events : 7C2 =21

P (X|X are not blue)= P(X|X is red or green)= 5C2/ 7C2 =**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 (X) | 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

Soln:

X: Count of Candies/number of Candies for a random child

E(X): Expected value of X

E(X) = ∑XP(X)

E(X) = ∑XP(X) = 3.09

**Ans. E (P(X)) = 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.

Soln:

Total Records: 32

Records is of Famous Cars and reviews on basis of Points, Score, Weigh based on some Performance factors

Points: all 32 records falls between 4.93-2.76 values. Deviation is least as compare to other parameters, bimodal, mean is closer to median

Score: all 32 records falls between 5.424-1.513 values. Deviation is in between other parameters, unimodal, mean is closer to median

Weigh: all 32 records falls between 22.9-14.5 values. Deviation is most of all parameters, bimodal, mean is closer to median.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | 3.596 | 3.22 | 17.85 |
| **variance** | 0.277 | 0.927 | 3.0934 |
| **std deviation** | 0.526 | 0.963 | 1.7588 |
| **Mode** | 3.07,3.92 | 3.44 | 17.02,18.90 |
| **Range (MAX-MIN)** | 4.93-2.76 | 5.424-1.513 | 22.9-14.5 |
| **Median** | 3.695 | 3.325 | 17.71 |

**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?

Soln:

X: one patient is chosen at random

P(X):1/9

**E(X) = ∑XP(X) = 145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

Soln:

Calculations performed in R studio

|  |  |  |
| --- | --- | --- |
|  | speed | dist |
| Skewness | -0.1106 | 0.75913 |
| Kurtosis | -0.6731 | 0.119397 |

Inference

* speed data set is slight negatively skewed distribution as value is closer to zero having max data saturated on slight middle right side of dataplot and negative kurtosis means peakedness of data is less it has wider peak and thinner tail.
* dist data set on the other hand is positively skewed distribution having max data saturated on left side of data plot and positive kurtosis means peakedness is seen having sharpen peak and fatty tails.

SP and Weight(WT)

Use Q9\_b.csv

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| SP | 1.55226 | 2.58307 |
| WT | -0.5922 | 0.72574 |

Inference:

* SP data set on the other hand is positively skewed distribution having max data saturated on left side of data plot and positive kurtosis means peakedness is seen havindg sharpen peak and fatty tails.
* WT data set is slight negatively skewed distribution as value is closer to zero having max data saturated on slight middle right side of dataplot and positive kurtosis means peakedness is seen having sharpen peak and fatty tails.

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



Soln:

From Histogram, one can see x axis contain continuous variable weights in ranges of 100units and on y axis frequencies in ranges of 50. Max data are saturated on left side and data are skewed on right side so positively skewed distribution of all data. Max data ranges between 50-150.

From boxplot, one can see max data are present in the lower side. Upper whisker is longer than lower. As an imaginary line is drawn over the boxplot starting from lower extreme axis to upper extreme over the inter quartile region we will get a curve which is skewed on upper region So data is skewed along the upper half so it is positively skewed distribution. There are outliers in the data set represented by circles above the upper extreme. (approximately 8)

**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?

Soln:

From Data we get,

N = 3,00,000 , n=2000 , ~~X~~ =200 , s=30

By t distribution method as σ not provided

µ can be calculated from t method

* 1-α = 0.94

t1-α,n-1 = 1.88 from R studio r code --- qt(0.97,1999)

~~X~~ ± t1-α,n-1\*(s/sqrt(n))

Substituting values we get

200+1.26, 200-1.26

201.26, 198.74

Average weight of adult male in mexico has 94 % chance to fall between 201.26-198.74 pounds range for any random sample of 2000 men

* 1-α = 0.98

t1-α,n-1 = 2.33 from R studio r code --- qt(0.99,1999)

~~X~~ ± t1-α,n-1\*(s/sqrt(n))

Substituting values we get

200+1.56, 200-1.56

201.56, 198.44

Average weight of adult male in mexico has 98 % chance to fall between 201.56-198.44 pounds range for any random sample of 2000 men

* 1-α = 0.96

t1-α,n-1 = 2.055 from R studio r code --- qt(0.98,1999)

~~X~~ ± t1-α,n-1\*(s/sqrt(n))

Substituting values we get

200+1.38, 200-1.38

201.38, 198.62

Average weight of adult male in mexico has 96 % chance to fall between 201.38-198.62 pounds range for any random sample of 2000 men

**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.

Soln:

|  |  |
| --- | --- |
| **Mean** | **41.17647** |
| **Median** | **41** |
| **Variance** | **24.96886** |
| **Std Dev** | **4.996885** |

1. What can we say about the student marks?

Soln:

Student marks data set contains students marks which are below first class marks 60% . Total students :18 and concerning their average to below average performance, provisions to be made to improve their marks by providing extra classes, more assignments, attendance improvement to minimize the no. in this range 34-60. This will improve overall average marks and value of median.

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

Soln:

When median=mode=mean there is no skewness and we get normally distributed data having data present on either side of median or mode equally, skewness=0.

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

Soln:

When mean >median , skewness is positive having value greater than zero.

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

Soln:

When mean >median , skewness is negative having value less than zero.

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

Soln:

Positive kurtosis mean sharper peakedness of data along saturated area and fatty tails along the curve.

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

Soln:

Negative kutosis value indicates widening of data along saturation zone and losing peakedness having thinner tails at extremes.

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



What can we say about the distribution of the data?

The data distribution follows negatively skewed as max data are saturated at the right , upper whisker is less than lower one , the inter quartile is towards right where 50% data is present. The Range is from 1-19 and median is 15.2 approx

What is nature of skewness of the data?

The nature of skewness is negatively skewed.

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

The IQR of data ranges from 10-18.1 so IQR is 8.1  
Q19) Comment on the below Boxplot visualizations?



From box plot, here mean= median= mode of both the plots and it follows normal distribution, 50% data are present on either side of mean 1) is the sample dataset 2) Is the population data set.

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

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) =0.3476.

Soln:

From R studio using pnorm()

P(MPG>38)=1 – P(MPG<38) = 1-0.6524=0.3476.

* 1. P(MPG<40) =0.7293

Soln

From R studio using pnorm(),

P(MPG<40)=0.7293

* 1. P (20<MPG<50) = 0.8988689

Soln:

From R studio using pnorm()

P(20<MPG<50) =P(MPG<50) – P(MPG<20)

= 0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Soln :

From Normal QQ plot in R studio of MPG data Cars.csv we infer it follows 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

Soln:

From Normal QQ plot in R studio of AT data and Waist data wc-at.csv we infer they do not follows normal Distribution .

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

Soln:

Zscore of 90% confidence interval from qnorm() in Rstudio : 1.644854

Zscore of 94% confidence interval from qnorm() in Rstudio : 1.880794

Zscore of 60% confidence interval from qnorm() in Rstudio : 0.2533471

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

Soln: from R studio qt() function used

|  |  |  |
| --- | --- | --- |
| 1-α | n | t score |
| **95** | **25** | **2.0639** |
| **96** | **25** | **2.171545** |
| **99** | **25** | **2.7969** |

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

n=18 , ~~x~~ = 260, µ=270 , s=90 ; n= sample size ; ~~x~~ = sample mean

µ = population mean

s= sample standard deviation

tscore = (~~x~~ - µ)/(s/sqrt(n))

we get

tscore= -0.4714

from r studio

p(t at n=18 <-0.4714 ) = 0.3217= 32.17% Hence, if the true bulb life were 270 days, there is a 32.17% chance that the average bulb life for 18 randomly selected bulbs would be less than or equal to 260 days.