Q1) Identify the Data type for the Following :

|  |  |
| --- | --- |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Ordinal |
| Weight of a person | Ratio |
| Weight of Gold | Ratio |
| Distance between two places | Ratio |
| Length of a leaf | Ratio |
| Dog's weight | Ratio |
| Blue Color | Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Ordinal |
| Number of times married | Discrete |
| Gender (Male or Female) | Nominal |

Q2) Identify the Data types, which were among the following

|  |  |
| --- | --- |
| 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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

**sol=>** Let S be the sample space then

S={H,T},{H,T},{H,T} then the event E is {HHT,HTH,THH}

Let X be a random variable denothing the two heads and one tail.

p(X=2)=probability of occurance of 2 heads and 1 tail.

=p(THH)

=1/2\*1/2\*1/2+1/2\*1/2\*1/2+1/2\*1/2\*1/2

**=3/8**

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

* Equal to 1
* Less than or equal to 4
* Sum is divisible by 2 and 3

**sol=>1)**If two dices were rolled,then total possible cases=36

Total Favourable cases (Having sum=1)=0

As minimum sum is 2 for outcome (1,1).

Hence,probability is 0.

**2)**The number of favorable outcomes=3

total number of possibilities=36

probability=The number if favorable outcomes/total number of possibilities

=3/36

=1/12

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?

**sol=>**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

=(7\*6)/(2\*1)

**=21**

Let E=Event of drawing 2 balls, none of which is blue.

therefore, n(E)=Number of ways of drawing 2 balls out of (2+3) balls.

=5c2

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

=10

therefore, p(E)=n(E)/n(S)

**=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

**sol=>**

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.

=>>No case has the variable mean=Median=Mode

Thus a seen in the graph(in Q7.file) dataset "score " and "weigh" has outlier.

**Use Q7.csv file**

**sol=>Inference Drawn:**

* The **mean** is useful for spotting trends in the data because we can compare means over a time period to spot trends. The mean is the most common measure of central tendency.
* The **median** divides a sample of data in half; it is the middle score. The median is a useful statistic if we think our data have some extreme cases. The median is not impacted by extreme cases, but the mean is.

Q8) Calculate Expected Value for the problem below

* 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?

**sol=>**To find expected value,first let the random variable x=thw weights (X) of patients at a clinic (in pounds)

mean/expected value of x

E(x)=Σ xp(x)/n

=108+110+123+134+135+145+167+187+199/9

1308/9

E(x)=145.3333

therefore the expected value of the weight of that patient is 145.3333

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

**Cars speed and distance**

**Use Q9\_a.csv**

**=>"dist" is positively skewed where as "speed" is negatively skewed**

**Thus dist has distribution of data concentrated on the left where as speed has distribution on the right.As seen in the graph(Q9.a) .**

**Both dist and speed has positively kurtosis.**

**sol=>(Q9.a)**

**For Speed**

Skewness = -0.1139548

Kurtosis = 2.422853

The “speed” data set is negatively skewed which means that most of the data points lie on the right side of the curve and tail is extended towards Left side. Moreover it is having a slightly high peak.

**For Distance**

Skewness = 0.7824835

Kurtosis = 3.248019

The “distance” data set is positively skewed which means that most of the data points lie on the left side of the curve and tail is extended towards right side. Moreover it is having high peak.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**=>"SP" is posively skewed where as "WT" is negatively skewed .**

**Thus SP has distribution of data concentrated on the right. As seen in the graph .**

**Both WT and SP has positive kurtosis.**

**sol=>Q9\_b**

**For Speed**

Skewness = 1.552258

Kurtosis = 2.583072

The “speed” dataset is representing negatively skewed distribution, maximum data points lie on the right side of the curve and the tail is extended towards left side as very few data points are lying in the left side. As kurtosis value is positive, which shows that the curve is having slightly higher peak

**For Weight**

Skewness= -0.5921721

Kurtosis= 0.7257402

The “weight” dataset is representing negatively skewed distribution, maximum data points lie on the right side of the curve and the tail is extended towards left side as very few data points are lying in the left side. As kurtosis value is positive, which shows that the curve is having much higher peak. If we analyze the histogram and box plot of the dataset, we can find some outlies in the upper extreme as well as lower extreme.

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



**sol=>**Histogram:

1)Chick weight data is right skewed or positively skewed-yes

2)More than 50% chick weight is between 50 to 150--yes

3)Most of the chick weight is between 50 to 100-- yes

4) The data is positively skewed.



**sol=>**1) The data is positively skewed.

2)There are outliers at upper 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?

=>n=2000

X=200

s=30

= 30

**sol=>**Confidence Interval Estimate= Z => 200 Z

94% Confidence: qnorm(0.97)

[1] 1.880794=Z

200 1.88\* =**198.74 – 201.26**

98% Confidence: > qnorm(0.99)

[1] 2.326348=Z

200 2.33\* =**198.44-201.56**

96% Confidence: > qnorm(0.98)

[1] 2.053749

200 2.05\* = **198.62-201.38**

**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**

* Find mean, median, variance, standard deviation.
* What can we say about the student marks?

**sol=>Find mean,median,variance,standard deviation=**

Mean = 41

Median= 40.5

Variance=25.529

Standard deviation = 5.053

* **What can we say about the student marks?**

Most of the data points lies around the mean but there are some extreme values (outliers) which is much higher than the mean.

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

**sol=>-** Skewness is 0. Symmetric

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

**sol=>**Skewness is positive

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

**sol=>**Skewness is negative.

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

**sol=>**It indicates that the curve is taller than the curve of normal distribution. Tail part does not have much data.

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

**sol=>**It indicates that the curve is flatter and most of the data point lies on the tail portion of the curve.

Q18) Answer the below questions using the below boxplot visualization

* What can we say about the distribution of the data?

**sol=>** The most of the data points are distributed towards right side of the curve.

* What is nature of skewness of the data?

**sol=>** negatively skewed

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

**sol=>**IQR = 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.

**sol=>-** The data is spread across wider range in 2 when compared to 1. Inter Quartile Range of 1 is less than 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

* P(MPG>38)

**sol=>** 0.347487

* P(MPG<40)

**sol=>**0.7294571

c. P (20<MPG<50)

**sol=>**0.148

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**sol=>No** it is not normal distribution. As from the histogram it is clearly mentioned that the curve is not symmetrical and not bell shaped. Rather it is unsymmetrical and flatter in shape

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

Dataset: wc-at.csv

**sol=>Waist Circumference**

Waist Circumference datasets does not follow normal distribution as the curve display in the histogram is not symmetrical and bell shaped. It is slight positively skewed data.

**Adipose Tissue**

Adipose Tissue dataset also does not follow normal distribution as it can be clearly found in histogram and boxplot. Data is positively skewed data.

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

**sol=>**For 90%,

From R,

qnorm(0.95)= 1.644854 = 1.64

For 94%,

qnorm(0.97)= 1.880794 = 1.88

For 60%,

qnorm(0.80)=0.8416212 = 0.84

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

**sol=>**:- n= 25, n-1 =25-1 =24

For 95%,

qt (0.975,24)= 2.063899

For 96%,

qt (0.98,24)=2.171545

For 99%,

qt (0.995,24)= 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 pt(tscore,df)

df degrees of freedom

**sol=>**:- **Considering it as a normal distribution,**

Given,

n = 18,

µ = 270 days,

xbar= 260 days,

S = 90 days

**t-score = -µ/(s/n^(1/2)),**

Putting all the values in above formula,

t-score= -0.471

df =n-1 = 18-1 =17

In R code,

**pt(tscore,df=n-1) = pt(-0.471, 17) = 0.321**

Probability is 32%