|  |  |
| --- | --- |
| 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(If checking on google map is shows x.xxx else it can also be discrete according to road signs) |
| Length of a leaf | Continuous |
| Dog's weight | Continuous |
| Blue Color | Categorical(Nominal) |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Categorical(Nominal) |

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 | Ordinal |
| High School Class Ranking | nominal |
| 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 | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

Probability=no. of favorable outcomes/total outcomes

Total outcomes=2^n=2^3=8

Favorable outcomes=3 (H, H, T), (H, T, H), (T, H, H)

3/8

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

Total number of outcomes for 2 dices= 6^n

N= no. of objects so, 6^2 as dices are 2-> 36

1. Equal to 1

0

As there cannot be a dice with a number 0 in it, so the min sum that we can get it 1 and max is 12

1. Less than or equal to 4

6/36

Pairs that result in a sum less than or equal to 4 are (1,1), (1,2), (1,3), (2,1), (2,2), and (3,1)

1. Sum is divisible by 2 and 3

6/36

Here we need to find pairs that are divisible by both 2 and 3 so those are 6,12

So dices which have sum of 6 and 12 are->

(3,3),(1,5),(5,1),(2,4),(4,2),(6,6)

So 6/36

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?

Total number of balls=2+3+2=7

P randomly picking 2 balls out of 7=> 7C2=7!/2!(7-2)!

P of randomly picking 2 balls where there is no blue ball=> 5C2=5!/2!(5-2)!

P that none of the balls is blue is=>

P of Randomly picking balls where 2 are missing/ P of random picking 2 balls

10/21 or 0.47 Ans

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

3.105

just multiply the candies count with their probability and sum them->

1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120=3.105

Expected Value is nothing but the expected Population mean with enough trials and samples.

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.

A screenshot of a table

Description automatically generated

**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-> 132.56(E(X) of a 1D data is its mean)

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Here as we see->**

**Skewness**

**A screenshot of a computer code

Description automatically generated**

**1) Speed is -ve so the probability distribution is towards the right side but there are some smaller values towards left side hence making out graph left skewed/left tailed.(These values are called outliers)**

**2)Dist is +ve skew, most of the values are on the left side of the mean but there are some outliers which are on the right, Tail is on the right side.**

**Kurtosis**

**A screenshot of a computer

Description automatically generated**

**1) A positive kurtosis means more pointy peak whereas negative means more slack peak, Though this description might sound a bit funny the thing is that with a sharper or slacker peak we the confidence interval of getting a std and mean decreases as the whole Probability distribution shrinks at a +ve kurtosis**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**A screenshot of a computer

Description automatically generated**

**Here what I can infer is that the skewness and kurtosis of SP is high, which means that the graph is right skewed/tailed but the kurtosis peak is at peak which means an outlier on the left side of the graph which needs to be treated.**

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



Rightskewed histogram, We can say that there are a few chicks whose weight is 300-400 but most of the weight lies between 50-200



There are a lot of outliers in the data which needs to be treated,the median is at 30% of the left of the data and long whiskers after erd Quartile so right skewed.

**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?

A white board with writing on it

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.

Mean=41

Median=40.5

Mode=41

Variance=25.52

Std=5.05

Mean =is avg of all values so 34+36+….56/n where n is total number of values

Median= is sort the values in ascending order and find the middle value by n/2 or n+1/2(odd)

Mode= It’s the most frequent element in the list

Variance=(x-xbar)^2/n for all x so it will be (34-41)^2/n +(36-41)^2/n+….

Std= its sqrt of Variance

1. What can we say about the student marks?

We can say that average marks of the students are 41

About 95% of the students lie between 31-51 marks

Mode is 41 so there might be some problems which were difficult for most students

Std of the sample is small if considering 5 marks as small that means the tendency of variation is low which is good for the overall class performance.

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

The data is equally distributed or symmetric i.e no skewness in the data.

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

Right Skewed because the mean is pulled by the larger values towards the right

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

Left skewed

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

It has smaller confidence interval for probability distribution as it has higher peak but wrt to outlier we can say that the data has more outliers around the mean.

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

It indicates that the data has more frequency of values at the center around the mean.

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



What can we say about the distribution of the data?

* Here we can Infer that the right side 25% has high count of values whereas the left side of the median is more distributed., median is about 15,min is 2, max is 19, 25th percentile is 10, 50th percentile is median, 75th percentile is 18

What is nature of skewness of the data?

Negatively skewed

What will be the IQR of the data (approximately)?   
  
that will be 75th percentile -25th percentile i.e 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.

Box 1 and box 2 have same median but the distribution in box 1 is smaller whereas in box 2 its more distribution or variation of values

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) -> 40.7%
  2. P(MPG<40) -> 75.3%

c. P (20<MPG<50)-> 85.18%

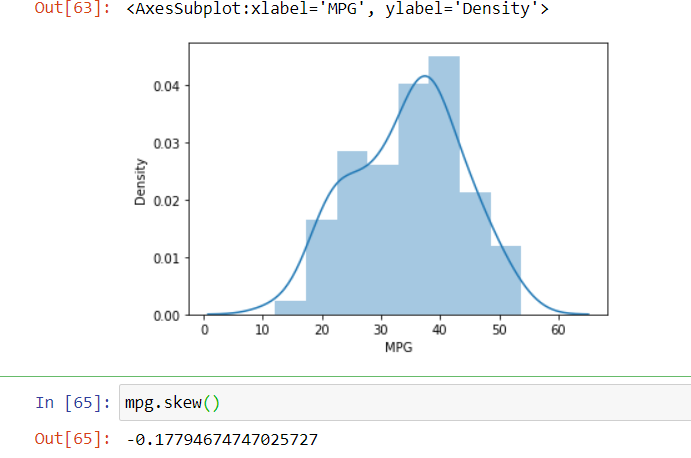
A screenshot of a computer

Description automatically generated

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



It does follow almost normal distribution, for practical reasons we consider anything. For theoretical cases the skewness 0 is considered Normal distribution but for practical reasons we consider anything between -0.5 to 0.5 as 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

Waist is bimodal and not normal, even if skewness is less and its symmetrical we cant consider this as normal distribution as it has 2 modes

A graph with a line

Description automatically generated

A screenshot of a computer screen

Description automatically generated

The AT also is not a normal distribution because first of all its not symmetrical and has 2 peaks, skew is positive and doesn’t form bell curve.

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

A screenshot of a computer

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

Df=Degree of freedom=n-1-> 25-1=24

1-alpha=P(X=x)

A screenshot of a computer

Description automatically generated

95% confidence interval->2.064

96% confidence interval->2.171

99% confidence interval->2.797

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

A white board with writing on it

Description automatically generated

A screenshot of a computer

Description automatically generated