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

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 Data |
| High School Class Ranking | Ordinal Data |
| Celsius Temperature | Interval Data |
| Weight | Ratio Data |
| Hair Color | Nominal Data |
| Socioeconomic Status | Orderly Data |
| Fahrenheit Temperature | Interval Data |
| Height | Ratio Data |
| Type of living accommodation | Orderly Data |
| Level of Agreement | Orderly Data |
| IQ(Intelligence Scale) | Orderly Data |
| Sales Figures | Ratio Data |
| Blood Group | Nominal Data |
| Time Of Day | Nominal Data |
| Time on a Clock with Hands | Orderly Data |
| Number of Children | Ratio Data |
| Religious Preference | Nominal Data |
| Barometer Pressure | Interval Data |
| SAT Scores | Ratio Data |
| Years of Education | Ratio Data |

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

Sol. Total outcomes possible = 8

Favourable outcomes = 3 i.e (HHT), (THH), (HTH)

Probability= Favourable outcomes/Total outcomes=3/8

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

1. Equal to 1 Sol. The probability is Zero
2. Less than or equal to 4 Sol. The probability is 6/36 or 1/6
3. Sum is divisible by 2 and 3 Sol. The probability is 24/36

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?

Sol. Probability=20/21

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. The expected number of candies is = 3.09 candies

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.

Sol.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | mean | median | mode | variance | Std.dev | Range |
| Points | 3.596563 | 3.695 | 3.07 | 0.285881 | 0.53467874 | 2.17 |
| Score | 3.21725 | 3.325 | 3.44 | 0.957379 | 0.97845744 | 3.911 |
| Weigh | 17.84875 | 17.71 | 17.02 | 3.193166 | 1.78694324 | 8.4 |
|  |  |  |  |  |  |  |

We can say that the weigh data is “Right Skewed” as mode<median<mean

We can also conclude that score data is “Left Skewed” as mean<median<mode

Points data is neither left nor right skewed.

**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?

Sol.

|  |  |  |
| --- | --- | --- |
| 108 | 0.111111 | 12 |
| 110 | 0.111111 | 12.22222 |
| 123 | 0.111111 | 13.66667 |
| 134 | 0.111111 | 14.88889 |
| 135 | 0.111111 | 15 |
| 145 | 0.111111 | 16.11111 |
| 167 | 0.111111 | 18.55556 |
| 187 | 0.111111 | 20.77778 |
| 199 | 0.111111 | 22.11111 |
|  | exp.value | 145.3333 |

The expected value is 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
|  | skewness | Kurtosis |
| Speed | -0.11751 | -0.50899 |
| Dist | 0.806895 | 0.405053 |

The speed data has negative skewness value in between -0.5 to +0.5 so it is approximately symmetrical with a longer tail on the left.

The distance data is positively skewed and its value lies in between 0.5 to +1 so, it is moderately skewed with a longer tail on the right.

The kurtosis value for the speed data is negative and hence less than 3 so it is platokurtic data.

The kurtosis value for the distance data is positive so its distribution is peaked and platokurtic data.

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | skewness | Kurtosis |
| SP | 1.61145 | 2.977329 |
| WT | -0.61475 | 0.950291 |

SP is highly skewed to the right(positively skewed) and has a kurtosis value less than 3 so it platykurtic data.

WT has skewness between -0.5 to -1 so it is moderately skewed.

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





Sol. The histogram is positively skewed to the right. The boxplot has 7 outlier values. The data below the median is concentrated and above the median there is large amount of variation in the chickweight data.

**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?

Sol. We have sample size(n)=2000, standard deviation=30, mean=200 pounds, Degree of freedom=(n-1) =1999, 

This is the output we obtain from python.

**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 | Variance | std dev | Median |
| 40.44444 | 13.43791 | 3.665775 | 40.5 |
|  |  |  |  |

1. What can we say about the student marks?

Sol. When we calculate the mode it comes 41. So, the mean, median and mode are almost same. So, the student data is perfectly symmetrical distributed.

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

Sol. The data has zero skewness and is symmetric.

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

Sol. It is right skewed data.

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

Sol. It is left skewed data.

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

Sol. This tells us that there is a peak in the distribution, it has outliers.

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

Sol. This tells us that the distribution is flat with less outliers.

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



What can we say about the distribution of the data?

Sol. To the left of the median there is variation in the data and to the right there is more concentration of data or we say the data is consistent.

What is nature of skewness of the data?

Sol. The above data is skewed to the left. There are more values to the left side as the left whisker is longer when compared to the right whisker.

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

Sol. The IQR value between 8 to 8.5   
  
  
Q19) Comment on the below Boxplot visualizations?



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

Sol. The boxplot 2 has a greater variation inn data when compared to boxplot 1.

The boxplot 1 is more consistent than boxplot2.

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)

Sol. Mean of the data MPG= 34.42208, std.dev=9.131445

a) we have z value=(38-mean)/std.dev=(38-34.42208)/ 9.131445=0.39

from the z table we get the prob value till 0.39=0.6517

so, required probability = 1-0.6517=0.35

b) we have z value=(40-mean)/std.dev=(40-34.42208)/ 9.131445=0.61

from the z table we get the prob value till 0.61 = 0.7291

so, required probability = 0.7291

c) we have for x=2, z value=(20-mean)/std.dev=(20-34.42208)/ 9.131445=-1.58

from the z table we get the prob value till -1.58 = 0.05705

again for x=50, z value=(50-mean)/std.dev=(50-34.42208)/ 9.131445=1.71

probability for 1.71= 0.95637

so finally req.value P(20<MPG<50)=( 0.95637-0.05705)=0.90

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

|  |  |
| --- | --- |
| *Column1* | |
|  |  |
| Mean | 34.42208 |
| Standard Error | 1.014605 |
| Median | 35.15273 |
| Mode | 29.62994 |
| Standard Deviation | 9.131445 |
| Sample Variance | 83.38328 |
| Kurtosis | -0.61168 |
| Skewness | -0.17795 |
| Range | 41.59942 |
| Minimum | 12.10126 |
| Maximum | 53.70068 |
| Sum | 2788.188 |
| Count | 81 |
|  | 0 |

If we look at the value of mean and the median for the MPG data, they are almost same. Also, the value of skewness lies in the interval (-0.5 to +0.5) so it is moderately symmetrical. So,we can say that the MPG data is normally distributed

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

Dataset: wc-at.csv

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Waist |  |  | AT |  |
| *Column1* | |  | *Column1* | |
|  |  |  |  |  |
| Mean | 91.90183 |  | Mean | 101.894 |
| Standard Error | 1.298728 |  | Standard Error | 5.487843 |
| Median | 90.8 |  | Median | 96.54 |
| Mode | 94.5 |  | Mode | 121 |
| Standard Deviation | 13.55912 |  | Standard Deviation | 57.29476 |
| Sample Variance | 183.8496 |  | Sample Variance | 3282.69 |
| Kurtosis | -1.10267 |  | Kurtosis | -0.28558 |
| Skewness | 0.134056 |  | Skewness | 0.584869 |
| Range | 57.5 |  | Range | 241.56 |
| Minimum | 63.5 |  | Minimum | 11.44 |
| Maximum | 121 |  | Maximum | 253 |
| Sum | 10017.3 |  | Sum | 11106.45 |
| Count | 109 |  | Count | 109 |
|  | 0 |  |  | 0 |

The waist data is normally distributed, the values of the mean and median appear to be almost same. The skewness value between -0.5 to 0.5 which makes the waist data almost symmetrical.

The AT data has a skewness value between 0.5 to 1, The value of mean and median also differ from each other.so, the AT data is not normally distributed.

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

Sol. For the confidence interval 90% we have value =(1+CL)/2=0.95

So, we have the value from the z-score table as 0.82894

For the confidence interval 94% we have the value =(1+CL)/2=0.97

So, we have the value from the z-score table as 0.83398

For the confidence interval 90% we have value =(1+CL)/2=0.8

So, we have the value from the z-score table as 0.78814

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

Sol. We have the sample size=25, Degree of Freedom=24,

t score(95%confidence) =2.064, t score (96% confidence) = 2.492,

t score(99% confidence)=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

Sol. We have n=18 bulbs, std.dev=90 days, sample mean=260 days, other mean value=270, Degree of Freedom=17;

the t value = (given mean-sample mean)/(std.dev/sqrt(n)) = 0.47

In R software using the above code we get the value as 0.6778361