|  |  |
| --- | --- |
| 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 | nominal |
| Number of kids | discrete |
| Number of tickets in Indian railways | discrete |
| Number of times married | discrete |
| Gender (Male or Female) | 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 | nominal |
| High School Class Ranking | ordinal |
| Celsius Temperature | interval |
| Weight | nominal |
| Hair Color | nominal |
| Socioeconomic Status | nominal |
| Fahrenheit Temperature | interval |
| Height | nominal |
| Type of living accommodation | ordinal |
| Level of Agreement | nominal |
| IQ(Intelligence Scale) | ratio |
| Sales Figures | ratio |
| Blood Group | interval |
| Time Of Day | ratio |
| Time on a Clock with Hands | ordinal |
| Number of Children | ordinal |
| Religious Preference | ratio |
| Barometer Pressure | ratio |
| SAT Scores | ratio |
| Years of Education | internal |

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

* Favourable outcomes are 3
* Favourable/total outcomes i.e 3/8
* Probability = 0.375

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: when 2 dices are rolled possible outcomes are 36
* a) Equal to 1; There is no such outcome, as minimum sum of 2 dices is 2(1,1)
* Therefor
* 1.
* e probability of getting the sum =1 is '0'.
* b) Less than or equal to 4; there are only 6 outcomes whose sum is equal to 4 or less than 4:{1,1; 1,2; 1,3; 2,1; 2,2; 3,1}
* So by formula, it is 6/36 i.e 1/6.
* c) Sum is divisible by 2 and 3; there are only 6 outcomes whose sum is
* divisible by both 2 and 3 i.e {1,5; 2,4; 3,3; 4,2; 5,1; 6,6}
* So by formula it is 6/36 i.e 1/6."

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 there are 7 balls(2+3+2) in the bag: n!/r!(n-r)! 7C2 = 7\*6/2= **21.**
* But we don’t want blue balls so 7-2=5 i.e5C2= 5\*4/2=**10**
* So,by formula= 10/21,so the probability that the ball is not blue is **10/21=0.47**

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

* Expected no of candies for randomly selected will be given by:

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

i.e: 0.015+0.8+1.95+0.025+0.06+0.24=3.090

= **Expected number of candies for a randomly selected child  = 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.

* Points scores weigh
* Mean = 3.5965625 3.21725 17.84875
* Median= 3.695 3.325 17.71
* Mode= 3.92 3.44 17.02
* Range= 0.534678736 0.978457443 1.786943236
* Standard deviation= 0.534678736 0.978457443 1.786943236
* Variance= 0.285881351 0.957378968 3.193166129

I would say that the average is 3.5965625 of dataset of points so the car models must be somewhere around 3.

In the above dataset the median value is odd so that it is easy to identify it.

In Range ,standard deviation and variance it requires the formulas so it is a two step process value.

|  |  |
| --- | --- |
|  |  |

**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

* Expected value= Summation(prob\*value)

There are 9 patients, so probability of selecting one patient=1/9

So, (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

145.33

**Expected Value of the Weight of that patient is 145.33**

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

**Cars speed and distance**

* **Skewness of car speed -0.1139548**
* **Skewness of car distance 0.7824835**
* **Kurtosis of car speed 2.422853**
* **Kurtosis of car speed 3.248019**

**Use Q9\_a.csv**

**SP and Weight(WT)**

* **Skewness of SP= 1.67545**
* **Skewness of WT=-0.63103**
* **Kurtosis of SP=3.24148**
* **Kurtosis of WT=0.941127**

**Use Q9\_b.csv**

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

**Ans- This histogram of Chickweight and weight are right skewed means all the larger amount of data are on the left side highest frequency is 200 of this histogram.**

**Because of the right skewed after the highest peak there are decrease in the histogram graph.This is asymmetric data,data are not equally distributed.This is vertical boxplot**

**In the boxplot we can clearly see there are some outlier present in this data set,so the outliers are present in the dataset they can cause problem error or sampling.**



* The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side.

-> The boxplot has outliers on the maximum 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?

**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,variance=25.52941,Standard deviation=5.052941

1. What can we say about the student marks?

* Student marks are in accending order,as the student solving more test the marks are improving slowly.
* First test obtained by this student is 34 after 17 tests he improve marks by 22 marks i.e 56 marks

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

* Nature of skewness if mean and median are equal in data that it is called normal distribution because in data the values of data are equally distributed in both right and left side of the bar graph that’s why it is called normal distribution.The distribution has zero skewness

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

* Nature of skewness when mean is greater then mode is called positive skewness because the highest peak of the graph is in the left side of the mean.
* This is right skewness

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

* The nature of skewness when median is greater then mean is called negative skewness because the highest peak of the bar graph is in the right side of graph.
* This left skewness

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

* In positive kurtosis at first the value is lowest at the x axis,when it reaches the highest bar we see the decrease in value of bars.
* The peakedness in positive kurtosis are in center of the bar graph then mean and median are equal.The highest bar is at the center.

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

* It indicates in symmetric distribution bar graph that the measure of peakedness are in between the peaks and thinner tails.
* The measurement of bar are going to be in between the peaks of the graph.

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



What can we say about the distribution of the data?

* MOST OF THE DATA IS ON THE LEFT SIDE OF THE BOX. MEDIAN IS ABOUT 15

What is nature of skewness of the data?

* **POSITIVELY SKEWED**

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

* **18-10=8(UQ-LQ)**

Q19) Comment on the below Boxplot visualizations?



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

* here are no outliers and both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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) -> Probability of MPG>38 is 0.4074
  2. P(MPG<40)->probability of MPG<40 is 0.7530

P (20<MPG<50) -> 0.89899

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

* Yes it is normally distributed

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

Dataset: wc-at.csv

* For Waist- yes, it is normally distributed

For AT: It is right skewed and does not follow normal distribution

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

* for 90%= 1.6448
* For 94%=1.8807
* For 60%=0.8416

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

* T-score for sample size of 25-95% 2.0638
* 96%- 2.1715
* 99%-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

t**=**(260**-**270)**/**(90**/**18**\*\***0.5)=-0.4714

p\_value**=**1**-**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)=0.32167