|  |  |
| --- | --- |
| 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* |

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 | *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 | *Interval* |
| SAT Scores | *Interval* |
| Years of Education | *Ratio* |

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

*Total possible outcomes = 23 = 8*

*Possibilities of 2 heads and 1 tail = [HHT,HTH,THH] = 3*

***So probability = 3/8***

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

Total possible outcomes = (possible numbers)No. of dice = 62 = 36

1. Equal to 1

*Probability = 0*

1. Less than or equal to 4

*Possible outcomes = [(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)]*

*So probability = 6/36 = 1/6*

1. Sum is divisible by 2 and 3

*It means sum should be divisible by 6*

*Possible outcomes = [(1,5)(2,4),(3,3),(4,2),(5,1),(6,6)]*

*So probability = 6/36 = 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 outcomes = 7C2 = 7!/(2!\*5!) =* ***21***

*Total outcomes for no blue = ignore 2 blue balls = take random 2 balls from 2 red and 3 green = 5C2 = 5!/(2!\*3!) = 120/12 =* ***10***

*So, probability of getting no blue =* ***10/21 = 0.476***

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 number of candies for a randomly selected child = ∑*P*(*Xi*​)×*Xi*​​  
= (1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120) = **3.09**

I copied the 2 columns in excel and used the sumproduct function, instead of calculating manually

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.

**Use Q7.csv file**

***Please check attached assignement1-Q7.ipynb***

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?

*Expected value = sum(probability\*value)=(1/9)\*108+…….+(1/9)\*199*

*= 1/9 \* (sum of all weights) = m,ean =* ***145.33***

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

**Cars speed and distance**

**Use Q9\_a.csv**

***Please check Assignment 1 - Q9a for code***

**SP and Weight(WT)**

**Use Q9\_b.csv**

***Please check Assignment 1 - Q9b for code***

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



1. *This histogram is right skewed*
2. *The population size is appx. around 600 Chicks, 200 of which is having weight at around 50-100*



*Interpretation is :*

1. *All 4 parts of the plot are uneven, means in quartile grp 1 and 2, data is very close and similar, but in quartile grp 3 data is different and in quartile grp 4 it is very much different.*
2. *Outliers are available in upper whisker*

**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?

As sample size is greater than 30 and STD is known we will go with Z value.

Formula for confidence interval is (mean +/- (Z \*STD/root of n))

94% confidence interval

100-94 = 6

6/2 = 3

94 + 3 = 97

Look for 0.97 value in Z table = 1.89

Conf interval for 94%= 200+/- (1.89\*30/root of 2000) = 200+/- 1.2679

***So Conf interval for 94% is (198.73, 201.2679)***

***OR Using Python = stats.norm.interval(0.94,200,30/np.sqrt(2000))***

Same way found the conf interval for 98%

Conf interval for 98%= 200+/- 2.33\*30/root of 2000) = 200+/- 1.563

***So Conf interval for 98% is (198.437, 201.563)***

***OR Using Python = stats.norm.interval(0.98,200,30/np.sqrt(2000))***

Conf interval for 96%= 200+/- 2.06\*30/root of 2000) = 200+/- 1.3819

***So Conf interval for 96% is (198.6181, 201.3819)***

***OR Using Python = stats.norm.interval(0.96,200,30/np.sqrt(2000))***

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

2 students have got better marks as 49 and 56

**Please check attached Assignment 1 Q 12**

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

*Skewness is 0*

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

*Positive skewness*

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

*Negative skewness*

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

*Most of the data is located at the tails, rather than mean*

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

*Not sure about interpreting negative kurtosis, may be data is more deviated from mean*

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

*Negative skewness – left skewed*

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

*The data in first boxplot is less distributed and closer to the mean. The data in the second boxplot is much distributed and deviated from mean. Both have the same mean.*

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) = 33/81
  2. P(MPG<40) = 61/81

c. P (20<MPG<50) = 69/81

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Please check attached Assignment 1 Q 21**

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

Dataset: wc-at.csv

**Please check attached Assignment 1 Q 21**

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

Consider

Z score of 90% CI : stats.norm.ppf(0.95)

Considering default two tailed distribution

Z score of 90% confidence interval = stats.norm.ppf(0.95) = 1.645

Z score of 94% confidence interval = stats.norm.ppf(0.97) = 1.88

Z score of 60% confidence interval = stats.norm.ppf(0.80) = 0.8416

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

Df = 25-1 = 24

stats.t.ppf(0.95,24) = 1.71

stats.t.ppf(0.96,24) = 1.828

stats.t.ppf(0.99,24) = 2.49

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

Given :

Population mean µ = 270

Sample size n= 18

Sample mean x= 260

Sample std s= 90

As sample size is less than 30, we will go with t dist.

t for given data = (x-µ)/(s/√n) = (260-270)/(90/4.24) = -0.47

P(x=260) = pt(-0.47,17) = 0.322

So answer = 0.322