|  |  |
| --- | --- |
| 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 | Nominal |
| 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?

Ans3) The tossing of three at a time. These are the number of possible outcomes [HHH,THH,TTH,HTT,HHT,THT,HTH] So out of these outcomes we have the probability of getting Two heads and One tail is 3/8=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 2and 3

Ans4) total possible outcome =62=36=62=36

a) Favorable outcome *(sum equal to 1*)**= 0***{i.e. not possible that sum always exceed to 1}*

Required probability=036 =0=036 =0

b) Favorable outcome*(sum equal to 4)* **= 3***{i.e. (1,3)(2,2)(3,1)}*

Required probability =336 =112336 =112

c) Favorable outcome *(sum less than 13)***= 36** *{i.e. there are maximum sum is 12 of (6,6)}*

Required probability =3636 =1

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?

Ans5) Probability of selecting 2 balls such that none of them is blue.

Here are total 7 balls.

5 balls are not blue.

Select 2 balls from those 5 balls (these are the favorable cases)

Select 2 balls from those 7 balls (these are total cases)

P = 5c2/7c2

P= 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

Ans6) Expected number of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12

= 0.015 + 0.8  + 1.95 + 0.025 + 0.06 + 0.24

=      3.090

=  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.

**Use Q7.csv file**

**Ans7)**

|  |  |  |  |
| --- | --- | --- | --- |
| Mean | 3.596563 | 3.217250 | 17.848750 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Median | 3.695000 | 3.325000 | 17.710000 |
|  | Mode | 3.920000 | 3.440000 | 17.020000 |
|  | S.D | 0.534679 | 0.978457 | 1.786943 |
|  | Variance | 0.285881 | 0.957379 | 3.193166 |

df.describe()

(#will describe the data sets by providing values like mean, std, max and min, etc)

df.mode()(#will show the mode fo each column)

df.var() (#will provide the variance for given dataset for columns with numerical values)

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?

Ans8) Ans: EV=Σx/n =(108+110+123+134+135+145+167+187+199)/9=145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

****

Ans9)

print('skewness value for speed and distance is', np.round(df1.speed.skew(), 2),

'and', np.round(df1.dist.skew(), 2), 'respectively')

skewness value for speed and distance is -0.12 and 0.81 respectively

print('Kurtosis value for speed and distance is', np.round(df1.speed.kurt(), 2),

'and', np.round(df1.dist.kurt(), 2), 'respectively')

Kurtosis value for speed and distance is -0.51 and 0.41 respectively

Inferences: as you can see from the above data, there is a huge difference in the kurtosis values when e1071 and moments package are compared with each other. This is due to different equations used by the packages to find kurtosis.

**Q.9b)**

****

**SP and Weight(WT)**

Ans9)

print('skewness value for SP and WT(weight) is', np.round(df2.SP.skew(), 2),

'and', np.round(df2.WT.skew(), 2), 'respectively')

skewness value for SP and WT(weight) is 1.61 and -0.61 respectively

print('Kurtosis value for SP and WT(weight) is', np.round(df2.SP.kurt(), 2),

'and', np.round(df2.WT.kurt(), 2), 'respectively')

Kurtosis value for SP and WT(weight) is 2.98 and 0.95 respectively

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



Ans10) The given histogram is Right skewed

The given histogram is in bell curve

The distribution is normal



Ans10) The above boxplot suggests that the distribution has lots of outliers towards upper extreme

**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?

Ans11) Confidence interval for 94% is [198.73-201.26]

Confidence interval for 96% is [198.62-201.37]

Confidence interval for 98% is [198.43-201.56]

**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?

Ans12) df = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

df = pd.DataFrame(df)

df.mean()

0 41.0

df.median()

0 40.5

df.std()

0 5.052664

df.var()

0 25.529412

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

Ans13) Skewness = 0, Symmetric

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

Ans14) Right Skewed

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

Ans15) Left Skewed

Q16) What does positive kurtosis value indicates for adata ?

Ans16) Sharp Peak, Thick Tails

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

Ans17) Broad Peak, Thin Tails

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

Ans17)



What can we say about the distribution of the data?

* Not a Normal Distribution

What is nature of skewness of the data?

* Left Skewed

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

* 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.

Ans19)1) The median of the two boxplots are same approximately 260.

2) The boxplots are not skewed in +ve or –ve direction.

3) Outliers doesn’t exist in both of the boxplots.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

Ans20)

print("Probabilty that 'MPG' > 38 = ", np.round(1-stats.norm.cdf(38, loc = df4.MPG.mean(), scale = df4.MPG.std()), 3))

Probabilty that 'MPG' > 38 = 0.348

print("Probabilty that 'MPG' < 40 = ", np.round(stats.norm.cdf(40, loc = df4.MPG.mean(), scale = df4.MPG.std()), 3))

Probabilty that 'MPG' < 40 = 0.729

print("Probabilty that 20 <'MPG' < 40 = ", np.round((1-stats.norm.cdf

(20, loc = df4.MPG.mean(), scale = df4.MPG.std())) -

(stats.norm.cdf(40, df4.MPG.mean(), scale = df4.MPG.std())) , 3))

Probabilty that 20 <'MPG' < 40 = 0.214

print("Probabilty that 'MPG' < 40 = ", np.round(stats.norm.cdf(70, loc = 60, scale = 10), 5))

Probabilty that 'MPG' < 40 = 0.84134

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans21) import statsmodels.api as smf

import pylab as py

smf.qqplot(df["MPG"],line='45')

py.show()



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

Dataset: wc-at.csv

Ans)

import statsmodels.api as smf

import pylab as py

smf.qqplot(df5["Waist"],line='45')

py.sho



Ans)

import statsmodels.api as smf

import pylab as py

smf.qqplot(df5["AT"],line='45')

py.show()



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

Ans22) Z value for 90% confidence interval is 1.645

Z value for 94% confidence interval is 1.88

Z value for 60% confidence interval is 0.842

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

Ans 23)

95% Confidence:

>qt(0.975,df=24)

2.063899

96% Confidence:

>qt(0.98,df=24)

2.171545

99% Confidence:

>qt(0.995,df=24)

2.79694

Q 24**)**A Government companyclaims 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

Ans24) x = mean of the sample of bulbs =  260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

The probability of the bulbs lasting less than 260 days on average of 0.3218