Q1) Identify the Data type for the Following:

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

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

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

**ANS**. While tossing three coins the probability that two heads and one tail are obtained is 3/8 because we have total 8 possible outcomes out of which there are only 3 possibilities to come HHT,THH,HTH

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

1. Equal to 1

**ANS.**

If we rolled two dice then there is no probability that sum is equals to 1.So its 0/36=0 chances.

1. Less than or equal to 4

**ANS.**

When we rolled 2 dice, there are 3/36 chances that sum is equals to 4.

Total possibilities=36

Out of them getting 4= 2+2, 1+3, 3+1

So it will be 1/6\*1/6 + 1/6\*1/6 + 1/6\*1/6 =3/36

1. Sum is divisible by 2 and 3

**ANS.**

When sum is divisible by 2 and 3 possible outcome will be:-

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

So, its 5/36

**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?

**ANS.**

P=5C2/7C2

5/2!\*3!/7/2!/5!

10/21

So there is 10/21 chances that none of the balls drawn is blue

**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

**ANS.**

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\*0.120

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

= 3.090

**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**

**ANS.** df.mean() # It gives us the average/central idea of our data set. These values are representing out entire dataset.

Points 3.596563

Score 3.217250

Weigh 17.848750

dtype: float64

df.median() # It represents the mid value. Mostly we impute median when there is outlier in dataset because median doest not get affected by outliers.

Points 3.695

Score 3.325

Weigh 17.710

dtype: float64

df.mode() # the most occurred value is called mode.

Unnamed: 0 Points Score Weigh

0 AMC Javelin 3.07 3.44 17.02

1 Cadillac Fleetwood 3.92 NaN 18.90

2 Camaro Z28 NaN NaN NaN

3 Chrysler Imperial NaN NaN NaN

4 Datsun 710 NaN NaN NaN

5 Dodge Challenger NaN NaN NaN

6 Duster 360 NaN NaN NaN

7 Ferrari Dino NaN NaN NaN

8 Fiat 128 NaN NaN NaN

9 Fiat X1-9 NaN NaN NaN

10 Ford Pantera L NaN NaN NaN

11 Honda Civic NaN NaN NaN

12 Hornet 4 Drive NaN NaN NaN

13 Hornet Sportabout NaN NaN NaN

14 Lincoln Continental NaN NaN NaN

15 Lotus Europa NaN NaN NaN

16 Maserati Bora NaN NaN NaN

17 Mazda RX4 NaN NaN NaN

18 Mazda RX4 Wag NaN NaN NaN

19 Merc 230 NaN NaN NaN

20 Merc 240D NaN NaN NaN

21 Merc 280 NaN NaN NaN

22 Merc 280C NaN NaN NaN

23 Merc 450SE NaN NaN NaN

24 Merc 450SL NaN NaN NaN

25 Merc 450SLC NaN NaN NaN

26 Pontiac Firebird NaN NaN NaN

27 Porsche 914-2 NaN NaN NaN

28 Toyota Corolla NaN NaN NaN

29 Toyota Corona NaN NaN NaN

30 Valiant NaN NaN NaN

31 Volvo 142E NaN NaN NaN

df.var() # with this we will get to know the variation/flutuation in our dataset. This fluctuation is only can be measured in quantity and we denote it as variance. It simply tells us how far each data point is from mean.

Points 0.285881

Score 0.957379

Weigh 3.193166

df.std() # It’s simply the square root of variance that we have calculated just before.

Points 0.534679

Score 0.978457

Weigh 1.786943

dtype: float64

**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. Expected Value  =  sigma( probability  \* Value )

 sigmaP(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

Ex  108, 110, 123, 134, 135, 145, 167, 187, 199

P(x)  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

Expected Value  =  (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 = 145.33

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

Cars speed and distance

Use Q9\_a.csv

SP and Weight(WT)

Use Q9\_b.csv

**ANS.**

Before imputing skewness and kurtosis, we first need to calculate Mean,Std dev and Mode

**SK = mean – mode/std dev**

**df.skew()**

Points 0.292780

Score 0.465916

Weigh 0.406347

dtype: float64

**df.kurtosis()**

Points -0.450432

Score 0.416595

Weigh 0.864931

dtype: float64

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



**ANS.** In the above histogram frequency is on the y-axis and chick weight is on x-axis. Here we can see clearly that most of the chick weights belongs to 50 – 100. Most of the data is right skewed which means it is positively skewed.Here mean is greater than the median. Between 300 – 400 we can observe outliers whose values are approximately between 0 – 20 frequency.



**ANS.** The above box plot tells us that the data is positively skewed.Q3-Q2 is greater than Q2-Q1. As we can see the variation is very low here. Even the data is not uniformaly distributed. lower 25% data is bunched tightly in between Q1-Q2 which means most of the data is present here. That’s why the bar in histogram from 50-100 is taller than others. There are many outliers on the upper side of the plot.

The higher numbers are at the lower side which indicated the data is positively 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?

**ANS**

Import pandas as pd

Import numpy as np

From scipy import stats

#### average weight of adults in mexico with 94% C.I

Stats.t.ppf(0.94,1999)

=1.555

Standard error=30/√2000

=0.67

Now, put all the values in the formula :-

X̅ ± t1-α \* s/√n

200-1.555\*0.67

200+1.555\*0.67

Hence the values we get are:-

[198.738, 201.261]

###### average weight of adults in mexico with 96% C.I

Stats.t.ppf(0.96,1999)

=1.751

Standard error=30/√2000

=0.67

Now, put all the values in the formula :-

X̅ ± t1-α \* s/√n

200-1.751\*0.67

200+1.751\*0.67

Hence the values we get are:-

[198.82, 201.173]

###### average weight of adults in mexico with 98% C.I

Stats.t.ppf(0.98,1999)

=2.055

Now, put all the values in the formula :-

X̅ ± t1-α \* s/√n

200-2.055\*0.67

200+2.055\*0.67

Hence the values we get are:-

[198.62, 201.37]

**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?

**ANS.**

**1.Mean-** Sum of Observation / Total numbers of Observations

Import numpy as np

np.mean([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

Our mean = 41.0

**2.Median-** numberof observations n = 18(even)

40+41/2 = 81/2

= 40.5

np.median([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

**3.** **Variance**-

np.var([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

= 24.1111111111

**4.std dev-**

np.std([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

= 4.910306620885412

**2.**

After seeing the student marks we can observe that most of the students secured 41 marks and only 1 student got maximum marks in class which is 56.

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

ANS. When mean = median we can say that there is no skewness in our dataset. So, it’s a symmetrical distribution.

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

ANS. When mean > median it means the tail is towards the right and it is positively skewed.

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

ANS. When median > mean it means the tail is towards our left and the data is negatively skewed.

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

ANS.when Kurtosis is > 0 it’s a positive Kurtosis. Here most of the points are located in tail as compared to near mean.Here the chances of having Outliers is far more than outher kurtosis.

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

ANS. When kutrtosis < 0. Here tail is near the mean which indicates more data points are near mean.

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



**i).**What can we say about the distribution of the data?

**ANS**. As we can see clearly that our median is shifted towards the shorter whisker. Which means the data is skewed towards the left. Right side whisker is very short as compared to the left sided whisker. So this distribution of data is negatively skewed.

Here, mean<median<mode

We observed more data in between Q1 and Q2 as compared with Q2 and Q3

Q2-Q1 > Q3-Q2

**ii).**What is nature of skewness of the data?

**ANS**. Here nature of skewness is negative which is towards left. As the data is towards the left so the mean is toward the left which is smaller than the median.

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

**ANS.** IQR=Q3-Q1

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.

**ANS**. Here we can clearly see that the 1st boxplot is very much tightly clustered around the median as compared with the 2nd plot as the 2nd represents very much dispersed values because it is really spread out.

But if we talk about the median then it’s clear that both the plots have same median values approx 260. As the median is in the centre of both. Boxplot 2 is uniformally distributed because the amount of the data in every quadrant is same .

In boxplot 1 tons of the data is present in small space between Q1 and Q3.

**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)

**ANS.** car=pd.read\_csv('C:/Users/RIG1/Desktop/DS ASSIGNMENTS/cars.csv')

df=car

df.MPG.mean() ---------------- 34.422075728024666

df.MPG.std() ----------------------- 9.131444731795982

from scipy import stats

**a).**# P(MPG>38)

a=stats.norm.cdf(38,34.42,9.13)

1-a

0.34748702501304063

**b).**# P(MPG<40)

stats.norm.cdf(40,34.42,9.13)

0.7294571279557076

**c).**# P(20<MPG<50)

a1=stats.norm.cdf(20,34.42,9.13)

b1=stats.norm.cdf(50,34.42,9.13)

c1=b1-a1

c1

0.8989177824549222

**Q 21)** Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANS.** import matplotlib.pyplot as plt

plt.hist(car['MPG'])

Here we can clearly see that the MPG of cars do not follow 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

**ANS.**

df1=pd.read\_csv('C:/Users/RIG1/Desktop/DS ASSIGNMENTS/wc-at.csv')

plt.hist(df1['AT'])

plt.hist(df1['Waist'])

As we have plotted a histogram for these columns and we have found that both the columns do not follows Normal Distribution.

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

**ANS.**

Z score of 90% confidence interval - 1.281

Z score of 94% confidence interval - 1.554

Z score of 60% confidence interval – 0.253

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

**ANS** Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

T score for 95% C.I. = 2.064

T score for 99% C.I. = 2.492

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

**ANS.**

X=260

s= 90

n=18

u(mew)=270

260-270/90/sqrt 18

= -10/90/sqrt 18

= -0.423