|  |  |
| --- | --- |
| 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 | Continuous |
| Number of times married | Discrete |
| Gender (Male or Female) | Continuous |

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

**Ans**=3/8=0.375

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

1. Equal to 1

**Ans=**0

1. Less than or equal to 4

**Ans=** The probability of getting<=4 is 1/6 When we tow dies are thrown ,the probability less than or equal to 4 is (1,1),(1,2),(1,3),(2,1),(3,1) ,Therfore

* Number of favourable outcomes=6
* Total number of possibilities=36

Probability= Number of favourable outcomes / Total number of possibilities

                    =6/36

                    =1/6

1. Sum is divisible by 2 and 3

**Ans**=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?

**Ans=**

Total number of balls =7 balls

N (Event (2 balls are drawn randomly from bag) = 7! / 2! \* 5!

= (7\*6\*5\*4\*3\*2\*1) /

(2\*1) \* (5\*4\*3\*2\*1)

N (Event (2 balls are drawn randomly from bag) = (7\*6)/ (2\*1) = 21

If none of them drawn 2 balls are blue = 7 – 2 = 5

N (Event (None of the balls drawn is blue) = 5! / 2! \* 3! = (5\*4) / (2\*1)

= 10

P (None of the balls drawn is blue) = N (Event (None of the balls drawn is blue) /

N (Event (2 balls are drawn randomly from

bag)

= 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

**Ans**=0.015+0.8+1.95+0.025+0.06+0.24=3.09

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weight>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **POINT** | **SCORE** | **WEIGHT** |
| **Mean** | 3.60 | 3.22 | 17.85 |
| **Median** | 3.70 | 3.33 | 17.71 |
| **Mode** | 3.914 | 3.49 | 17.65 |
| **Variance** | 0.29 | 0.96 | 3.91 |
| **Std Derivation** | 0.53 | 0.98 | 1.79 |
| **Range** | 2.76,4.93 | 1.513,5.424 | 14.5,22.9 |

**Ans= Mean=3.596563**

**Median=3.695**

**Mode=”Numeric”**

**Varience=0.2858814**

**Standard Daviation=0.5346787**

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 of weight for a random selected patient =145.33

**Explanation:-**

Expected value = ∑(probability \*value)

Probability of selecting each patient = 1/9

Expected value

=(1/9)108+(1/9)110+(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

**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=**Skewness and kutosis of car speed and distance is as follow :

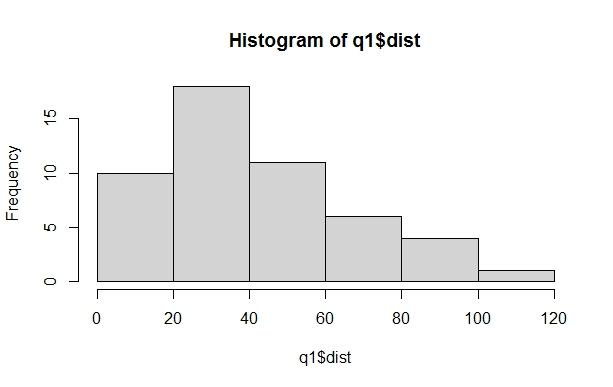
Skewness=-0.111(car speed ) ,0.759(distance)

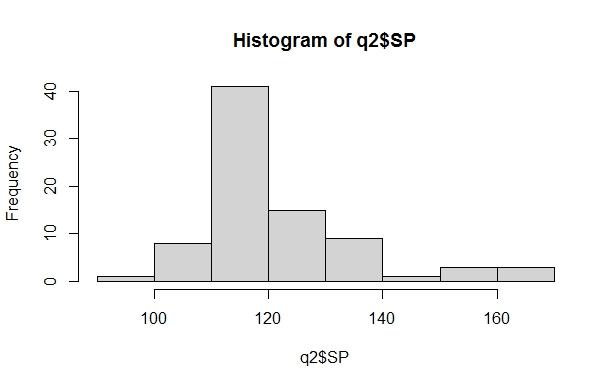
Kutosis=2.42(car speed),3.24(distance)

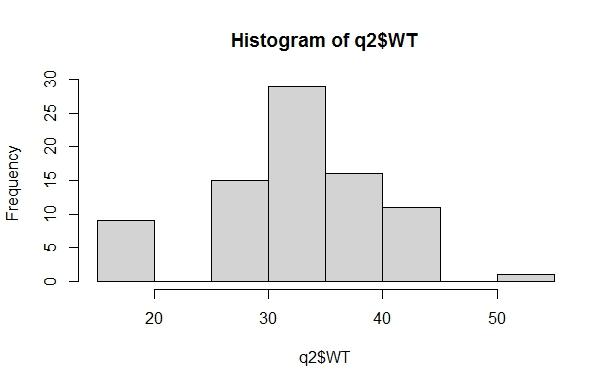
Skewness and kutosis of SP and weight (WT) data are as follow:

    Skewness=1.55(SP), -0.59(weight)

    Kutosis=5.72(SP),3.87(weight)

****

****

****

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

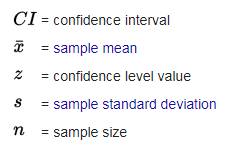
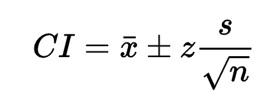


**Ans**=The histogram and boxplot in Fig is positively skewed on right side.

i.e mean and median of the data is greater than mode.

**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**=



Given : x=200, s=30, n=2000

1. The 94% confidence interval is (198.739,201.62) (z=1.8808)
2. The 96% confidence interval is (198.622,201.378) (z=2.0537)
3. The 98% confidence interval is (198.439,201.561) (z=2.3263)

**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?
3. **Ans**=Mean=41 ,median=40.5

Variance=25.529

Standard deviation=5.05

1. Repeately obtained are 36,38,40,41 and 42

Skewness =1.42

i.e.positive.

We don’t have the outliers and the data is slightly skewed towards right because mean >median

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

**Ans**=No skewness is present because of that we have a perfect symmetrical distribution. If the distribution is symmetric, then the mean =median, and the distribution has Zero Skewness

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

**Ans**=Skewness and tail is present at towords right because of that extreme values affect the mean more than the median

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

**Ans**=Skewness and tail is towords left

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

**Ans**=The positive kurtosis indicates a distribution where more number are located in the tails of the distribution instead of around the mean. It means that the curve is more peaked and it is Leptokurtic.

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

**Ans**=A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution.Negative kurtosis means the curve will be flatter and broader.

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



What can we say about the distribution of the data?

**Ans**=The date is not actually equally distributed across the plane. Their might be outliers influencing the data. Median of the data is approximately 14.7. Approximate 25% data is on the left sight in between 0 to 10. Approximate 50% data is in between 10 to 18. Approximate 25% data is after 18

What is nature of skewness of the data?

**Ans**=The data will be left skewed since whisker length on the upper extreme is higher than the data on the lower extreme.

Median>Mean

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

**Ans**=Q1=10

Q2=14.7

Q3=18

IQR=Q3-Q1=18-10=8(Appro.)

Q19) Comment on the below Boxplot visualizations?



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

**Ans=**

In this Fig of boxplot. Boxplot 1 is  positively skewed that’s means mean and median is greater than mode . And in boxplot 2 Fig normal distribution that means the skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero.

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)

**Ans**: Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38,

loc= q20.MPG.mean(), scale= q20.MPG.std()),3)

print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

* 1. P(MPG<40)

**Ans**= prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc =

q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

* 1. P (20<MPG<50)

**Ans**: prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20))

p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<50)=',(prob\_MPG\_less\_than\_50))

P(MPG<50)= 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20)

print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))

P(20<MPG<50)= 0.013000000000000012

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans**= MPG of Cars follows 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**=Adipose Tissue(AT) and Waist does not follow Normal Distribution.





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

**Ans**=1)Z value for 90% confidence interval

Z score of 60% CI = 0.84

         Z score of 90% CI = 1.645

         Z score of 94% CI = 1.881

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

**Ans=**

# t score for 95% confidence interval

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))

T score for 95% Confidence Interval = -2.0639

# t value for 94% confidence interval

print('T score for 94% Confidence Inteval =',np.round(stats.t.ppf(0.03,df=24),4))

T score for 94% Confidence Inteval = -1.974

# t value for 99% Confidence Interval

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))

T score for 95% Confidence Interval = -2.7969

|  |  |
| --- | --- |
| **Confidence Interval** | **T Score** |
| 95% | -2.06 |
| 96% | -2.17 |
| 99% | -2.79 |

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=**

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.

: import numpy as np

Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

(260-270)/90/np.sqrt(18))

t\_score = -0.471

stats.t.cdf(t\_score, df = 17)

0.32 = 32%