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

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

P(2 heads and 1 tail)=3/8

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: 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: 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:

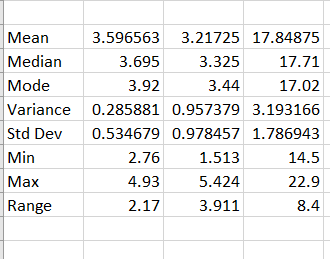
(1\*0.015)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.12)=3.09

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**



The observations that can be made from the dataset are

1. The mean, median and mode of Weight is more different from the other 2 which indicates that there can possibly be outliers in the column Weigh
2. Also relatively High variance of weight indicates that the data is spread over a large range
3. Also the standard deviation for points and score are quite low compared to that of weight which indicates that the datapoints in points and score are more close and are tightly clustered whereas the datapoints of weight are spread

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?

Solution: Expected value is the mean

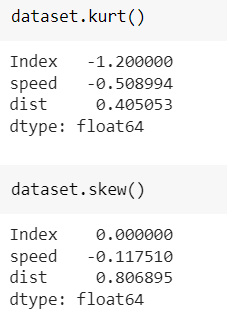
(108+110+123+134+135+145+167+187+199)/9

145.33333333

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

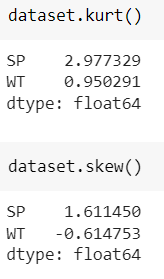
**Cars speed and distance**

**Use Q9\_a.csv**



**SP and Weight(WT)**

**Use Q9\_b.csv**



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



From the Histogram we can observe that the frequency if Chickweight is highest in the range 50-100 and lowest between 350-400. It also forms a graph that is more skewed towards the right because there are many datapoints with less frequency and high weights which can also be outliers in this case. Therefore the Mean>Median>mode because the distribution is right skewed. We can also get info from the plot that maximum number of datapoints lie within the 50-200 range of weight. Then as the weight increases the frequency is less.

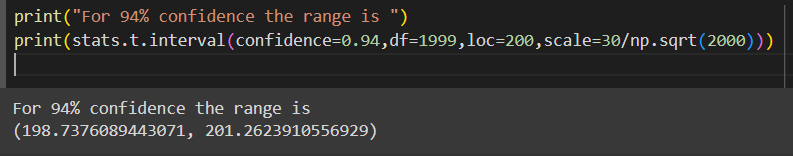


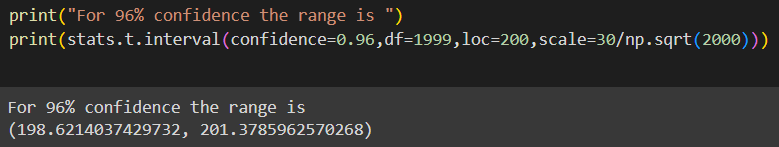
The information obtained from the box plot is that it has many outliers to the maximum side. It also means that the data is right skewed and also the median lies close to the Lower Quartile. The outliers are abovev the Upper Extereme.

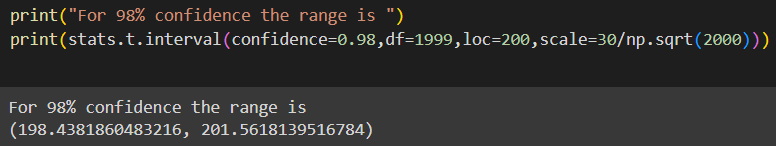
**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

Since the population standard deviation is not given we have to use t-distribution.

From python we can find it using the following code







**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=41.0

Median=40.5

Mode=41

Variance=25.529411764705884

Standard deviation=5.05266382858645

1. What can we say about the student marks?

There are no outliers In the dataset. We can say that it is slightly right skewed because mean>median

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

No skewness

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 ?

It has a narrow peak and is heavy tailed meaning that there are more outliers in the data. It is called Leptokurtic

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

Negative value of kurtosis indicates that the distribution is broader with thin tails and the number of outliers are less as most of the datapoints are within the 1 standard deviation. It is called Platykurtic

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



What can we say about the distribution of the data?

There are no outliers in the data. Most of the datapoints are within the range of 10 to 18 with the median(50 percentile) towards the upper quartile.

What is nature of skewness of the data?

The data seems to be left skewed as there are more number of data points between the 25 percentile and the 50 percentile when compared to 50 percentile and 75 percentile.

What will be the IQR of the data (approximately)?   
The Inter Quartile Range is upper quartile-lower quartile

Here IQR is 8

Q19) Comment on the below Boxplot visualizations?



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

Some of the major inferences that can be drawn from the visuals are

1. There are no outliers
2. The median for both the plots is approximately the same
3. There is no skewness in the data that is they are normally distributed.
4. There are more number of datapoints in the 2nd compared visual compared to that of the first

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)

0.34759392515827137

* 1. P(MPG<40)

0.7293498762151609

c. P (20<MPG<50)

0.8988689169682047

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

It does form a normal distribution curve which is slightly left skewed and is Platykurtic

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

Dataset: wc-at.csv

They do form normal distribution curves which is slightly right skewed and is Platykurtic

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

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

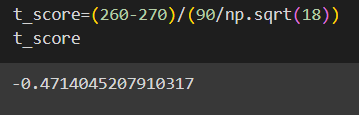
Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

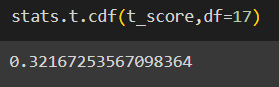
rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

Wkt t\_score=(x-mean)/(sample stddev/square root of sample size)



To calculate the pvalue:



32% probability that 18 randomly selected bulvs will have an avg lige of no more than 260