Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete numerical variable |
| Results of rolling a dice | Discrete numerical variable |
| Weight of a person | Continuous numerical variable |
| Weight of Gold | Continuous numerical variable |
| Distance between two places | Continuous numerical variable |
| Length of a leaf | Continuous numerical variable |
| Dog's weight | Continuous numerical variable |
| Blue Color | Categorical variable |
| Number of kids | Count variable |
| Number of tickets in Indian railways | Count variable |
| Number of times married | Count variable |
| Gender (Male or Female) | Categorical variable |

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

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

ANS:

P(2heads and 1 trail)={HHH,HHT,HTH,THH,HTT,TTH,THT,TTT}

Number of coins tossed=3

Number of heads=2

Number of trials=1

{HHT,HTH,THH}

P=(3/2)\*(1/2)\*\*2\*(1-1/2)=3\*1/4\*1/2

Probability =3/8

probability that two heads and one tail when tossing three coins is 3/8

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

1. Equal to 1

Ans: Probability of getting a sum of 1 is 0

1. Less than or equal to 4

Ans : Sum is less than or equal to 4 are:

1.(1,1)-sum=2

2.(1,2)-sum=3

3.(2,1)-sum=3

4.(1,3)-sum=4

5. (2,2)-sum=4

6.(3,1)-sum=4

Number of favorable out comes 6 and Total number of outcomes is 36 by rolling 2 Dice. Therefor probalititly is:

Number of favorable outcomes = 6/36 = 1/6

Total number of outcomes

1. Sum is divisible by 2 and 3

Ans : Need to find the common multiplies of 2 and 3, which is 6:

1.(1,5)-sum=6

2.(2,4)-sum=6

3.(3,3)-sum=6

4.(4,2)-sum=6

5. (5,1)-sum=6

There are 5 favorable outcomes out of total of (6\*6=36 times) possible outcomes when rolling two dice. Therefor probalititly is:

Number of favorable outcomes = 5/36

Total number of outcomes

therefore, the probalititly that the sum of the numbers rolled on two dice is divisible by both 2 and 3 is 5/36

5) 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=2red+3green+2blues=7 balls

S= two balls are drawn at random

N(s)= C (7,2) = 7\*6/2\*1=42/2=21

N(s)=21

A= none of the ball drawn is blue

N(a)=C (5,2) =5\*2/2\*1 = 20/2 =10

So, probability is that none of the ball is drawn Is blue is:

P =Number of ways to draw two balls without blue = 10/21

Total number of ways to draw two balls

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 : To calculate the expected number of candies for a randomly selected child, we multiply each candy count by its corresponding probalititly and then sum up the results.

Expected number of candies((E)):

E= (1\*0.015) +(4\*0.20) +(3\*0.65) +(5\*0.005) +(6\*0.01) +(2\*0.120)

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

E= 3.135

So, the excepted number of candies for a randomly selected child is 3.135.

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 : 1.Mean:

points=3.59656

Score=3.21725

Weight=17.8487

**2.** Median:

points=3.6950

Score=3.325

Weight=17.71

3.Mode:

points=the mode is 3.07 and 3.92 appears thrice in the data.

Score=3.44

Weight=17.02 and 18.90 appears thrice in the data.

4. Variance:

points=3.59656

Score=3.21725

Weight=17.8487

4.standard deviation:

points=0.5346

Score=0.9784

Weight=1.7869

4.range: max-min

points=2.17

Score=3.911

Weight=8.399

**Comments and inferences:**

1. **For** points **,** thevalues are quite close to each other with a variance and standard deviation, indicating consistency in the data.
2. **Score** has a slightly higher variance compared to points, suggesting some variability in the scores.
3. The weight variable has the highest variance and standard deviation among the three, indicating the most significant variability in the data.
4. Mode is only applicable to points as it has repeated values.
5. The range indicates the spread of the data, with weight having the widest range.
6. Overall, points show the least variability among the three variables, while weight show the highest variability.

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 : Given Weights(x) of patients: x= {108, 110, 123, 134, 135, 145, 167, 187, 199}

Number of patients (n) =9

Now, to find the expected value (E), you sum up all the weights and divide by the number of patients:

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

E=1308/9=145.33 pounds

The expected value of the weight of a randomly chosen patient is approximately

145.33 pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

Ans :

**To find skewness, kurtosis ,upload the csv file in jupyter notebook**

**df.skew() is used to find Skewness in jupyter**

**Skewness: Index=0.0000**

**cars speed= -0.117510,**

**distance=0.806895**

**dtype: float64**

**df.kurt() is used to find Kurtosis in jupyter**

**Kurtosis= Index=-1.200000**

**cars speed= -0.508994**

**distance= 0.405053**

**dtype: float64**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**df.skew() is used to find Skewness in jupyter**

**Skewness:**

**sp = 1.611450**

**wt = -0.614753**

**dtype: float64**

In [26]:

**df.kurt() is used to find Kurtosis in jupyter**

**Kurtosis=**

**sp = 2.977329**

**wt = 0.950291**

**dtype: float64**

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



The histogram appears to have for bins, each with a width of 100 units.chickweight$weight on the X-axis and frequency in Y- axis of data. There is very less data on the range 300 to 400 . the nature of skewness in given histrogram is positive skewness so the large number of the data is present on the right side of the median.



Ans:

In given boxplot mean is greater than median so the nature of skewness is positive skewness. As there positive skewness the large number of data is present on the right side of the boxplot. There are outliers on the upper extreme. As mean is greater than median the skewness is unsymmetric.

**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 : x\_bar(sample weight)=200, n=2000,, standard deviation=30

1.critical z value(94%)

stats.norm.ppf(1-0.06/2) =1.89

2. critical z value(98%)

stats.norm.ppf(1-0.02/2) =2.32

3. critical z value(96%)

stats.norm.ppf(1-0.04/2)=2.05

import math

1. pop mean lower limit value with CI 94%

(200-(1.89\*(30/math.sqrt(2000)))) =198.73

2.pop mean upper limit value with CI 94%

(200+(1.880\*(30/math.sqrt(2000)))) =201.26

3.pop mean lower limit value with CI 98%

(200-(2.32\*(30/math.sqrt(2000))))=198.44

4. pop mean upper limit value with CI 98%

(200+(2.32\*(30/math.sqrt(2000))))= 201.556

5.pop mean lower limit value with CI 96%

(200-(2.05\*(30/math.sqrt(2000)))) = 198.62

6. pop mean upper limit value with CI 96%

(200+(2.05\*(30/math.sqrt(2000))))=201.375

**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.

Ans:

**Mean**=**34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56**

**Mean= sum of all scores**

Number of scores

Mean **= 738/18= 41 Mean= 41**

**Median= the data is already sorted, the median is the middle value. For an odd number of values, it’s the middle one, and for an even number, it’s the average of the two middle values**

**Median = ( Number of data+1/5)**

**= (18+1)/2 =9.5**

**= (40+41)/2=40.5 Median=40.5**

**Variance = 24.1**

**standard deviation = square of variance = 4.91**

1. What can we say about the student marks?

Ans :Mean : the average score of the student is 41.

Median: 50% of the student ‘s scores are below 40.5

Variance: the scores are some what spread out from the mean, with a variance of 24.1

Standard deviation: the average deviation from the mean score is approximately 4.91.

Based on these statistics, we can say that the student’s marks are relatively centered around the mean score of 41, with some variation and few higher scores.

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

Ans :When the mean and median of a dataset are equal, the skewness of the data is typically close to zero. This suggests that the distribution is approximately symmetric. However, it’s possible for the distribution to still exhibit some degree of skewness, albeit minor , depending on the exact shape of the data distribution.

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

When mean is greater than the median, it indicates that the distribution is positively skewed. This means that there is a longer tail on the right side of the distribution, pulling the mean towards higher values compared to the median.

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

Ans : When the median is greater than the mean, it indicated that the distribution is negatively skewed. This means that there is a longer tail on the left side of the distribution, pulling the median towards lower values compared to the mean.

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

Ans :A positive kurtosis value indicates that the distribution of the data has heavier tails and a sharper peak than a normal distribution. This is often referred to as leptokurtic. It means that there are more extreme values (outliers) in the data compared to a normal distribution.

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

Ans : A negative kurtosis value indicates that the distribution of the data has lighter tails and a flatter peak than a normal distribution. This is often referred to as platykurtic. it means that data has fewer extreme values (outliers) compared to a normal distribution.

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



What can we say about the distribution of the data?

Ans:

The distribution of the data in the above Boxplot is skewed mostly to the Right side and data distributed less to the left side. The medium between the number line is 10 to 18.

What is nature of skewness of the data?

Ans:

The nature of the skewness of the boxplot is negatively skewed.it is a type of distribution where the data is mostly distributed to the right side.

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

The Inter quantile range (IQR) for the above boxplot is :

Q1=10

Q2=15

Q3=18

Formula : IQR= Q3-Q1

=18-10

IQR= 8  
  
Q19) Comment on the below Boxplot visualizations?



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

Ans:

Boxplot 1:

The lower quartile for boxplot1 is (Q1):250

The upper quartile for the boxplot1 is (Q3):275

The inter –quartile range is (IQR)=Q3-Q1

= 275-250

= 25

Lower extreme=Q1-1.5\*IQR

=250-1.5\*25

=212.5

Upper extreme= Q3+1.5\*IQR

=275+1.5\*25

=312.5

Boxplot 2:

The lower quartile for boxplot1 is (Q1):225

The upper quartile for the boxplot1 is (Q3):300

The inter –quartile range is (IQR)=Q3-Q1

= 300-225

= 75

Lower extreme=Q1-1.5\*IQR

=225-1.5\*75

=112.5

Upper extreme= Q3+1.5\*IQR

=300+1.5\*75

=412.5

Inference:

From the above 2 boxplots the median for boxplot1 is 25 and median for boxplot2 is 75. The outliers for boxplot 1 lies below 212.5 lower limit and above 237.5 upper limit. The outliers for boxplots 2 is below 112.5 lower limit and above 412.5 upper limit. In the boxplot 1 right skewed and in boxplot 2 is symmetric.

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

ANS: import pandas as pd

import numpy as np

import scipy.stats as stats

* 1. P(MPG>38)

1-stats.norm.cdf(38,df.MPG.mean(),df.MPG.std())

=0.3475939

* 1. P(MPG<40)

stats.norm.cdf(40,df.MPG.mean(),df.MPG.std())

=0.729349

* 1. P (20<MPG<50)

p1=stats.norm.cdf(0.20,df.MPG.mean(),df.MPG.std())

p2=stats.norm.cdf(0.50,df.MPG.mean(),df.MPG.std())

P2-p1

=1.243096

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans :import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df1=pd.read\_csv('Cars.csv')

df1

df1['MPG'].mean().round(3)

=34.442

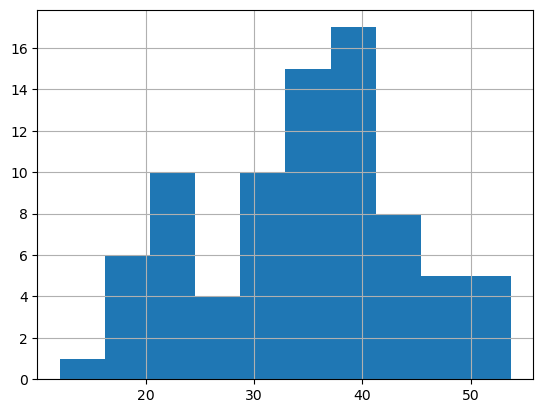
df1['MPG'].median()

=35.152

df1['MPG'].mode()

=0 29.629936

df1['MPG'].hist()



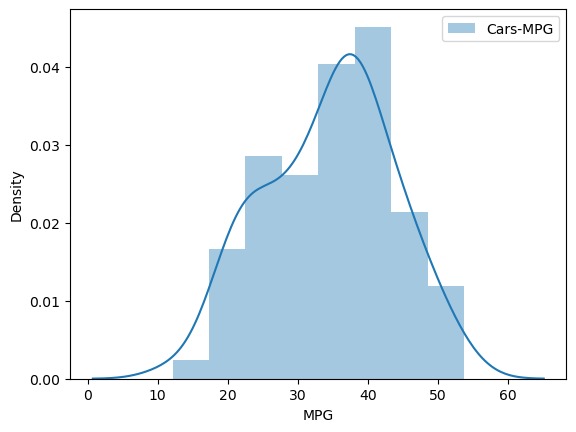
Sns.displot(df1[‘MPG’],label=’Cars-MPG’)

Plt.xlabel(‘MPG’)

Plt.ylabel(‘Density’)

Plt.grid(True)

Plt.plot()



df1['MPG'].skew()

=0.1779467

df1['MPG']. kurt ()

=0.611678

Inference:

From the data histogram and plots we can say that data in MPG of cars is fairly symmetrical,I.e it is normaly distributed.

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 pandas as pd

df=pd.read\_csv('WC-at.csv')

df

df. Mean()

**Waist 91.901835**

**AT 101.894037**

**dtype: float64**

**df**.median()

**Waist 90.80**

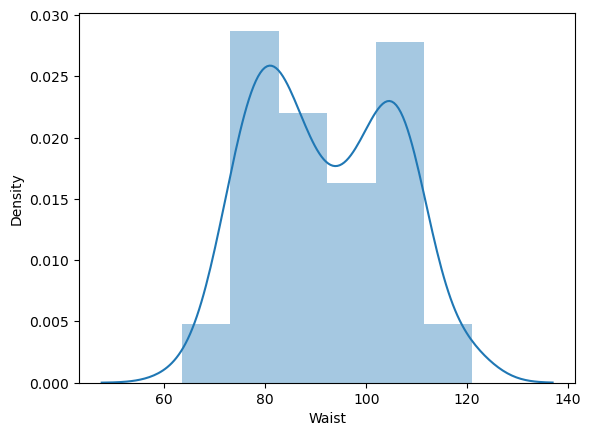
**AT 96.54**

**dtype: float64**

**df.mode()**

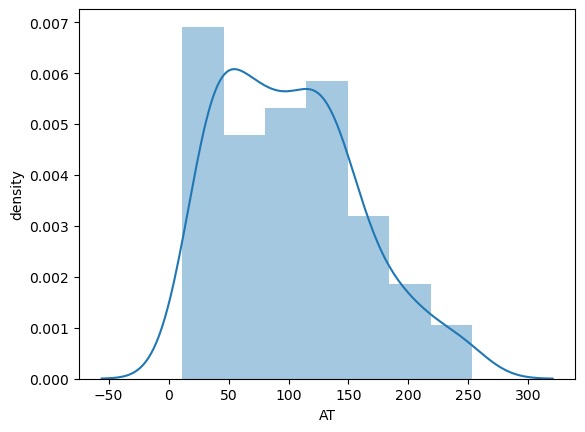
|  |  |  |
| --- | --- | --- |
|  | WAIST | AT |
| **0** | **94.5** | **121.0** |
| **1** | **106** | **123.0** |
| **2** | **108.5** | **NAN** |

sns.distplot(df['Waist'])

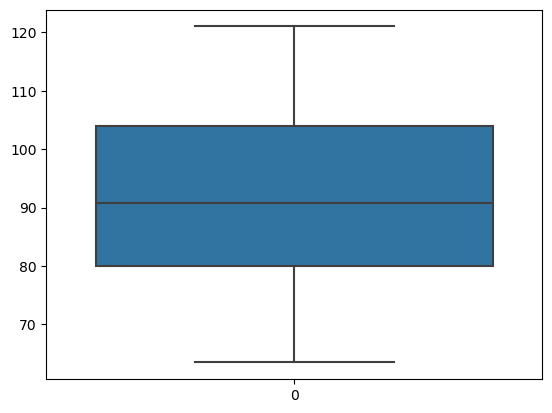


sns.displot(df['AT'])

plt.show()



sns.boxplot(df['Waist'])



Inference:

From the ploting the data is negatively kurtosis, positively skewed.

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

Here are the approximate Z-score for the given confidence intervals:

Stats.norm.ppf (1- alpha /2) alpha=0.9

1. For a 90% confidence interval:

90%= (1-0.9) =0.1

Stats.norm.ppf (1-0.1/2) =1.645.

The z-score corresponding to a 90% confidence interval is approximately 1.645.

1. For a 94% confidence interval:

94%= (1-0.94) =0.06 alpha =0.94

Stats.norm.ppf (1-0.06/2) =1.88

The z-score corresponding to a 94% confidence interval is approximately 1.88.

1. For a 60% confidence interval:

60%= (1-0.6) =0.4 alpha =0.6

Stats.norm.ppf (1-0.4/2) =0.84.

The z-score corresponding to a 60% confidence interval is approximately 0.84.

These values are based on the standard normal distribution.

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

Here are the approximate T-score for the given confidence intervals:

With sample size of 25, degree of freedom (df) is 24. n=25

Df=n-1,25-1=24

Df=24

Stats.norm.ppf (1- alpha /2, df) alpha =0.05

1.For a 95% confidence interval:

95%= (1-0.95) =0.05

Stats.norm.ppf (1-0.05/2,24) =2.064

The z-score corresponding to a 95% confidence interval is approximately 2.064.

2.For a 96% confidence interval:

96%= (1-0.96) =0.04 alpha =0.04

Stats.norm.ppf (1-0.04/2,24) =2.178

The z-score corresponding to a 96% confidence interval is approximately 2.178.

3. For a 99% confidence interval: alpha =0.01

99%= (1-0.99) =0.01

Stats.norm.ppf (1-0.01/2,24) =2.797

The z-score corresponding to a 99% confidence interval is approximately 2.797.

These vales represent the number of standard errors away from the sample mean the confidence interval extends.

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

x\_bar=260, n=18, pop\_mean=270, s=90

df=(n-1)

df=17

formula:

tscore=(x\_bar-pop\_mean)/(s/math.sqrt(n))

tscore=-0.4714

now, we use the R code pt (tscore, df) to find the probability:

R code

Tscore<- -0.4714

Df<- 17

P\_value<- pt (tscore, df)

P\_value

Using R code, we find that the probability (p-value) that 18 randomly selected bulbs would have an average life of no more than 260 days, if the Ceo ‘s claim were true, is approximately 0.3217