Q1) Identify the Data type for the Following:

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

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

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

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

Ans. When 3 coins are tossed, the possible outcomes are { HHH, HHT, HTT, HTH, TTT, TTH, THH, THT }

Probability that two heads and one tail are obtained - 3/8.

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

Ans. Two Dice are rolled then the number of sample space are = 36

{ (1,1), (1,2), (1,3), (1,4), (1,5), (1,6),

(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),

(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),

(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),

(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),

(6,1), (6,2), (6,3), (6,4), (6,5), (6,6) }

1. Equal to 1 :- 0/36 = 0
2. Less than or equal to 4 :- The possibility of getting less than or equal to 4 is (1,1), (2,1), (1,2), (1, 3), (2, 2), and (3,1) :- 6/36 = 1/6
3. Sum is divisible by 2 and 3 :- The possible outcomes are = (1,5), (3,3) , (4,2), (5,1), (6,6) :- 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.

Let the sample be S

Total numbers of ball is (2+3+2 = 7)

Given two balls are drawn out of 7

So, n(S) = 7C2 = (7\*6) / (2\*1) = 21

Let E be the event of drawing 2 balls, none of which is blue.

So, n(E) = Number of ways of drawing 2 balls out of (2 + 3) balls.

So, n(S) = 5C2 = (5\*4) / (2\*1) = 10

P(E) = n(E) / n(S) = 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.

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count X | Probability P(X) | X\*P(X) |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.20 | 0.80 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.120 | 0.24 |
|  |  |  | 3.09 |

So, from table expected number of candies for a randomly selected child is 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.

Ans.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.59656 | 3.21725 | 17.8487 |
| Median | 3.6950 | 3.3250 | 17.7100 |
| Mode | 3.9200 | 3.4400 | 17.0200 |
| Variance | 0.2858 | 0.9573 | 3.1931 |
| Std Deviation | 0.5346 | 0.9784 | 1.7869 |
| Range | 2.76 - 4.93 | 1.513 - 5.424 | 14.5 - 22.9 |

Thus as we can see from above dataset mean, median and mode are not same so data is positively skewed.

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.

There are 9 patients

Probability of selecting each patient = 1/9

|  |  |  |
| --- | --- | --- |
| Weights of Patients X | Probability P(X) | X\*P(X) |
| 108 | 1/9 | 12 |
| 110 | 1/9 | 12.22 |
| 123 | 1/9 | 13.66 |
| 134 | 1/9 | 14.88 |
| 135 | 1/9 | 15 |
| 145 | 1/9 | 16.11 |
| 167 | 1/9 | 18.55 |
| 187 | 1/9 | 20.77 |
| 199 | 1/9 | 22.11 |
|  |  | 145.3 |

So, Expected Value of the Weight of that patient - 145.3

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

Cars speed and distance

Ans.

data = pd.read\_csv('Q9\_a.csv')

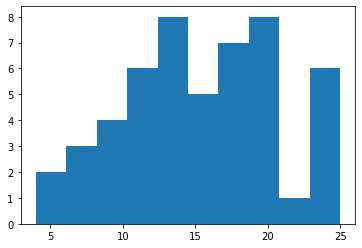
data.skew()

speed : -0.117510 dist : 0.806895

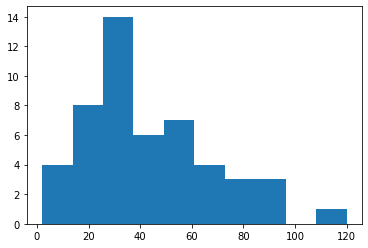
data.kurtosis()

speed : -0.508994 dist : 0.405053

plt.hist(data['speed']) #Negatively skewed



plt.hist(data['dist']) #Positively skewed

****

SP and Weight(WT)

data = pd.read\_csv('Q9\_b.csv')

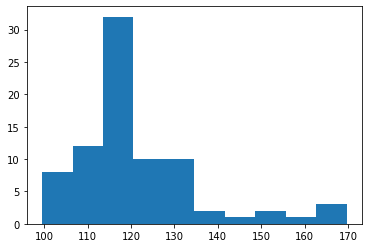
data.skew()

SP: 1.611450 WT: -0.614753

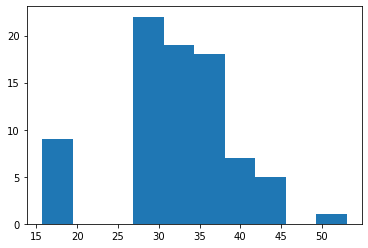
data.kurtosis()

SP: 2.977329 WT:0.950291

plt.hist(data['SP']) #Positively skewed



plt.hist(data['WT']) #Negatively skewed



Q10) Draw inferences about the following boxplot & histogram



Ans. Above histogram is positively skewed. And outliers lies on the right of histogram.



Ans. In above boxplot outliers lies above upper extreme and also we can see that there is huge difference between max value and median value.

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. python code -

print("94% confidence interval:-",stats.norm.interval(0.94 ,loc = 200,scale = 30))

94% confidence interval:- (143.57619175546247, 256.42380824453755)

print("98% confidence interval:-",stats.norm.interval(0.98 ,loc = 200,scale = 30))

98% confidence interval:- (130.2095637787748, 269.7904362212252)

print("96% confidence interval:-",stats.norm.interval(0.96 ,loc = 200,scale = 30))

96% confidence interval:- (138.38753268104531, 261.61246731895466)

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. Python code :-

import numpy as np

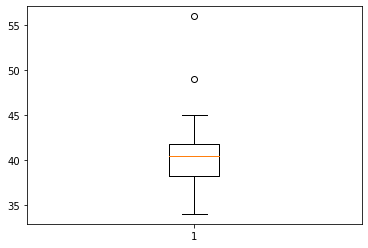
import pandas as pd

import matplotlib.pyplot as plt

df = pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

df.mean() :- 41.0, df.median() :- 40.5, df.var() :- 25.52, df.std() :- 5.05

plt.boxplot(df)



So, from above plot and information we can say that mean of marks of student is 41 which is slightly greater than median 40.5.

There are two outlier ~ 49,56.

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

Ans. Symmetric in nature.

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

Ans. Positively skewed in nature.

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

Ans. Negatively skewed in nature.

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

Ans. Positive kurtosis value implies that the peak is pointer and tail is thicker.

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

Ans. Negative kurtosis value implies that the peak is wider and tail is thinner.

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



What can we say about the distribution of the data?

Ans. The above box plot show that the median is closer to the upper quartile and the distribution of the data is Skewed.

What is nature of skewness of the data?

Ans. Negatively skewed in nature.

What will be the IQR of the data (approximately)?   
Ans. Q1=10 , Q3=18

IQR of the data = 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. In above boxplot (1) we can see that most of point lies between 258 to 277 and in boxplot (2) we can see that most of point lies between 225 to 310, Also we can see that median of both lies at 260.

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

P(MPG>38)

1 - st.norm.cdf(38,data['MPG'].mean(),data['MPG'].std())

Ans. 0.3475939251582705

P(MPG<40)

st.norm.cdf(40,data['MPG'].mean(),data['MPG'].std())

Ans. 0.7293498762151616

P (20<MPG<50)

st.norm.cdf(50,data['MPG'].mean(),data['MPG'].std()) - st.norm.cdf(20,data['MPG'].mean(),data['MPG'].std())

Ans. 0.8988689169682046

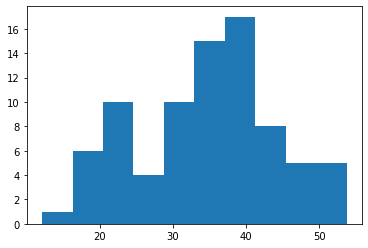
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Ans.

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

plt.hist(data['MPG'])



data['MPG'].mean()

34.422075728024666

data['MPG'].median()

35.15272697

data['MPG'].skew()

-0.17794674747025727

As, we can see the MPG of Cars is negatively skewed, it doesn’t follow Normal

Distribution. (Mean, Median of a Normal Distribution are equal)

b) Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist)

from wc-at data set follows Normal Distribution.

Ans.

data = pd.read\_csv('wc-at.csv')

data.median()

Waist 90.80

AT 96.54

data.mean()

Waist 91.901835

AT 101.894037

As we can see Waist Circumference (Waist) median is not equal to mean.Thus

Waist Circumference (Waist) from wc-at data set doesn’t follows Normal

Distribution. Also we can see that Adipose Tissue (AT) median is not equal to

mean. Thus Adipose Tissue (AT) from wc-at data set doesn’t follows Normal

Distribution.

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

Ans.

from scipy import stats as st

print('z Score of 90%:',st.norm.ppf(0.90))

print('z Score of 94%:',st.norm.ppf(0.94))

print('z Score of 60%:',st.norm.ppf(0.60))

output :-

z Score of 90%: 1.2815515655446004

z Score of 94%: 1.5547735945968535

z Score of 60%: 0.2533471031357997

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

Ans.

from scipy import stats as st

print('t Score of 95%:',st.t.ppf(0.05,25-1))

print('t Score of 96%:',st.t.ppf(0.04,25-1))

print('t Score of 99%:',st.t.ppf(0.01,25-1))

output :-

t Score of 95%: -1.7108820799094282

t Score of 96%: -1.8280511719596344

t Score of 99%: -2.4921594731575762

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

Ans.

Mean of the sample of bulbs =  260

Population mean = 270

Standard deviation of the sample = 90

Number of items in the sample = 18 (square root or 18 = 4.24)

t = (260 – 270) / (90/4.24)

t = - 0.471

df = 18 - 1 = 17

python code :- stats.t.cdf (t,df)

stats.t.cdf(-0.47,17)

0.32216394448907903

So, **0.322** is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days.