|  |  |
| --- | --- |
| 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 | Categorical |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Categorical |

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

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

Ans- There coins are tossed so the possibility of outcomes will be: HHH, HHT, HTH, THH, TTT, TTH, THT and HTT

The probability of getting two Heads and one tail = (interested events / total no of outcomes) = 3/8 = 37 %

**In Python**

def event\_probability(event\_outcomes, sample\_space):

probability = (event\_outcomes / sample\_space) \* 100

return round(probability, 1)

tevents = 8

ievents = 3

HT\_probability = event\_probability(ievents,tevents)

print('Probability of getting 2 heads & 1 tails is:',str(HT\_probability) + '%')

Probability of getting 2 heads & 1 tails is: 37.5%

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

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

Ans –

Two Dice are rolled, so the no of outcomes = 6 \* 6 = 36

1. When we rolled two dice the probability of sum is minimum 2, because it starts from (1,1) so the probability that sum is equal to 1 = ZERO
2. The probability that sum is less than or equal to 4

According to above condition the outcomes are (1,1),(1,2),(1,3),(2,1),(2,2) and (3,1)

So now the probability that sum is less than or equal to 4 = 6/36 =16.66 %

1. The probability that sum is divisible by 2 and 3

For the above condition the outcomes are: (1,5),(2,4),(3,3),(4,2),(5,1),(6,6)

So now the probability that sum is divisible by 2 and 3 = 6/36 = 16.66%

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 = (2 + 3 + 2) = 7  
Let S be the sample space  
Then, n(S) = Number of ways of drawing 2 balls out of 7

n(S)=7C2

n(S)=(7×6) / (2×1)= 21

Let E = Event of 2 balls, none of which is blue  
∴ n(E) = Number of ways of drawing 2 balls out of (2 + 3) balls

n(E)=5C2

n(E)=(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 –

Expected number of candies for a randomly selected child

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

= 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, Weigh>

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

**Use Q7.csv file**

Ans-

import pandas as pd

df = pd.read\_csv('D:\\Study\\Assignments\\Q7.csv')

df

df.describe()

df.var()

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | 3.596563 | 3.217250 | 17.848750 |
| **Median** | 3.695000 | 3.325000 | 17.710000 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Variance(s2)** | 0.285881 | 0.957379 | 3.193166 |
| **Standard Deviation(s)** | 0.534679 | 0.978457 | 1.786943 |
| **Range** | 2.17 | 3.911 | 8.4 |

**Inferences:**

Here in this case of different models of cars data, most type of cars have average points of 3.596563 , score of 3.217150 and weigh of 17.848750. Also here in this scenario the standard deviation is very low in points and score so chances of presence of outliers in both the case is very low and comparing to weigh there is little bit higher standard deviation so may be some outliers are present.

Somehow data points in every case have less spread so most of the data points lie near to the median.

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:**

The weights(X) of patients at a clinic (in pounds) are 108,110, 123, 134, 135, 145, 167, 187 and 199

There are 9 patients and also their weights are different

So the probability of choose each patient = 1/9

Here E(X) - 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 = ∑ (probability x Value)

 ∑ P(x).E(x)

Expected Value = ((1/9) x 108) + ((1/9) x 110) + ((1/9) x 123) + ((1/9) x 134) + ((1/9) x 135) + ((1/9) x 145) + ((1/9) x 167) + ((1/9) x 187) + ((1/9) x 199)

= (1/9) x (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9) x (1308)

= 145.33 ~ 145

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:**

import pandas as pd

Q9\_a = pd.read\_csv("D:\\Study\\Assignments\\Q9\_a.csv")

Q9\_a

Q9\_a.skew(axis = 0, skipna = True)

Q9\_a.kurt(axis = 0, skipna = True)

Skewness of car speed = -0.117510

Skewness of distance = 0.806895

Kurtosis of car speed = -0.508994

Kurtosis of distance = 0.405053

**Inferences:**

* For car speed skewness is negative and also the kurtosis is negative, which suggests that the distribution is more towards left. It means the distribution is left skewed or negative skewed. Here in negative skewed mean is less than median. As taking kurtosis into consideration it shows that the distribution has broad peak and thin tail.
* For the distance travel by the car skewness is positive and also the kurtosis is positive, which suggests that the distribution is more towards right. It means the distribution is right skewed or positive skewed. Here in positive skewed mean is greater than median. As taking kurtosis into consideration it shows that the distribution has pointed peak and wide tail.

**SP and Weight (WT)**

**Use Q9\_b.csv**

**Ans:**

import pandas as pd

Q9\_b = pd.read\_csv("D:\\Study\\Assignments\\Q9\_b.csv")

Q9\_b

Q9\_b.skew(axis = 0, skipna = True)

Q9\_b.kurt(axis = 0, skipna = True)

Skewness of SP = 1.611450

Skewness of WT = -0.614753

Kurtosis of SP = 2.977329

Kurtosis of WT = 0.950291

**Inferences:**

* For SP skewness is positive and also the kurtosis is positive, which suggests that the distribution is more towards right. It means the distribution is right skewed or positive skewed. Here in positive skewed mean is greater than median. As taking kurtosis into consideration it shows that the distribution has pointed peak and wide tail.
* For WT skewness is negative and the kurtosis is positive, which suggests that the distribution is more towards left. It means the distribution is left skewed or negative skewed. Here in negative skewed mean is less than median.

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



**Inference:**

From this above Histogram and Box plot, it shows that the distribution has outliers at the end (means in histogram tail side and in box plot at in upper extreme). The distribution is positive skewed or right 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:**

Here total no of sample men(n) = 2000

The avg weight of person in sample(X̅) = 200

Standard deviation of sample() = 30

Confidence Interval =

For 94% of CI value Z score = 1.89

Confidence interval for 94% = 200 ± (1.89 x (30/))

=198.73 to 201.27

For 98% of CI value Z score = 2.33

Confidence interval for 98% = 200 ± (2.33 x (30/))

=198.43 to 201.56

For 96% of CI value Z score = 2.06

Confidence interval for 96% = 200 ± (2.06 x (30/))

=198.62 to 201.38

**OR in Python**

from scipy import stats

import numpy as np

from math import sqrt

ci\_94 = stats.norm.interval(0.94,200,scale = (30/sqrt(2000)))

print('Weight at 94% confidence interval is:',np.round(ci\_94,4))

Weight at 94% confidence interval is: [198.7383 201.2617]

ci\_98 = stats.norm.interval(0.98,200,scale = (30/sqrt(2000)))

print('Weight at 98% confidence interval is:',np.round(ci\_98,4))

Weight at 98% confidence interval is: [198.4394 201.5606]

ci\_98 = stats.norm.interval(0.98,200,scale = (30/sqrt(2000)))

print('Weight at 98% confidence interval is:',np.round(ci\_98,4))

Weight at 96% confidence interval is: [198.6223 201.3777]

**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:**

import pandas as pd

import statistics as sts

st = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

sts.mean(st)

sts.median(st)

round(sts.variance(st),4)

round(sts.stdev(st),4)

Mean = 41

Median = 40.5

Variance = 25.5294

Standard Deviation = 5.0527

1. What can we say about the student marks?

**Ans:** The student score 41 mark most of the time. He scores average 41 mark.

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

**Ans-** The Nature of skewness is zero.

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

**Ans-** When mean > median, the nature of skewness is Positive. It means right skewed.

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

**Ans-** When the median > mean, the nature of skewness is Negative. It means left skewed.

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

**Ans-** Positive kurtosis indicates that the distribution is peaked and possess thick tails. It means most of the data located on the tail side. And it also indicates that the distribution has heavier tails than the normal distribution.

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

**Ans-** Negative kurtosis value for a data indicates that the distribution has lighter tails than the normal distribution.

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



What can we say about the distribution of the data?

**Ans:** Most of the data distributed between of “10 to 18.3”.

What is nature of skewness of the data?

**Ans:** Nature of skewness of the data is Negative Skewness. It means left skewed.

What will be the IQR of the data (approximately)?   
**Ans:** IQR = 8.2 (approximately). 50 % of data lies in between IQR range.  
  
Q19) Comment on the below Boxplot visualizations?



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

**Ans:** From the both box plot the mean is around 262.5 for both cases. The distribution in both the cases looks like symmetric distribution. We can also say that both the box plot are normally distributed.

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)

**In Python:**

import pandas as pd

cars = pd.read\_csv('D:\\Study\\Assignments\\Cars.csv')

cars.head()

**HP MPG VOL SP WT**

**0 49 53.700681 89 104.185353 28.762059**

**1 55 50.013401 92 105.461264 30.466833**

**2 55 50.013401 92 105.461264 30.193597**

**3 70 45.696322 92 113.461264 30.632114**

**4 53 50.504232 92 104.461264 29.889149**

cars["MPG"].mean()

34.422075728024666

cars["MPG"].std()

9.131444731795982

#a. P(MPG>38)

from scipy import stats

1-stats.norm.cdf(38,34.42,9.13)

0.34748702501304063

#b. P(MPG<40)

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

0.7294571279557076

#c. P(20<MPG<50)

stats.norm.cdf(50,34.42,9.13)-stats.norm.cdf(20,34.42,9.13)

0.8989177824549222

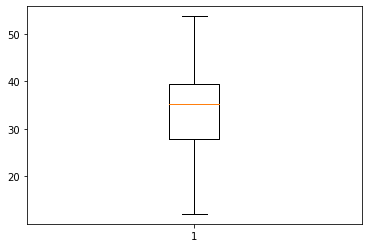
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

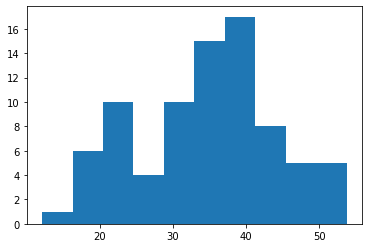
Dataset: Cars.csv

import matplotlib.pyplot as plt

plt.boxplot(cars['MPG'])



plt.hist(cars['MPG'])



From this above box plot and histogram we can say the 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:**

import pandas as pd

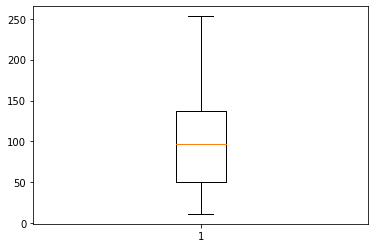
import matplotlib.pyplot as plt

wc\_at = pd.read\_csv('D:\\Study\\Assignments\\wc-at.csv')

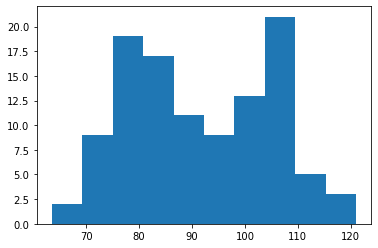
plt.hist(wc\_at["AT"])



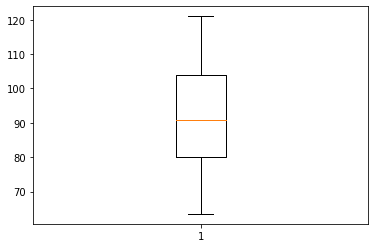
plt.boxplot(wc\_at["AT"])



plt.hist(wc\_at["Waist"])



plt.boxplot(wc\_at["Waist"])



From the above histogram and box plot for both AT & Waist of wc-at data set , it shows that both AT & Waist follows normal distribution.

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

**Ans:**

from scipy import stats

stats.norm.ppf(0.95)

Z scores of 90% confidence interval = 1.65

from scipy import stats

stats.norm.ppf(0.97)

Z scores of 94% confidence interval = 1.89

from scipy import stats

stats.norm.ppf(0.80)

Z scores of 60% confidence interval = 0.85

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

stats.t.ppf(0.975,24)

t score of 95% confidence interval for sample size of 25 = 2.064

from scipy import stats

stats.t.ppf(0.98,24)

t score of 96% confidence interval for sample size of 25 = 2.172

from scipy import stats

stats.t.ppf(0.995,24)

t score of 99% confidence interval for sample size of 25 = 2.797

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:**

Population mean,µ = 270

Sample size, n = 18

Sample mean, x̅ = 260

Standard deviation, s =90

t score = (x̅- µ)/(s/sqrt(n))

=(260-270)/(90/sqrt(18))

= -10/21.23

= -0.47

df= degrees of freedom = n-1 = 18-1= 17

**Probability**

pt(tscore,df)

pt(-0.47,17)

ans = 0.3221639