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

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

🡺If three coins tossed same time

The possible combinations are, HHH,HHT,HTH,THH,TTT,TTH,THT,HTT=8

Number of combinations having Two heads and one tail are=HHT,THH,HTH=3

:probability that two heads and one tail is

**P=3/8 or 0.375**

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

**🡪**

If two Dice are rolled the numbers of outcomes are=6^2=36

a)Probablity of sum equal to one=**0**

b)Probablity that sum is less that or equal to 4 is=(1,1),(1,2)(1,3),(2,1),(2,2),(3,1)=**6/36=1/6**

**c)** Probablity that sum is divisible by 2 and 3 is=6**/36=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?

🡪Total number of balls=7

Number of ways of drawing 2 balls out of 7 balls

(R,R),(R,G),(R,B),(G,B),(G,G),(B,B)=6

=7\*6/2=41/2=21

Now Draw two balls in which no one is blue means are 5 balls(2red,3 Green)

(R,R),(G,G),(R,G),(G,R)=4

=5\*4/2=20/2=10

**the probability that none of the balls drawn is blue=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

🡪 Expected number of candies for a randomly selected child are

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

=**3.09 Candies**

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-**

**🡪**

**A)For Points-**

Mean=3.5965

Median=3.695

Mode=3.92

Variance-=0.285

Std deviation=0.534

**B)For Score-**

Mean=3.2172

Median=3.325

Mode=3.44

Variance-=0.9573

Std deviation=0.9784

**C)For Weigh-**

Mean=17.8487

Median=17.71

Mode=17.02

Variance-=3.1932

Std deviation=1.7869

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?

🡪Sum of all weights=1308

Mean=1308/9=145.33 this is the expected value of wt of that patient

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

**Cars speed and distance**

**Use Q9\_a.csv**

**🡪**

**data1=book1**

**summary(data1);data1**

**installed.packages('dplyr')**

**install.packages('ggplot2')**

**install.packages("moments")**

**skewness(data1)**

**kurtosis(data1)**

|  |  |  |
| --- | --- | --- |
|  | **speed** | **distance** |
| **mu** | **15.4** | **42.98** |
| **sigma** | **5.2876** | **25.76** |
| **Skewness** | **-0.11** | **0.78** |
| **Kurtosis** | **2.42** | **3.24** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**SP and Weight(WT)**

**Use Q9\_b.csv**

**🡪data1=book2**

**summary(data1);data1**

**installed.packages('dplyr')**

**install.packages('ggplot2')**

**install.packages("moments")**

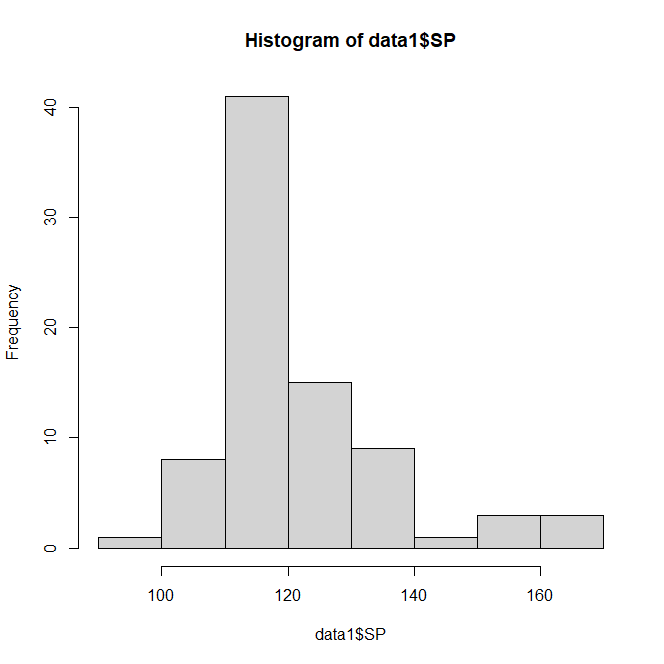
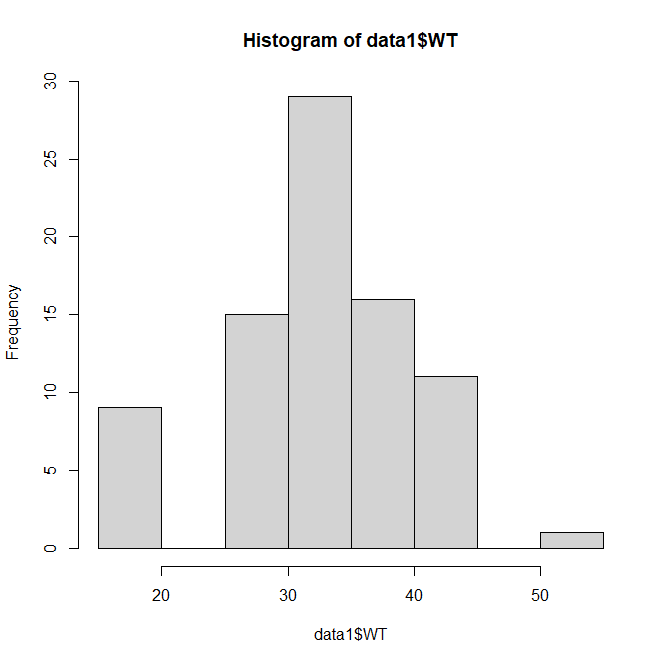
**skewness(data1)**

**kurtosis(data1)**

**hist(data1$WT)**

**hist(data1$SP)**

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| **mu** | **121.54** | **32.41** |
| **sigma** | **14.18** | **7.49** |
| **Skewness** | **1.58** | **-0.60** |
| **Kurtosis** | **5.72** | **3.18** |



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



* The histograms peak has right skew and tail is on right.Mean>median. We have outliers on the higher side.



🡪The boxplot has outliers on the maximum side.

**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?

🡪for 94%

conf\_94=stats.norm.interval(alpha=0.94,loc=200,scale=30/np.sqrt(2000))

print(np.round(conf\_94,0))

print(conf\_94)

-->interval range is(198.73-201.26)

🡪for 98%

conf\_98=stats.norm.interval(alpha=0.98,loc=200,scale=30/np.sqrt(2000))

print(np.round(conf\_98,0))

print(conf\_98)

--interval range is(198.43-201.56)

🡪for 96%

conf\_96=stats.norm.interval(alpha=0.96,loc=200,scale=30/np.sqrt(2000))

print(np.round(conf\_96,0))

print(conf\_96)

--interval range is(198.62-201.36)

**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?

🡪1)Mean=41

Median=40.50

Variance=25.5294

Std Deviation=5.052

2)We don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

🡪 nature of skewness is Symmetric

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

🡪 nature of skewness is Right Skew

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

🡪 nature of skewness is Left Skew

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

🡪Positive kurtosis indicates that more data is crowded around mean value.Chances of finding extreme value is less.

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

🡪Negative kurtosis indicates that more data is crowded around mean value.Chances of finding extreme value is more.

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



What can we say about the distribution of the data?

🡪It is not a normal distribution of data.

What is nature of skewness of the data?

🡪It is left skewed

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

=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.  
🡪Median of both boxplots is same that 260

Both has no outliers.

The boxplots are not skewed in any direction

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)
  3. P (20<MPG<50)

🡪import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

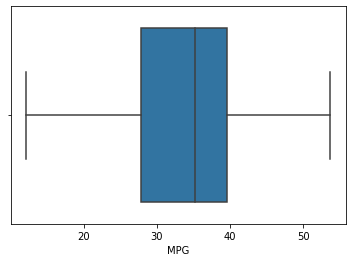
from scipy import stats

from scipy.stats import norm

cars=pd.read\_csv("/content/Cars (1).csv")

cars

sns.boxplot(cars.MPG)



#for P(MPG>38)

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

🡪**0**.**34759**

# P(MPG<40)

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

**🡪0.72934**

# P (20<MPG<50)

stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())

🡪**1.24309 e-05**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

🡪 import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

from scipy.stats import norm

cars=pd.read\_csv("/content/Cars (1).csv")

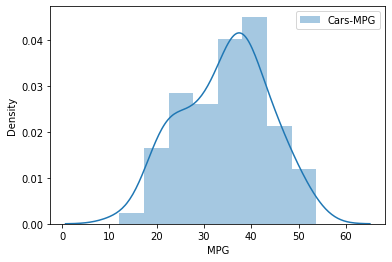
cars

sns.distplot(cars.MPG, label='Cars-MPG')

plt.xlabel('MPG')

plt.ylabel('Density')

plt.legend();



cars.MPG.mean()

**🡪 34.4220**

cars.MPG.median()

**🡪 35.1527**

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

Dataset: wc-at.csv

🡪 import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

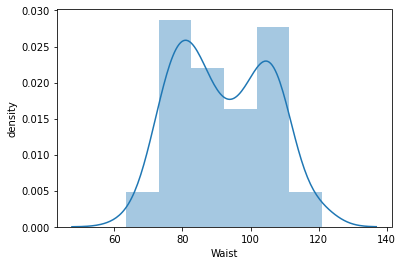
%matplotlib inline

wcat=pd.read\_csv('/content/wc-at (1).csv')

wcat

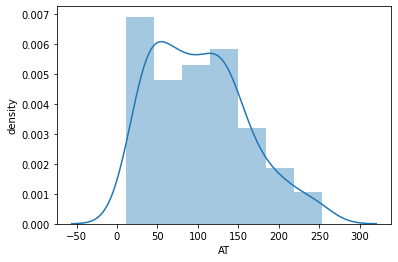
sns.distplot(wcat.Waist)

plt.ylabel('density');



sns.distplot(wcat.AT)

plt.ylabel('density');



wcat.Waist.mean() , wcat.Waist.median()

**- (91.9018, 90.8)**

wcat.AT.mean() , wcat.AT.median()

**- (101.8940, 96.54)**

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

🡪 **from** scipy **import** stats

**from** scipy.stats **import** norm

*# Z-score of 90% confidence interval*

stats**.**norm**.**ppf(0.95)

1.6448

*# Z-score of 94% confidence interval*

stats**.**norm**.**ppf(0.97)

1.8807

*# Z-score of 60% confidence interval*

stats**.**norm**.**ppf(0.8)

0.8416

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

🡪 **from** scipy **import** stats

**from** scipy.stats **import** norm

*# t scores of 95% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.975,24) *# df = n-1 = 24*

* 2.06389

*# t scores of 96% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.98,24)

* 2.1715

*# t scores of 99% confidence interval for sample size of 25*

stats**.**t**.**ppf(0.995,24)

* 2.7969

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

🡺mean x=260

Population mean y=270

Standard Deviation=90

Sample size N=18

**from** scipy **import** stats

**from** scipy.stats **import** norm

*# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days*

*# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days*

*# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))*

t**=**(260**-**270)**/**(90**/**18**\*\***0.5)

t

**-0.4714**

*# Find P(X>=260) for null hypothesis*

*# p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function*

p\_value**=**1**-**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)

p\_value

**0.3216**

*# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function*

p\_value**=**stats**.**t**.**sf(abs(**-**0.4714),df**=**17)

p\_value

**0.3216**