|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | DISCRETE |
| Results of rolling a dice | DISCRETE |
| Weight of a person | CONTINOUS |
| Weight of Gold | CONTINOUS |
| Distance between two places | CONTINOUS |
| Length of a leaf | CONTINOUS |
| Dog's weight | CONTINOUS |
| 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 | INTERVAL |
| Weight | RATIO |
| Hair Color | NOMINAL |
| Socioeconomic Status | ORDINAL |
| Fahrenheit Temperature | INTERVAL |
| Height | RATIO |
| Type of living accommodation | NOMINAL |
| Level of Agreement | ORDINAL |
| IQ(Intelligence Scale) | INTERVAL |
| Sales Figures | RATIO |
| Blood Group | NOMINAL |
| Time Of Day | ORDINAL |
| Time on a Clock with Hands | INTERVAL |
| Number of Children | RATIO |
| Religious Preference | NOMINAL |
| Barometer Pressure | ORDINAL |
| SAT Scores | INTERVAL |
| Years of Education | RATIO |

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

When 3 coins are tossed,

No. of outcomes= 8([HHH], [HHT], [HTH], [HTT], [THH], [THT], [TTH],[TTT])

The occurrence of 2 heads and a tail=3 ([HHT], [HTH], [THH])

The probability = no. of favorable outcomes/ no. of total possible outcomes= 3/8

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

1. Equal to 1 🡺 0
2. Less than or equal to 4 🡺 1/6 (6/36)
3. Sum is divisible by 2 and 3 🡺 1/6 (6/36)

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?

🡺5C2/7C2 = 20/42 = 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 no. of candies for a random child = 1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

=> 0.015+0.8+1.95+0.025+0.06+0.24 = 3.090

Which is,

Therefore no. of candies for a child that is randomly selected is 3.090

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**

Mean = (sum of observations)/(No. of observations)

Median = middle value if n=odd and average of middle 2 values if n=even

Mode = most repeated value

Range= maximum value – minimum value

In R,

#Calculating mean

> mean(Q7$Points)

[1] 3.596563

>

> mean(Q7$Score)

[1] 3.21725

>

> mean(Q7$Weigh)

[1] 17.84875

#Calculating median

|  |
| --- |
| > median(Q7$Points)  [1] 3.695  >  > median(Q7$Score)  [1] 3.325  >  > median(Q7$Weigh)  [1] 17.71  #Calculating mode  # There is no function called mode() to calculate mode of the data  > install.packages('modeest') #modeest=mode estimation  package ‘modeest’ successfully unpacked and MD5 sums checked  The downloaded binary packages are in  C:\Users\kssow\AppData\Local\Temp\RtmpSCtR4d\downloaded\_packages  > library(modeest)  # mfv- most frequent value  > mfv(Q7$Points)  [1] 3.07 3.92  >  > mfv(Q7$Score)  [1] 3.44  >  > mfv(Q7$Weigh)  [1] 17.02 18.90  # Calculating variance  > var(Q7$Points)  [1] 0.2858814  >  > var(Q7$Score)  [1] 0.957379  >  > var(Q7$Weigh)  [1] 3.193166  #Calculating Standard deviation  > sd(Q7$Points)  [1] 0.5346787  >  > sd(Q7$Score)  [1] 0.9784574  >  > sd(Q7$Weigh)  [1] 1.786943  #Calculating Range  > range(Q7$Points)  [1] 2.76 4.93  >  > range(Q7$Score)  [1] 1.513 5.424  >  > range(Q7$Weigh)  [1] 14.5 22.9  #Visualizing the data using histograms  > hist(Q7$Points)    > hist(Q7$Score)      > hist(Q7$Weigh) |
|  |

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?

Pi=1/9

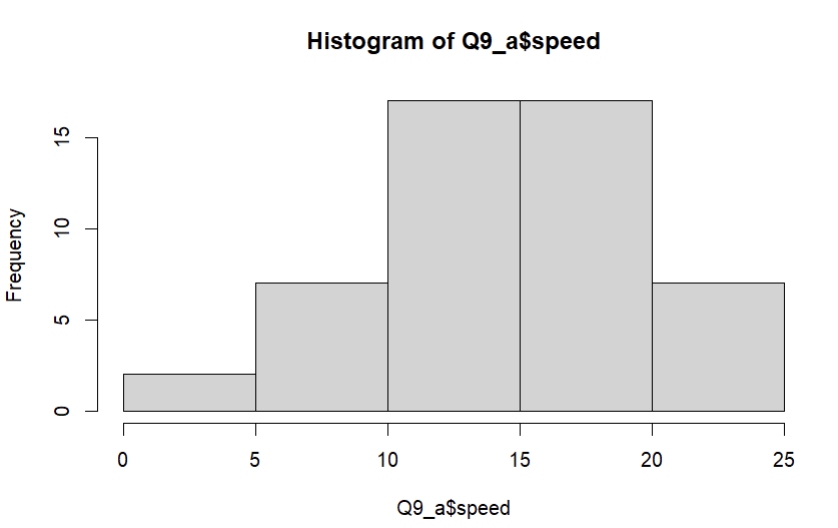
∑Pi\*xi = (1/9) (108+110+123+134+135+145+167+187+199) = (1/9)(1308) = 145.33

Therefore, the expect ed value of weight of the patient = 145.33

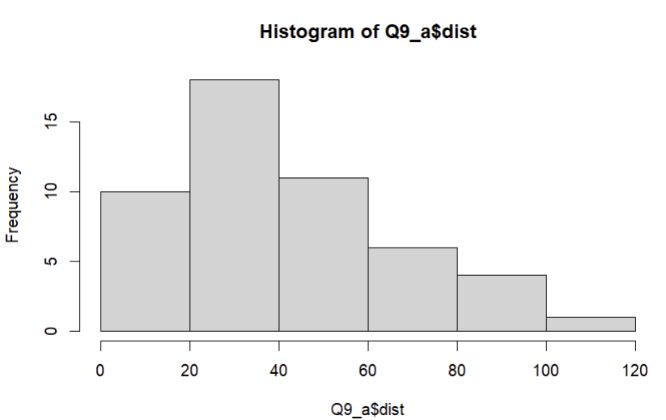
**Q9)** Calculate Skewness, Kurtosis & draw inferences on the following data Cars speed and distance

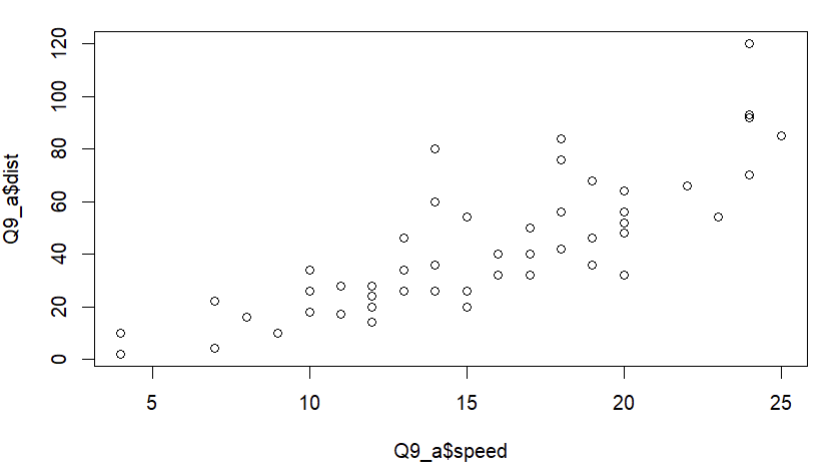
Use Q9\_a.csv







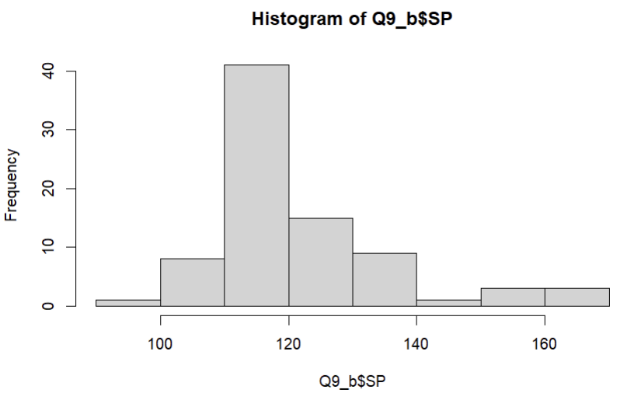




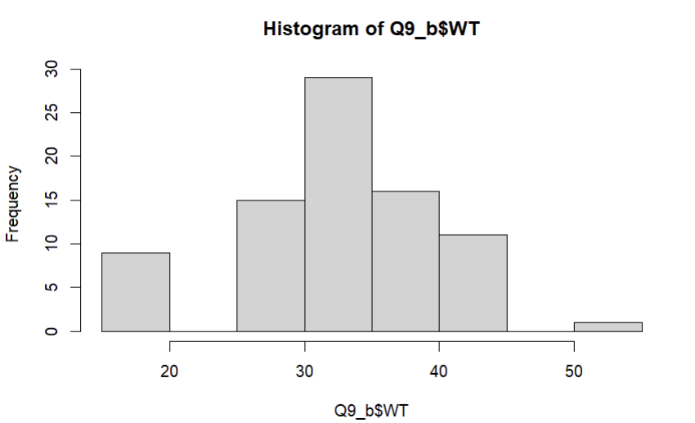
* According to the plot above, the best fitting line between the speed and distance is a straight line.
* Also, the observations in the speed column are left skewed and the max no. of observations are on the right side of the histogram. Therefore, the speed of the most of the cars is above 15.
* Since the kurtosis is less than 3 for WT, the peak is slightly flatter
* The observations in the distance column are right skewed and the maximum no. of observations are on the left side of the histogram. Therefore, most of the cars cover the distance below 60.
* Since the kurtosis is slightly greater than 3 for distance, the peak is slightly sharper and there are very few values concentrated at the tails than usual.

SP and Weight (WT)

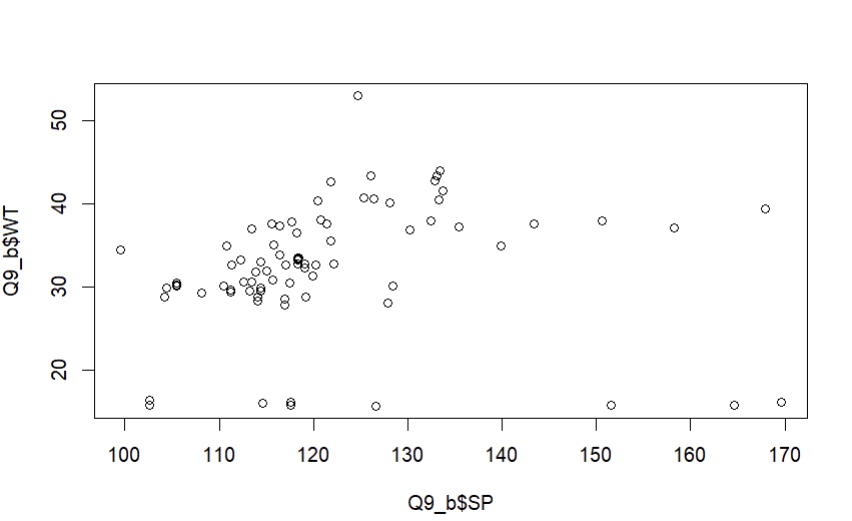
Use Q9\_b.csv











* According to the plot above, the best fitting line between the speed and distance is a straight line.
* Also, the observations in the WT column are left skewed and the max no. of observations are on the right side of the histogram. Therefore, the values of WT are mostly above 25.
* Since the kurtosis is slightly greater than 3 for WT, the peak is slightly sharper and there are very few values concentrated at the tails than usual.
* The observations in the SP column are right skewed and the maximum no. of observations are on the left side of the histogram. Therefore, the values of SP are mostly below 150.
* Since the kurtosis is greater than 3 for SP, the peak is sharper and most of the values concentrated at the tails than usual.

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



* The histogram given is right skewed hence the value of the skewness will be positive.
* As the most of the observations are on the left side of the histogram , indicationg most of the Chick’s weight are below 200.



* It gives us information about lower boundary,1st quartile, median(2nd quartile), 3rd quartile, upper boundary and outliers.
* The given box plot is a left skewed boxplot.
* That is because length of the lower whisker is longer than the upper whisker and the median value is closer to the upper bound.

**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?

stats.norm.interval(0.94,200,30/(2000\*\*0.5))

(198.738325292158, 201.261674707842)

stats.norm.interval(0.96,200,30/(2000\*\*0.5))

(198.62230334813333, 201.37769665186667)

stats.norm.interval(0.98,200,30/(2000\*\*0.5))

(198.43943840429978, 201.56056159570022)

In [ ]:

**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.

X= c (34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)

mean(x)

41

median(x)

40.5

mvf(x)

41

sd(x)

5.052664

1. What can we say about the student marks?

* The average score of all the students is 41
* Most of the students has scored 41 marks
* The marks of 5 students mostly vary from the average score.

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

The data is perfectly symmetrical at the center and skewness=0

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

The data is right skewed and skewness>0

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

The data is left skewed and skewness<0

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

* + Indicates sharper peak
  + Expected kurtosis>3
  + More observations are concentrated around the tail

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

* + The peak is flatter
  + Expected kurtosis<3
  + Less observations are present around the tails

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



What can we say about the distribution of the data?

The data is not perfectly distributed

What is nature of skewness of the data?

Left skewed

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

IQR=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.

The median of boxplot 1 = the median of boxplot 2

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

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

Mean=cars['MPG'].mean()

sd=cars['MPG'].std()

from scipy import stats

print(1-stats.norm.cdf(38,Mean,sd))

print(stats.norm.cdf(40,Mean,sd))

print(stats.norm.cdf(50,Mean,sd)-stats.norm.cdf(20,Mean,sd))

Results:

0.34759392515827137

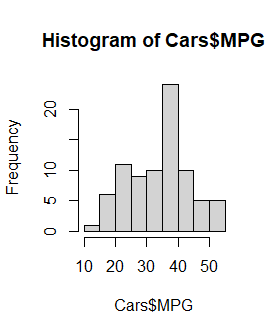
0.7293498762151609

0.8988689169682047

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

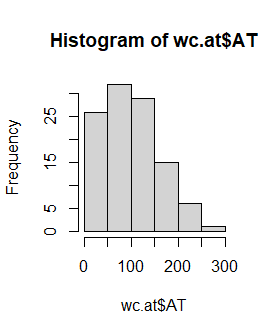
Dataset: Cars.csv

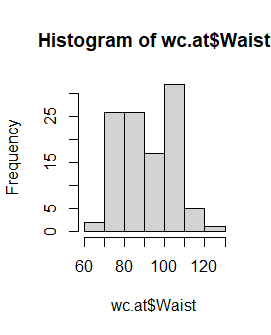


* The data doesn’t follow normal distribution as the curve is not a smooth bell shape

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

Dataset: wc-at.csv





Both of the data doesn’t follow normal distribution

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

stats.norm.ppf(0.90)

stats.norm.ppf(0.94)

stats.norm.ppf(0.60)

Results:

1.2815515655446004

1.5547735945968535

0.2533471031357997

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

stats.t.ppf(0.95,24)

stats.t.ppf(0.96,24)

stats.t.ppf(0.99,24)

Results:

1.7108820799094275

1.8280511719596342

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

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

from math import sqrt

from scipy import stats

#identifing t-score value

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

print('t-score=',t)

#find P(X>=260)

p=1-stats.t.cdf(abs(t),17)

print(p)

Output:

t-score= -0.4714045207910317

0.32167253567098353