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

**Name: - Arsha Vinod**

**Data Science Assignment 1**

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 | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

H = head T = tail

HHH HTH HHT THH

TTT THT TTH HTT

Total no of outcomes = 8

Two heads and one tail = 3

* **3/8**

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
4. Two dices = 6\*6 =36 (Total outcomes)

No of favorable outcomes = 0 since minimum sum is (1, 1) = 2

Hence probability = **0**

1. (1,1) , (2,1) ,(3,1) ,(2,1) ,(2,2) , (3,1) = 6 outcomes

Total outcomes = 36

6/36 = **1/6**

1. (1,5),(2,4),(3,3),(4,2),(5,1),(4,6) = 6 outcomes

Total outcomes = 36

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 balls = 2+3+4 = 7

No of way in which 2 balls draws out of 7

7C2 = = 21

No of non-blue balls = 5

5C2 = = 10

Probability that none of the balls 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 = 1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

= **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**

Mean

cars.mean()

Points 3.596563

Score 3.217250

Weigh 17.848750

Median

cars.median()

Points 3.695

Score 3.325

Weigh 17.710

Mode

cars.mode()

0 3.07

1 3.92

Variance

cars.var()

Points 0.285881

Score 0.957379

Weigh 3.193166

Standard Deviation

cars.std()

Points 0.534679

Score 0.978457

Weigh 1.786943

Range = max – min

points\_range=cars.Points.max()- cars.Points.min()

Points\_Range

score\_range = cars.Score.max()-cars.Score.min()

score\_range

weigh\_range = cars.Weigh.max() - cars.Weigh.min()

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?

No of patients = 9

=

**= 145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

skewness(data$speed)

[1] -0.1139548

skewness(data$dist)

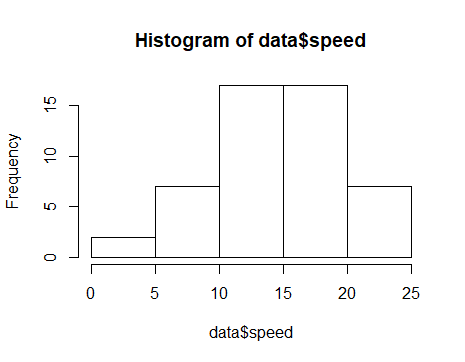
[1] 0.7824835

kurtosis(data$speed)

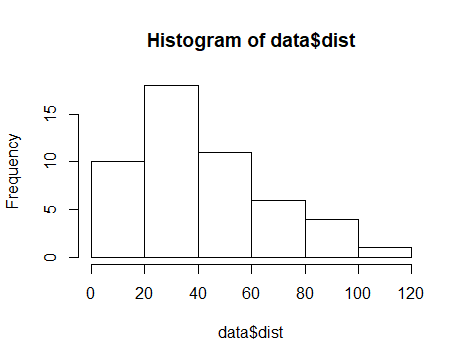
[1] 2.422853

kurtosis(data$dist)

[1] 3.248019



It is negatively skewed.



It is positively skewed.

**SP and Weight(WT)**

**Use Q9\_b.csv**

Skewness(data1$SP)

[1] 1.581454

Skewness(data1$WT)

[1] -0.6033099

kurtosis(data1$SP)

[1] 5.723521

kurtosis(data1$WT)

[1] 3.819466

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



Right side skewed or positive skewed.



The boxplot is a positive skewed and it has lots of outliers towards upper extreme.

**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?

Sample mean = 200

Sample SD = 30

n = 2000

Confidence interval =

94% -> qnorm(0.97) = 1.880794

= 200 1.88 \*

**(198.74 - 201.25)**

98% -> qnorm(0.99) = 2.326348

=200 2.32 \*

**(198.44 - 201.55)**

96% -> qnorm(0.98) = 2.053749

= 200 2.05 \*

**(198.62 – 201.37)**

**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?
3. Mean = mean(num)

=**41**

Median = median(num)

= **40.5**

Variance = var(num)

= 25.52941

Standard Deviation = sd(num)

=5.052664

1. In this the mean is greater than median and the distribution will be skewed towards the right.

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

It will be symmetric and it has zero skewness.

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

The distribution will be positive skewed.

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

The distribution will be negative skewed.

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

The distribution will be longer. The peak will be higher and wider tails

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

The distribution is shorter. The peak will be less and thinner tails

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

What is nature of skewness of the data?

Negative skewed

What will be the IQR of the data (approximately)?   
IQR = upper quartile – lower quartile

= 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 distribution is symmetric.

In both of the boxplot there are no outliers.

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)

Mean -> mean(cars$MPG)

= 34.42208

SD -> sd(cars$MPG)

= 9.131445

1. P(MPG>38)

pnorm(38,34.42,9.13)

0.652513

1-p(mpg<38)= 1-0.65 =0.35

1. P(MPG<40)

Pnorm(40,34.42,9.13)

= 0.7294571

c) P (20<MPG<50)

pnorm(20,34.42,9.13) - pnorm(50,34.42,9.13)

= 0.8989178

Q 21) Check whether the data follows normal distribution

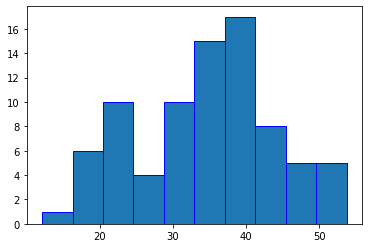
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Yes, it follows normal distribution

plt.hist(cars['MPG'],edgecolor="blue")

plt.show()



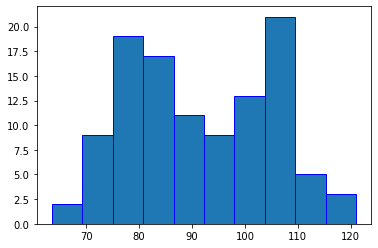
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Yes,Waist and AT follows Normal Distribution.

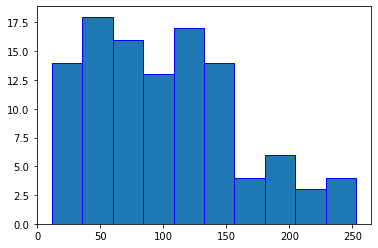
plt.hist(wt['Waist'],edgecolor="blue")

plt.show()



plt.hist(wt['AT'],edgecolor="blue")

plt.show()



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

90% -> qnorm(0.95)

[1] 1.644854

94% -> qnorm(0.97)

[1] 1.880794

60% ->qnorm(0.8)

[1] 0.8416212

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

95% ->

Z value

n=25

df = n-1 =25-1 =24

From the t table =**0.264**

96%- > 1+0.96/2 =0.98

qt(0.98,24)

= **2.171545**

99% ->

n=25

df = n-1 =25-1 =24

From the t table =**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

µ=270, =260, sd = 90, n =18

df = n-1

= 18-1

=17

tscore = = =-10/21.22

= -0.47

pt(-0.47, 17)

= 0.3221639

Probability = **32%**