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

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 | Interval |
| SAT Scores | Interval |
| Years of Education | ratio |

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

Ans: Probability = Favorable outcomes/Total outcomes

Total outcomes= {(TTT)(HHH)(THH)(HTT)(HTH)(THT)(HHT)(TTH)}

We have 3 outcomes with 2 heads and 1 tail

Probability = 3/8

= 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

Ans: a) Equal to 1 = 0

b) Less than or equal to 4 ={(1,1)(1,2)(1,3)(2,1)(2,2)(3,1)}

=6/36=0.166

c) Sum is divisible by 2 and 3 ={(1,5)(2,4)(3,3)(4,2)(6,6)}

=5/36=0.138

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: Balls = {R,R,G,G,G,B,B}

N(s)Number of ways of drawing 2 balls out of 7 = 7C2 = (7x6)/(2x1) = 21

N(E)Event of 2 balls, none of which is blue

Number of ways of drawing 2 balls out of (2+3) balls =5C2=(5x4)/(2x1)=10

Probability = 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

Expected number of candies for a randomly selected child = (1\*0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12)

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24 = 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.

**Ans:**

Mean - Points: 3.596563, Score: 3.21725, Weigh: 17.84875

Median - Points: 3.695, Score: 3.325, Weigh: 17.71

Mode - Points: 3.07 3.92, Score: 3.44, Weigh: 17.02 18.90

Variance - Points: 0.2858814, Score: 0.957379, Weigh: 3.193166

Standard

deviation - Points: 0.5346787, Score: 0.9784574, Weigh:1.786943

Range - Points: 2.17, Score: 3.911, Weigh: 8.4

**Comments:**

1. In points column, median>mean, it means the distribution is negatively skewed.
2. In score column, median>mean, it means the distribution is negatively skewed.
3. In weigh column, mean>median, it means the distribution is positively skewed.

**Use Q7.csv file**



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: Probability of selecting each patient = 1/9

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

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

= 145.33

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

**Cars speed and distance**

**Ans: Skewness**: speed: - 0.1139548 , dist: 0.7824835

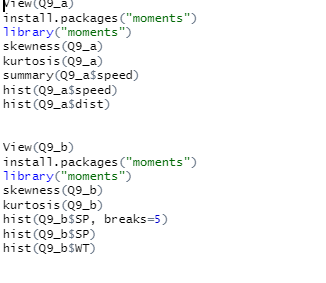
For speed, value of skewness is negative, hence it is left skewed.

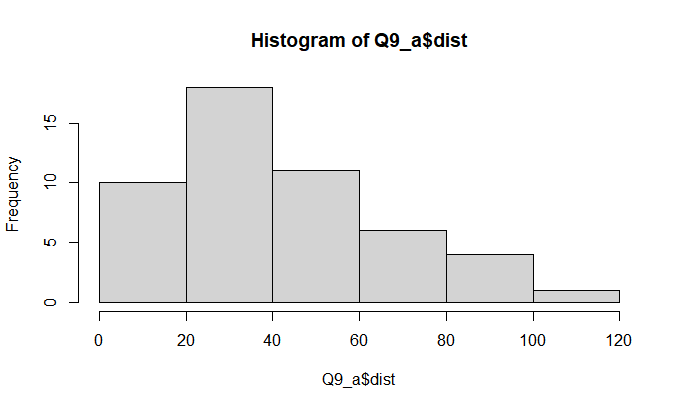
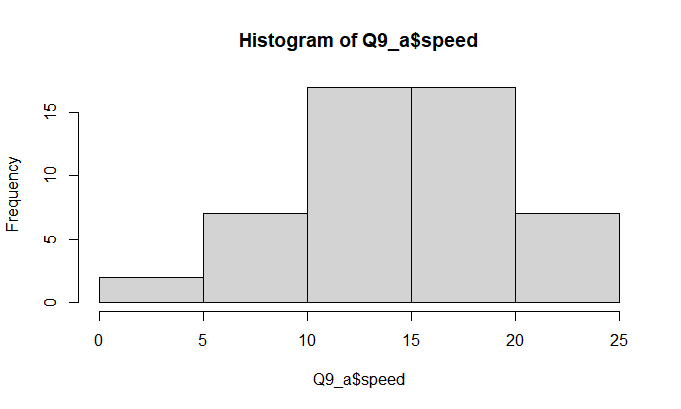
For dist, value of skewness is positive, hence it is right skewed.

**Kurtosis**: speed: 2.422853, dist: 3.248019

For speed, value of kurtosis is <3, somewhat less peaked, has light tails

For dist, value of kurtosis>3, somewhat more peaked, has heavy tails



****

**SP and Weight(WT)**

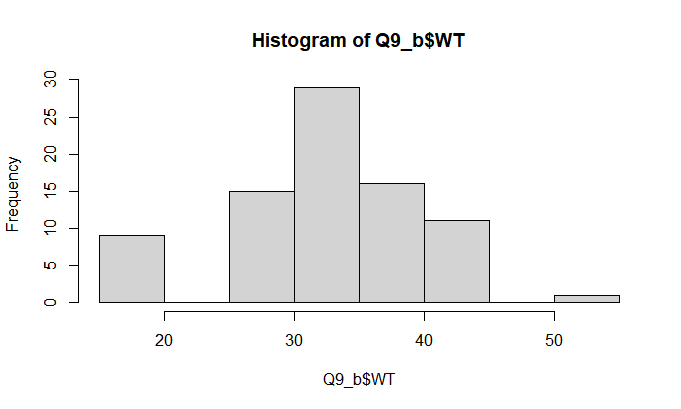
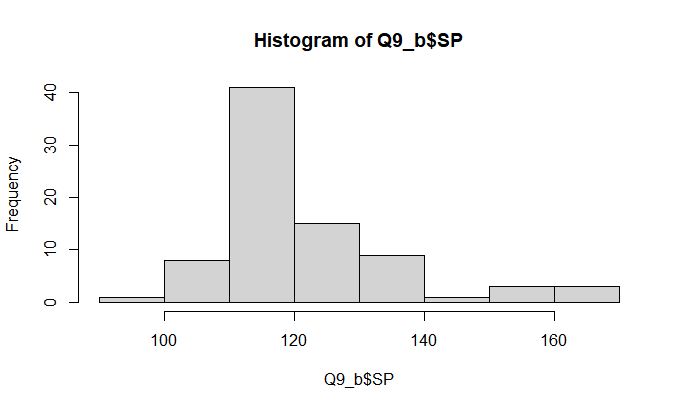
**Skewness:** SP: 1.5814537, WT: - 0.6033099

For SP, value of skewness is positive, hence it is right skewed.

For WT, value of skewness is negative, hence it is left skewed.

**Kurtosis**: SP: 5.723521, WT: 3.819466

Kurtosis values for both SP and WT are >3, which means they have heavy tails and are more peaked.



**Use Q9\_b.csv**

Q10) Draw inferences about the following boxplot & histogram



**Ans**: With the above histogram, we can conclude that it is a right skewed histogram and the skewness value will be always positive. In the interval 50-100 , we could see more number of chick weights.



**Ans:** From the boxplot given we can see lower limit, upper limit, median, 1st quartile, 3rd quartile, the box contains 50% data points, bottom whisker is small in length compared to upper whisker. Few data points have exceeded beyond Upper limit called outliers. Not normally distributed and is right skewed as mean > median.

**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:

Sample n = 2000

Sample mean x = 200

Sample standard deviation = 30

For 94% C.I >stats.norm.interval(0.94,200,30/(2000\*\*0.5)) = (198.7383, 201.2616)

For 98% C.I> stats.norm.interval(0.98,200,30/(2000\*\*0.5))=(198.4394, 201.56056)

For 96% C.I> stats.norm.interval(0.96,200,30/(2000\*\*0.5))=(198.6223, 201.37769)

**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:** Mean: 41, Median: 40.5, variance: 25.52941, standard deviation: 5.052664

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

**Ans:** Skewness=0/Perfectly symmetrical

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

**Ans:** Right skewed/ Positive skewness, most of the data will be lying on left side of the plot.

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

**Ans**: left skewed/negative skewness, most of the data will be lying on the right side of the plot.

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

**Ans**: Thinner /sharp peak and wider tails

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

**Ans:** Flat peak and tinner tails

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



What can we say about the distribution of the data? **Ans**: The data is not equally distributed

What is nature of skewness of the data? **Ans:** Left-skewed since left whisker is larger in length than right whisker.

What will be the IQR of the data (approximately)? **Ans:** 8  
  
  
Q19) Comment on the below Boxplot visualizations?



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

**Ans:** The box plot 2 is highly distributed compared to box plot 1. Both the box plots are symmetrical and median is same for both. In box plot 2 the data range is high and it will be difficult to make any predictions.

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)

Ans:

a) 1-stats.norm.cdf(38,loc=cars.MPG.mean(), scale=cars.MPG.std())

P(MPG>38) = 0.347 or 34.7%

1. stats.norm.cdf(40,loc=cars.MPG.mean(), scale=cars.MPG.std())

P(MPG<40) = 0.729 or 72.9%

1. stats.norm.cdf(50,loc=cars.MPG.mean(), scale=cars.MPG.std())-stats.norm.cdf(20,loc=cars.MPG.mean(), scale=cars.MPG.std())

P (20<MPG<50) = 0.898 or 89.8%

Q 21) Check whether the data follows normal distribution

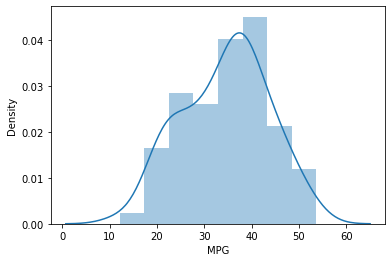
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: cars.MPG.mean() : 34.42

cars.MPG.median(): 35.15

Mean and median are almost same, hence it is normally distributed

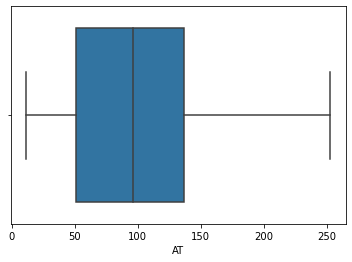


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

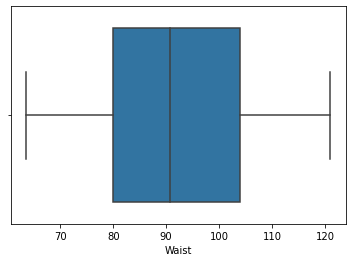
Dataset: wc-at.csv

Ans:

|  |  |  |
| --- | --- | --- |
|  | AT | Waist |
| Mean | 101.89 | 91.90 |
| Median | 96.54 | 90.80 |



For AT, since the right whisker is larger than the left one, it is right skewed.



For waist, both the whiskers are of same length and there’s only slight difference between mean and median, hence it is normally distributed.

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

Ans:

1. Z scores for 90% confidence interval: 1.6449
2. Z scores for 94% confidence interval: 1.8808
3. Z scores for 60% confidence interval: 0.8416

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

Ans: Sample size = 25, DOF = (n-1) = 24

1. T score for 95% confidence interval: 2.06
2. T score for 96% confidence interval: 2.17
3. T score for 99% confidence interval: 2.79

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: Sample size n = 18

Sample mean x = 260 days

Sample standard deviation s = 90 days

= 260-270/90/sqrt of 18

= -0.4714

pt(-0.4714, 18) = 0.321