|  |  |
| --- | --- |
| 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 | CONTINIOUS |
| 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 | INTERVAL |
| Blood Group | NOMINAL |
| Time Of Day | ORDINAL |
| Time on a Clock with Hands | RATIO/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:** HHH, TTT, HHT, TTH,HTH,THT,THH,HTT

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

**Ans:** a) Zero

b) 1/12

c) 5/36

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?

**Ans :** 5C2/7C2 ways.

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

= 0.015 + 1.95 + 0.025 + 0.06 + 0.24

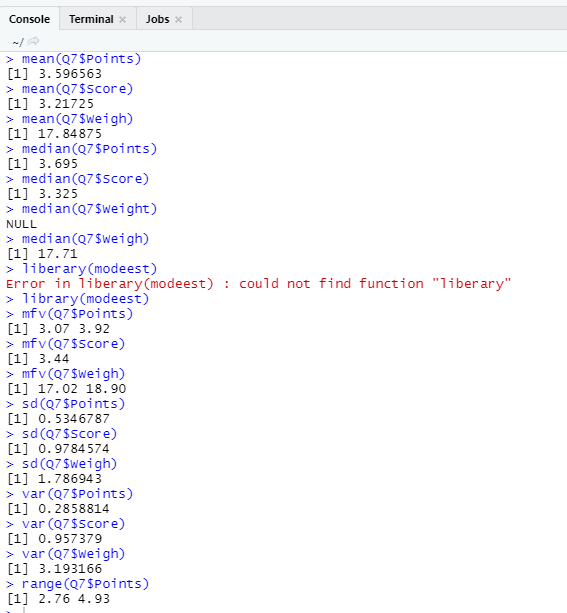
= 3.090

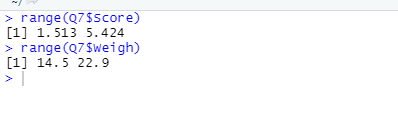
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weigh>

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

**Use Q7.csv file**

****

****

**Conclusion:** when compared to points and scores we can say that the weight column has more deviation or variance, also the values are greater compared to points and weights

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:** Expected value = ∑(probability \* Value)

∑P(x).E(x)

There are 9 patients, the probability of selecting each patient = 1/9

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.

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

=(1/9)(1308)

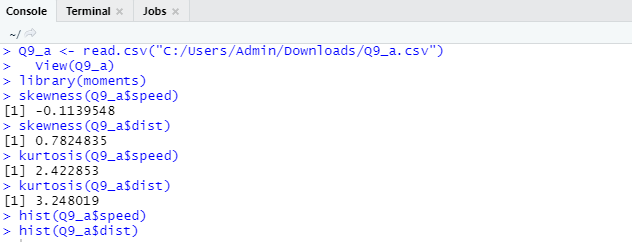
=145.33

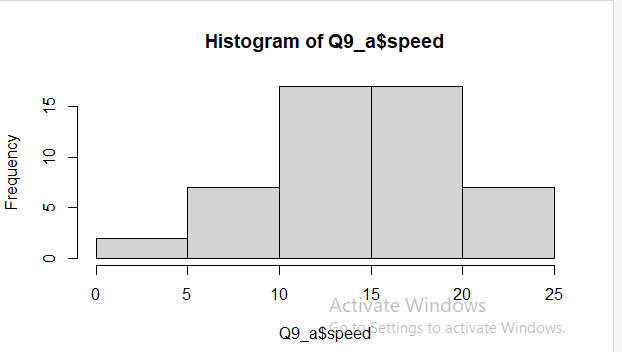
Expected value of weight of that patient = 145.33

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

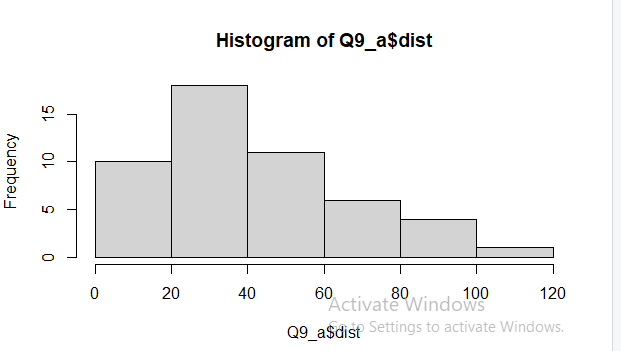
**Cars speed and distance**

**Use Q9\_a.csv**

****

****

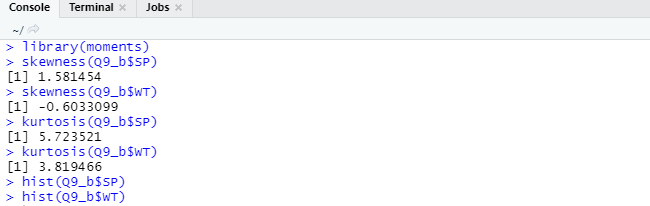
The nature of skewness for speed i.e mean > median implies the graph is left skewed and is negative.

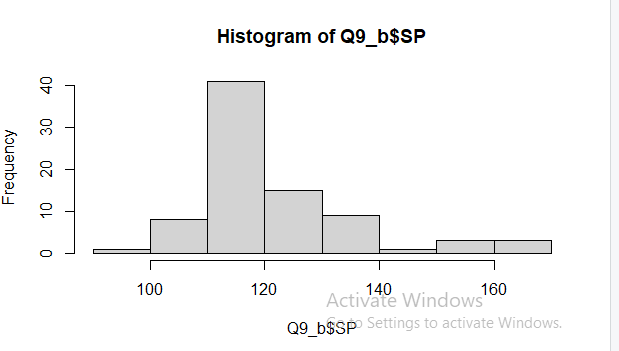
****

The nature of skewness for dist i.e mean <median implies the graph is right skewed and is positive.

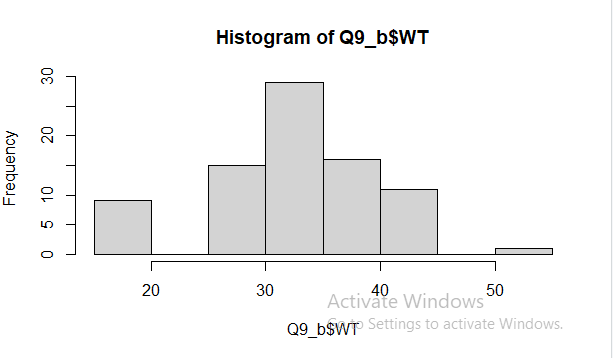
**SP and Weight(WT)**

**Use Q9\_b.csv**

****

****

The nature of skewness for SP i.e mean <median implies the graph is right skewed and is positive.

****

The nature of skewness =0 for WT as mean = median

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



**Ans:** Right Skewed and Positive valued



**Ans:** It is a 5 point charts

**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:** Given: n=2000, x(sample mean)=200, sd (sample)=30

Here 2000 is greater than 30 so the sample is normal and it is a simple random sample. Sample statistic is 200Confidence level is 94%, 98%, 96%.Standard error is 30/√2000= 0.67082

**Alpha (α)** =100-confidence level=6%, 2% & 4%=0.06, 0.02 & 0.04 respectively Critical Probability = 1 – α/ 2 = 0.97, 0.99 & 0.98 respectively

**Degree of freedom** = n - 1 = 2000 - 1 = 1999From t distribution critical values are = 1.882, 2.328 & 2.055 respectively

Margin of error (**94% confidence level**) = critical value (94% confidence level) \* standard error = 1.882 \* 0.67082 = 1.262

Margin of error (**98% confidence level**) = critical value (98% confidence level) \* standard error = 2.328 \* 0.67082 = 1.561

Margin of error (**96% confidence level**) = critical value (96% confidence level) \* standard error = 2.055 \* 0.67082 = 1.378

**Confidence interval = Sample statistics ±Margin of error**

Confidence interval (94% confidence level) = 200 ±1.262 =198.738 to 201.262

Confidence interval (98% confidence level) = 200 ±1.561 =198.439 to 201.561

Confidence interval (96% confidence level) = 200 ±1.378 = 198.622 to 201.378

**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:** Mean(μ) = (34 + 36 + 36 + 38 + 38 + 39 + 39 + 40 + 40 + 41 + 41 + 41 +41 + 42 + 42 + 45 + 49 + 56) / 18 = 738 / 18 = 41

Median = (40 + 41) / 2 = 40.5

Variance (σ2) = [(34-41)2+ (36-41)2+ (36-41)2+ (38-41)2+ (38-41)2+ (39-41)2+ (39-41)2+ (40-41)2+ (40-41)2+ (41-41)2+ (41-41)2+ (41-41)2+ (41-41)2+(42-41)2+ (45-41)2+ (45-41)2+ (49-41)2+ (56-41)2] / 18 = 434/18 = 24.111

Standard Deviation (σ) = (24.111)1/2= 4.9102)

1. What can we say about the student marks?

**Ans:** According to marks we can say that students who scored in between 34-39 are 7 in count, students who scored in between 40-45 are 9 in count, student who scored in between 46-51 is 1 in count and student who scored in between 52-57 is 1 in count. Hence if we calculate skewness and kurtosis that will be +1.686 and +3.953 respectively and are positively skewed and kurtosis is also positive.

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

**Ans:** Perfectly Symmetric

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

**Ans:** Right Skewed (Positive)

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

**Ans:** Left Skewed (Negative)

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

**Ans:** Positive Kurtosis means Peak is sharp

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

**Ans:** Negative Kurtosis means peak is flat.

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



What can we say about the distribution of the data?

**Ans:** The minimum of our data is <2 i.e we have at least one value <2 in our data

The max value is >18 i.e we have at least one value in our data set >18.

The median=x<16 but x>14. 50% of our data lies between 10 and 18

What is nature of skewness of the data?

**Ans:** Left Skewed and negative

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

**Ans:** Q1-Q3 = IQR

IQR=18-10=8

Q19) Comment on the below Boxplot visualizations?

Ans:



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

**Ans:** Both the boxplots are symmetric and have same median value around 262.5

Boxplot 1) have a range of values from 237 to 287 and Boxplot 2) have range of values between <200 to 325 i.e the variance is less for boxplot 1) when compared to boxplot 2) and hence boxplot 1) is more efficient to predict.

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)

**a=subset (MPG, MPG>38)**

**Show(a)**

**Ans: 33/81**

* 1. P(MPG<40)

**b=subset (MPG, MPG<40)**

**Show(b)**

**Ans: 67/81**

* 1. P (20<MPG<50)

**c=subset (20<MPG<50)**

**Show(c)**

**Ans: 69/81**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

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

**Ans: from scipy import stats**

Z score for 90%stats.norm.ppf (0.95) = 1.644

Z score for 94% stats.norm.ppf (0.94) = 1.55

Z score for 60% 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

**Ans: from scipy import stats**

T scores for 95% stats.t.ppf (0.95, df=24) = 1.710

T scores for 96% stats.t.ppf (0.98, df=24) = 2.171

T scores for 99% stats.t.ppf (0.995, df=24) = 2.796

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:** df = 18-1 =17

t - Critical value = stats.t.ppf(0.975, df=17) = 2.109

t – Statistic value = (260-270)/(90/np.sqrt(18)) = -0.4714

here t-value is < t value obtained with 17 degree of freedom and a t score of -0.471

the probability of bulbs lasting less than 260 days on **average of 0.3218** assuming the mean life of the bulbs is 300 days.