Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| 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** |

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 | **Nominal** |
| Fahrenheit Temperature | **Interval** |
| Height | **Ratio** |
| Type of living accommodation | **Nominal** |
| Level of Agreement | **Ordinal** |
| IQ(Intelligence Scale) | **Ordinal, Interval** |
| Sales Figures | **Ratio** |
| Blood Group | **Nominal** |
| Time Of Day | **Ratio** |
| 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?

**Let ‘S’ be the sample–space. Then S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}**

**Interested event: 3 (two heads and one tail) {HHT, HTH, THH}**

**Total Events: 8**

**Probability: 3/8**

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

1. Equal to 1--- **P(E) = n(E) / n(S) = 0 / 36 = 0**
2. Less than or equal to 4---- **P(E) = n(E) / n(S) = 3 / 36 = 1 / 12**
3. Sum is divisible by 2 and 3 --- **P(E) = n(E) / n(S) = 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?

**Interested events --- 10 (5c2)**

**Total Events --- 21(7c2)**

**Probability --- 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**

**= XPA + XPB + XPC+ XPD+ XPE+ XPF**

**=3.09**

|  |  |  |  |
| --- | --- | --- | --- |
| **CHILD** | **Candies count** | **Probability** | **XP** |
| **A** | **1** | **0.015** | **0.015** |
| **B** | **4** | **0.2** | **0.8** |
| **C** | **3** | **0.65** | **1.95** |
| **D** | **5** | **0.005** | **0.025** |
| **E** | **6** | **0.01** | **0.06** |
| **F** | **2** | **0.12** | **0.24** |
|  |  | **Total** | **3.09** |

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

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



**Ans:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dispersion** | **Points** | **Score** | **Weigh** |
| **Mean** | **3.596563** | **3.21725** | **17.84875** |
| **Mode** | **3.92** | **3.44** | **17.02** |
| **Median** | **3.695** | **3.325** | **17.71** |
| **Variance** | **0.2858814** | **0.957379** | **3.193166** |
| **Standard Deviation** | **0.5346787** | **0.9784574** | **1.786943** |
| **Range** | **2.76 - 4.93** | **1.513 - 5.424** | **14.5 - 22.9** |
| **Range Value(max-min)** | **2.17** | **3.911** | **8.4** |

Q8) Calculate Expected Value for the problem below

The weights (X) of patients at a clinic (in pounds), are 108, 110, 123, 134, 135, 145, 167, 187, and 199. Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**Ans:**

**You can think of an expected value as a mean, or average, for a probability distribution.**

**The mean is = Expected Value**

**Expected Value =**

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

**= 145.333.**

**It is not mandatory for the expected value to be present in sample space. As calculated above Expected Value of 145.33 is not present in our sample space.**

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

Cars speed and distance



**SP and Weight(WT)**



**Ans:**

**Skewness for speed= -0.8448909,**

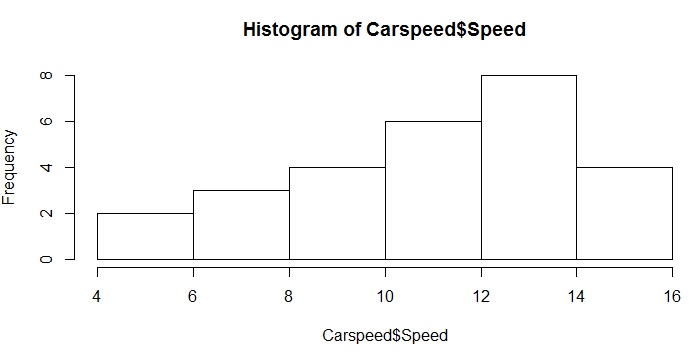
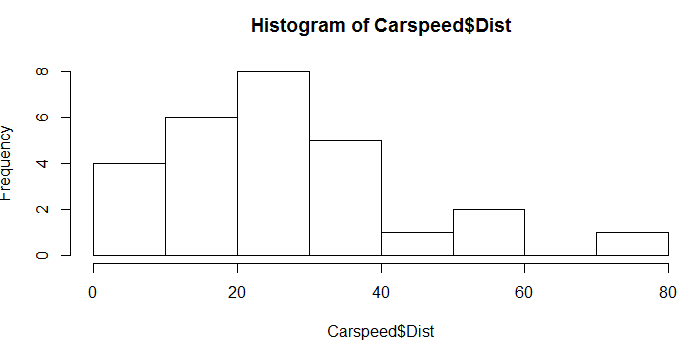
**Skewness value is negative so it is left skewed.**

**Kurtosis for speed=** **2.991396**

**And for Skewness for distance = 1.217917.**

**Skewness for distance is right skewed as it is Positive.**

**Kurtosis for distance = 4.816933**

**Skewness for SP= -0.4076944,**

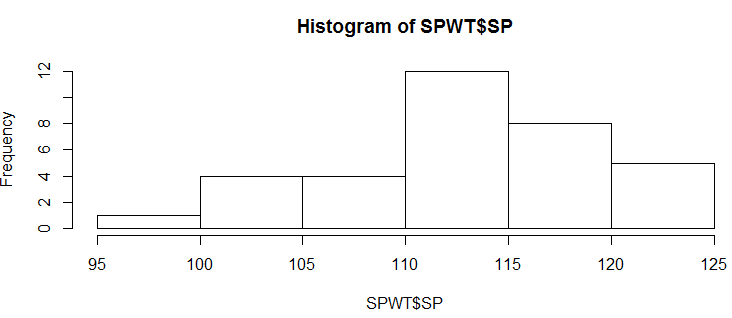
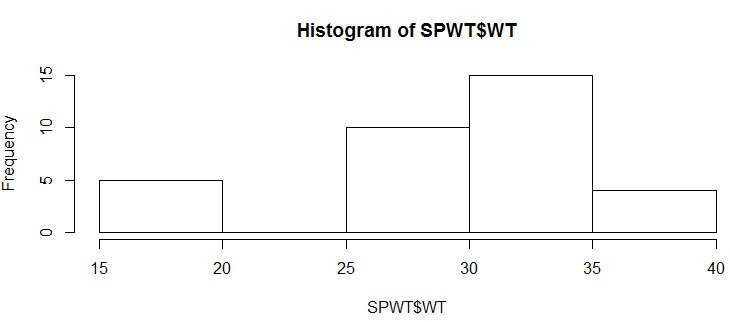
**Skewness value is negative so it is left skewed.**

**Kurtosis for SP= 2.086738**

**And for Skewness for WT= -1.28736**

**Skewness for distance is left skewed as it is negative.**

**Kurtosis for WT= 3.818813**

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



**Ans:**

**The most of the data points are concentrated in the range 50-100 with frequency 200.**

**And least range of weight is 400 somewhere around 0-10.**

**So the expected value the above distribution is around 75.**

**Skewness: - We can notice a long tail towards right so it is heavily right skewed.**



**Ans:**

**A boxplot is a graph that gives a good indication of how the values in the data are spread out.**

**First Quartile shows less data points.**

**Interquartile data points reflect Median is less than Mean, so data indicates right Skewness.**

**It seems the greater outliers on the upper side of box plot causing Skewness.**

**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:**

**As problem describes about average weight and we can assume normality from the setup, as the judgment for normality in a sample mean problem is the sample size.**

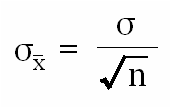
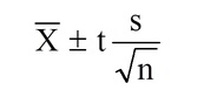
**Is *n* (the sample size) > 30? 2,000> 30, so it is normal.**

**The sample is normal and is a simple random sample.**

**Since we are trying to estimate the mean weight in the population, we choose the mean weight in our sample (200) as the sample statistic.**

**In this case, the confidence level is defined for us in the problem. We are working with a 94%, 96% and 98% confidence level.**

**Standard Error Formula: Margin of Error Formula:**

1. **The standard error:**= 30/(square root of 2000) =0 .67082
2. **The**[**degrees of freedom**](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom)**(df):** df = n - 1 = 2000 - 1 = 1999
3. **T value Calculation or critical value calculation:**

**Case1: 94% confidence Interval**

* Compute alpha (α):
  + For 94 % Confidence level α = 1 - (confidence level / 100) = 0.06
* The critical probability (p\*):
  + P\* = 1 – α/2 = 1 - 0.06/2 = 0.97

The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.97.

* From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 1.882.

The range of the confidence interval is defined by the

**Sample statistic +/- Margin of error**

And the uncertainty is denoted by the confidence level.

Therefore, this 94% confidence interval is 200 +/- 1.882.

Range= [198.118 201.882]

**Case2: 96% confidence Interval**

* Compute alpha (α):
  + For 96 % Confidence level α = 1 - (confidence level / 100) = 0.04
* The critical probability (p\*):
  + P\* = 1 – α/2 = 1 - 0.04/2 = 0.98
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.98.

From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 2.055.

The range of the confidence interval is defined by the

**Sample statistic +/- Margin of error**

And the uncertainty is denoted by the confidence level.

Therefore, this 94% confidence interval is 200 +/- 2.055.

Range= [197.945 202.055]

**Case3: 98% confidence Interval**

* Compute alpha (α):
  + For 98 % Confidence level α = 1 - (confidence level / 100) = 0.02
* The critical probability (p\*):
  + P\* = 1 – α/2 = 1 - 0.02/2 = 0.99
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.99.
* From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 2.328.

The range of the confidence interval is defined by the

**Sample statistic +/- Margin of error**

And the uncertainty is denoted by the confidence level.

Therefore, this 94% confidence interval is 200 +/- 2.328.

Range= [197.672 202.328]

**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?

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean** | **Median** | **Variance** | **Standard Deviation** |
| **41.0** | **40.5** | **24.1111** | **4.9103** |

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

**Skewness will be zero when mean and median of the data are equal.**

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

**Skewness is asymmetric and data will be skewed to right (positive skewed)**

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

**Skewness is asymmetric and data will be skewed to left (negative skewed)**

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

**The distribution of the data has heavier tails and presence of outliers than the normal distribution.**

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

**Kurtosis cannot negative, as it is having 4th power in its equation. It will be always positive. So, excess kurtosis term will be used as kurtosis for normal distribution is 3. The kurtosis value less than 3 indicates, the distribution of the data has lighter tails and lack of outliers than the normal distribution.**

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



What can we say about the distribution of the data?

**Sol: Let’s assume above box plot is about Age of the students in a school.**

**50% of the people are above 10 yrs. old and remaining is less. And students whose age is above 15 are approx. 40%.**

What is nature of Skewness of the data?

**Sol: Left skewed, median is greater than mean.**

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

**The Inter Quartile Range (IQR) is approximately 8 (18-10).**Q19) Comment on the below Boxplot visualizations?



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

Ans:

**The Mean and Median values are equal for both the Boxplot1 and Boxplot2. So it reflects their will be symmetrical in nature.**

**By observing both the plots, whisker’s level is high in boxplot 2 compare to Box plot 1.**

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)

**Ans:**

PMPG <- read.csv("D:/Data\_science/Assignments/Cars.csv")

library(dplyr)

library(psych)

library(modeest)

library(XML)

library(prob)

mean=mean(PMPG$MPG) # calculation of mean **# 34.42**

sigma=sd(PMPG$MPG) # calculation of standard deviation **# 9.13**

#pnorm=z=(X-mean)/sigma

#pnorm() gives z proportion below the mentioned value, so if we needs to get proportion

#above, it needs to subtract from 1

# Greater than==GT, Less Than==LT

prob\_MPG\_GT\_38 = 1-pnorm(38,mean,sigma) **# 0.34**

prob\_MPG\_LT\_40 = pnorm(40,mean,sigma) **# 0.73**

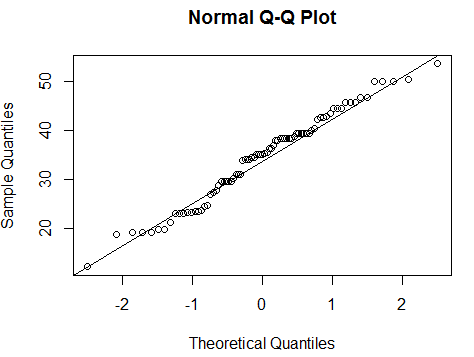
prob\_MPG\_GT\_20\_LT\_50= pnorm(50,mean,sigma)-pnorm(20,mean,sigma) **# 0.89**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

**Dataset: Cars.csv**

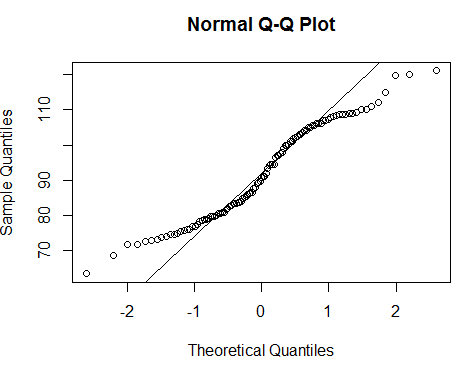
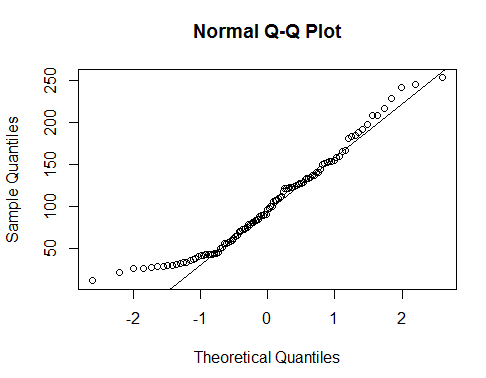
**Ans: The MPG of Cars does not follow Normal Distribution.**



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

**Dataset: wc-at.csv**

**Ans: The Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set does not follow Normal Distribution.**

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

**Ans:**

**Z score of 90% confidence interval is 1.65**

**Z score of 94% confidence interval is 1.55**

**Z score of 60% confidence interval is 0.85**

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

**Ans: t scores**

**For 95%= 1.96**

**For 96%= 2.5**

**For 99% = 2.47**

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:**

**t = [ x - μ ] / [ s / sqrt( n ) ]**

**p <- pt((260-270)/(90/sqrt(18)), 18) #0.3215**