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

**Q1) Identify the Data type for the Following:mukesh@excelr.com**

**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** | **Ordinal** |
| **Level of Agreement** | **Ordinal** |
| **IQ(Intelligence Scale)** | **Ratio** |
| **Sales Figures** | **Ratio** |
| **Blood Group** | **Nominal** |
| **Time Of Day** | **Interval** |
| **Time on a Clock with Hands** | **Interval** |
| **Number of Children** | **Nominal** |
| **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?**

**Interested number of outcomes = 3(two heads and one tail)**

**Total number of outcomes = 8**

**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**
4. **Interested number of events = 0**

**Total number of events = 36**

**Probability = 0/36 = 0**

**B) Interested number of events = 6**

**Total number of events = 36**

**Probability = 6/36 = 1/6 =0.1666**

1. **Interested number of events = 6**

**Total number of outcomes = 36**

**Probability = 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 number of events = 10(4+6+0)**

**Total number of events = 21(6+6+9)**

**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 value of child A = 1 \* 0.015 = 0.015**

**Expected value of child B = 4 \* 0.20 = 0.8**

**Expected value of child C = 3 \* 0.65 = 1.95**

**Expected value of child D = 5 \* 0.005 = 0.025**

**Expected value of child E = 6 \* 0.01 = 0.06**

**Expected value of child F = 2 \* 0.120 = 0.24**

**Expected number of candies for a randomly selected child is 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**

**file = pd.read\_csv("D:/study docs/Excel R/assignment/Assignment 1/Q7.csv")**

**file.mean()**

**Points 3.596563**

**Score 3.217250**

**Weigh 17.848750**

**dtype: float64**

**file.std()**

**Points 0.534679**

**Score 0.978457**

**Weigh 1.786943**

**dtype: float64**

**file.median()**

**Points 3.695**

**Score 3.325**

**Weigh 17.710**

**dtype: float64**

**file.var()**

**Points 0.285881**

**Score 0.957379**

**Weigh 3.193166**

**dtype: float64**

**file.Points.mode()**

**0 3.07**

**1 3.92**

**dtype: float64**

**file.Score.mode()**

**0 3.44**

**file.Weigh.mode()**

**0 17.02**

**1 18.90**

**points\_range = file.Points.max() - file.Points.min()**

**points\_range**

**2.17**

**score\_range = file.Score.max() - file.Score.min()**

**score\_range**

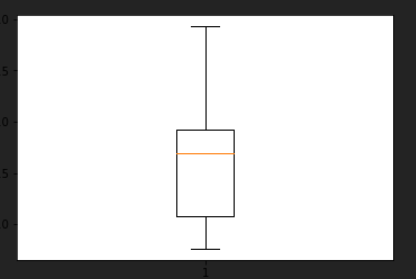
**3.9110000000000005**

**weigh\_range = file.Weigh.max() - file.Weigh.min()**

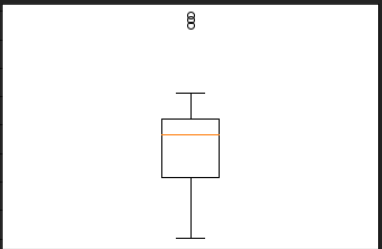
**weigh\_range**

**8.399999999999999**

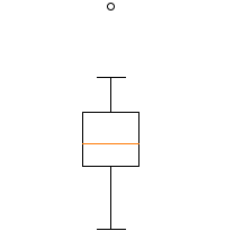
**plt.boxplot(file.Points)**

****

**plt.boxplot(file.Score)**

****

**plt.boxplot(file.Weigh)**

****

**For points column:**

**As mean and median are almost the same the data is quite symmetrical I.e data are evenly distributed.**

**Two modes illustrates that means it is bimodal.**

**Here,Low variance suggests that data points are close to each other and to the mean.**

**Here,Low standard deviation shows that data are clustered around the mean.**

**For score column:**

**As mean and median are almost the same the data is quite symmetrical I.e data are evenly distributed.**

**One mode shows that the data is unimodal.**

**Here,Low variance suggests that data points are close to each other and to the mean.**

**Here,Low standard deviation shows that data are clustered around the mean.**

**There are three outliers present.**

**For weigh column**

**As mean and median are almost the same the data is quite symmetrical I.e data are evenly distributed.**

**Two modes illustrates that means it is bimodel.**

**Here,Low variance suggests that data points are close to each other and to the mean.**

**Here,Low standard deviation shows that data are clustered around the mean.**

**There is an outlier.**

**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?**

**Probability of selecting each patient = 1/9**

**Ex  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**

**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)**

**= (1/9)  (  1308)**

**= 145.33**

**Expected Value of the Weight of that patient = 145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**file <- read.csv("D:/study docs/Excel R/CSV/Q9\_a.csv")**

**install.packages("e1071")**

**library(e1071)**

**skewness(file$speed)**

**#skewness of speed is -0.1105533**

**kurtosis(file$speed)**

**#kurtosis of speed is -0.5089944204057617**

**Speed column:**

**Here, negative skewness indicates that the mass of the distribution is concentrated on the right.**

**Here, negative kurtosis indicates that it has wider peak and thinner tail.**

**skewness(file$dist)**

**#skewness of dist is 0.80389**

**kurtosis(file$dist)**

**#kurtosis of dist is 0.4050525816795765**

**For dist column:**

**Here, positive skewness tells that the distribution is concentrated towards the left.**

**Here, positive kurtosis tells us that the distribution is peaked.**

**Q9\_b**

**file <- read.csv("D:/study docs/Excel R/CSV/Q9\_b.csv")**

**install.packages("e1071")**

**library(e1071)**

**skewness(file$SP)**

**#skewness of SP is 1.6114501961773586**

**kurtosis(file$SP)#can have outliers**

**#kurtosis of SP is 2.9773289437871835**

**For SP columns**

**Here, positive skewness tells that the distribution is concentrated towards the left.**

**Here, positive kurtosis tells us that the distribution is peaked.**

**skewness(file$WT)**

**#skewness of WT is -0.6147533255357768**

**kurtosis(file$WT)**

**#kurtosis of WT is 0.9502914910300326**

**For WT column:**

**Here, negative skewness indicates that the mass of the distribution is concentrated on the right.**

**Here, positive kurtosis tells us that the distribution is peaked.**

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

****

**The data is positively skewed**

**Most of chick weigh between 50 and 150**



**The data is positively skewed**

**There are outliers in the upper side**

**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?**

**Size of population = 3,000,000**

**Size of sample(n) = 2000**

**Mean of sample = 200**

**Std of sample = 30**

**import pandas as pd**

**import numpy as np**

**from scipy import stats**

**stats.norm.interval(0.94,loc = 200, scale = 30/np.sqrt(2000))**

**(198.738325292158, 201.261674707842)**

**stats.norm.interval(0.98,loc = 200, scale = 30/np.sqrt(2000))**

**(198.43943840429978, 201.56056159570022)**

**stats.norm.interval(0.96,loc = 200, scale = 30/np.sqrt(2000))**

**(198.62230334813333, 201.37769665186667)**

**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?**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**scores = pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])**

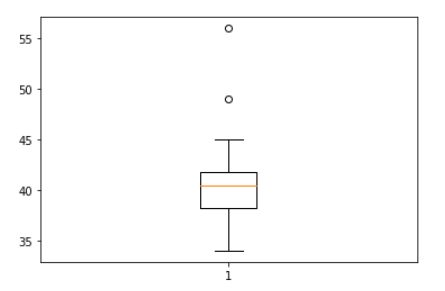
**scores.mean()#mean is 41.0**

**scores.var()#variance is 25.529411764705884**

**scores.median()#Median is 40.5**

**scores.std()#standard deviation is 5.05266382858645**

**plt.boxplot(scores)**

****

**II)Mean > Median, this implies that the distribution is slightly skewed**

**towards right.There are two outliers in student marks 49 and 56 and majority of the marks lie between 38-42.**

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

**If mean and median are equal,the distribution of skewness is 0 and there is normal distribution.**

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

**If mean is greater than median,the distribution is positively skewed.**

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

**If mean is greater than median,the distribution is negatively skewed.**

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

**Positive kurtosis indicate that there are outliers or that the distribution is peaked and has thick tails.**

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

**Negative kurtosis indicate wider peaks and thinner tails.**

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



**What can we say about the distribution of the data?**

**The distribution of data is concentrated more towards the right,is not symmetric(not normally distributed) and median is greater than the mean.**

**What is nature of skewness of the data?**

**The distribution is negatively skewed.**

**What will be the IQR of the data (approximately)?   
IQR = 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.**

**As the boxplot2 is longer than boxplot1 the data is more dispersed whereas data is less dispersed for boxplot1.**

**Larger range for boxplot2 indicates wider distribution and more scattered data.**

**Both boxplots are symmetric.**

**Both have the same median and no outliers are present.**

**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**

**import numpy as np**

**from scipy import stats**

**cars = pd.read\_csv('D:/study docs/Excel R/CSV/cars.csv')**

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

**#P(MPG>38) = 0.3475939251582705**

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

**#P(MPG<40) = 0.7293498762151616**

1. **1- stats.norm.cdf(0.20,loc = cars.MPG.mean(),scale = cars.MPG.std())- stats.norm.cdf(0.50,loc = cars.MPG.mean(),scale =cars.MPG.std())**

**# P (20<MPG<50) = 0.9998091179940826**

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

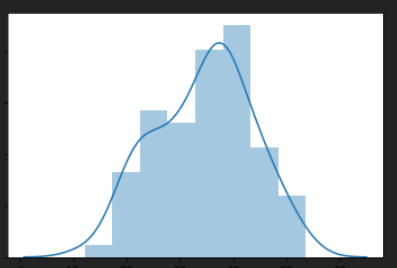
**Dataset: Cars.csv**

**import pandas as pd**

**import seaborn as sns**

**cars = pd.read\_csv("D:/study docs/Excel R/CSV/Cars.csv")**

**sns.distplot(cars.MPG)**

****

**cars.MPG.mean()#mean is 34.422075728024666**

**cars.MPG.median()#median is 35.15272697**

**As the values of mean and median are almost the same we can say that MPG approximately normal distribution but not an ideal normal distribution.**

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

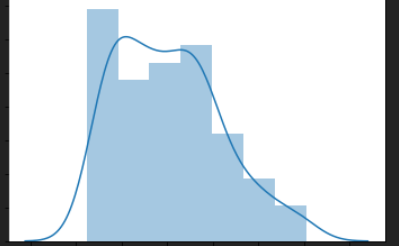
**Dataset: wc-at.csv**

**import pandas as pd**

**import seaborn as sns**

**wc = pd.read\_csv("D:/study docs/Excel R/CSV/WC\_AT.csv")**

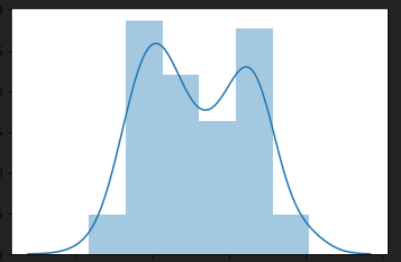
**sns.distplot(wc.AT)**

****

**wc.AT.mean()#mean is 101.89403669724771**

**wc.AT.median()#median is 96.54**

**sns.distplot(wc.Waist)**

****

**wc.Waist.mean()#mean is 91.90183486238533**

**wc.Waist.median()#median is 90.8**

**Since for AT and Waist the mean and median are roughly equal we can say that they follow normal distribution**

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

**from scipy import stats**

**from scipy.stats import norm**

**stats.norm.ppf(0.95)**

**1.6448536269514722**

**stats.norm.ppf(0.97)**

**1.8807936081512509**

**stats.norm.ppf(0.8)**

**0.8416212335729143**

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

**stats.t.ppf(0.975,df = 24)**

**2.0638985616280205**

**stats.t.ppf(0.98,df = 24)**

**2.1715446760080677**

**stats.t.ppf(0.995,df = 24)**

**2.796939504772804**

**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**

**mean of sample =  260**

**μ = 270**

**s = 90**

**n = number of items in the sample = 18**

**t = 260 - 270/(90/sqrt(18)) = -0.471**

**Degrees of freedom is 17**

**rcode <- pt(-0.471,17)**

**rcode**

**0.321814**

**Probability is 0.321814**