|  |  |
| --- | --- |
| 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: Saloni Surendra Alshi**

**Assignment 1:** **Basic Statistics\_Level 1**

**Q1) Identify the Data type for the Following:**

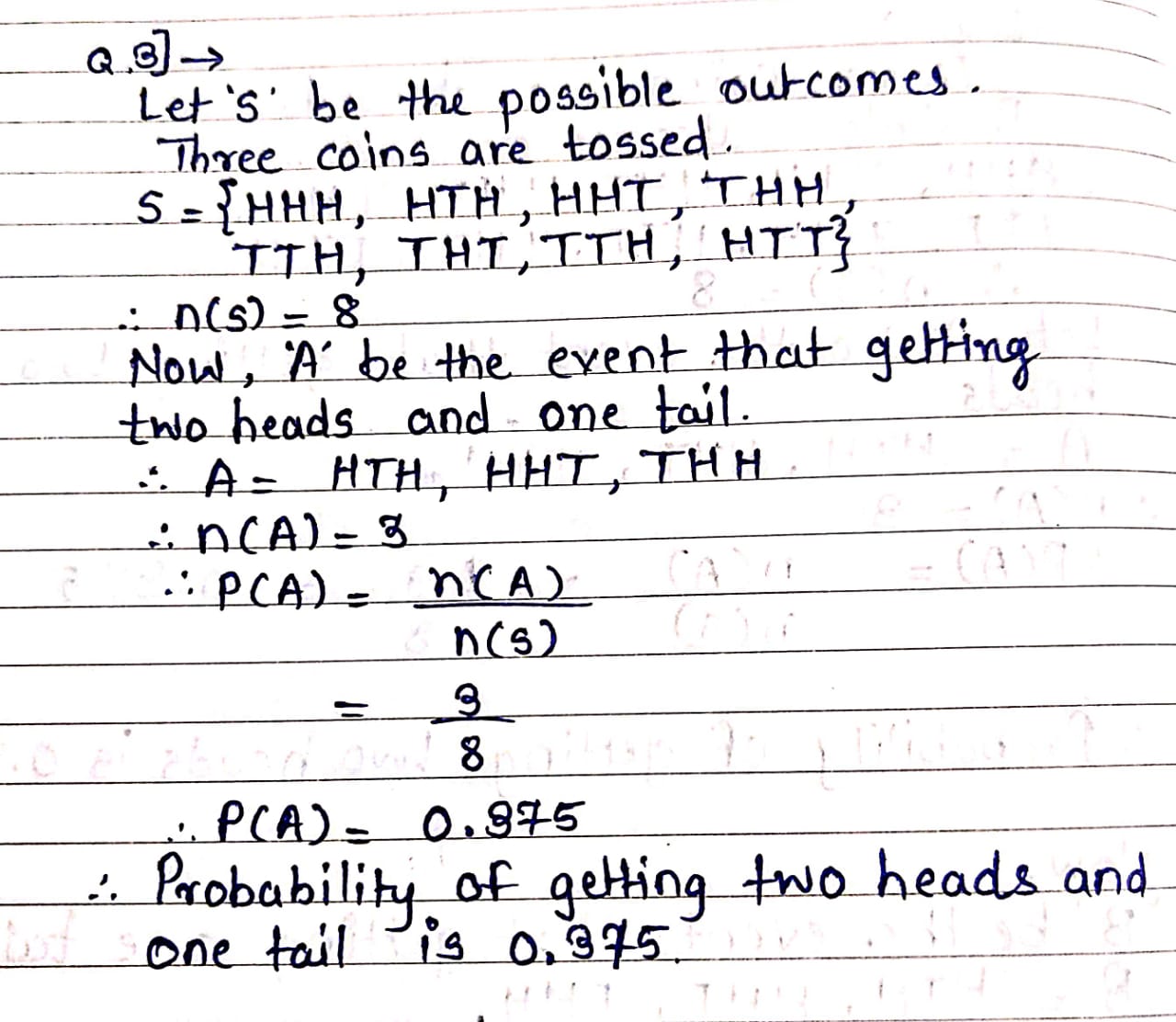
**Q2) Identify the Data types, which were among the following**

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

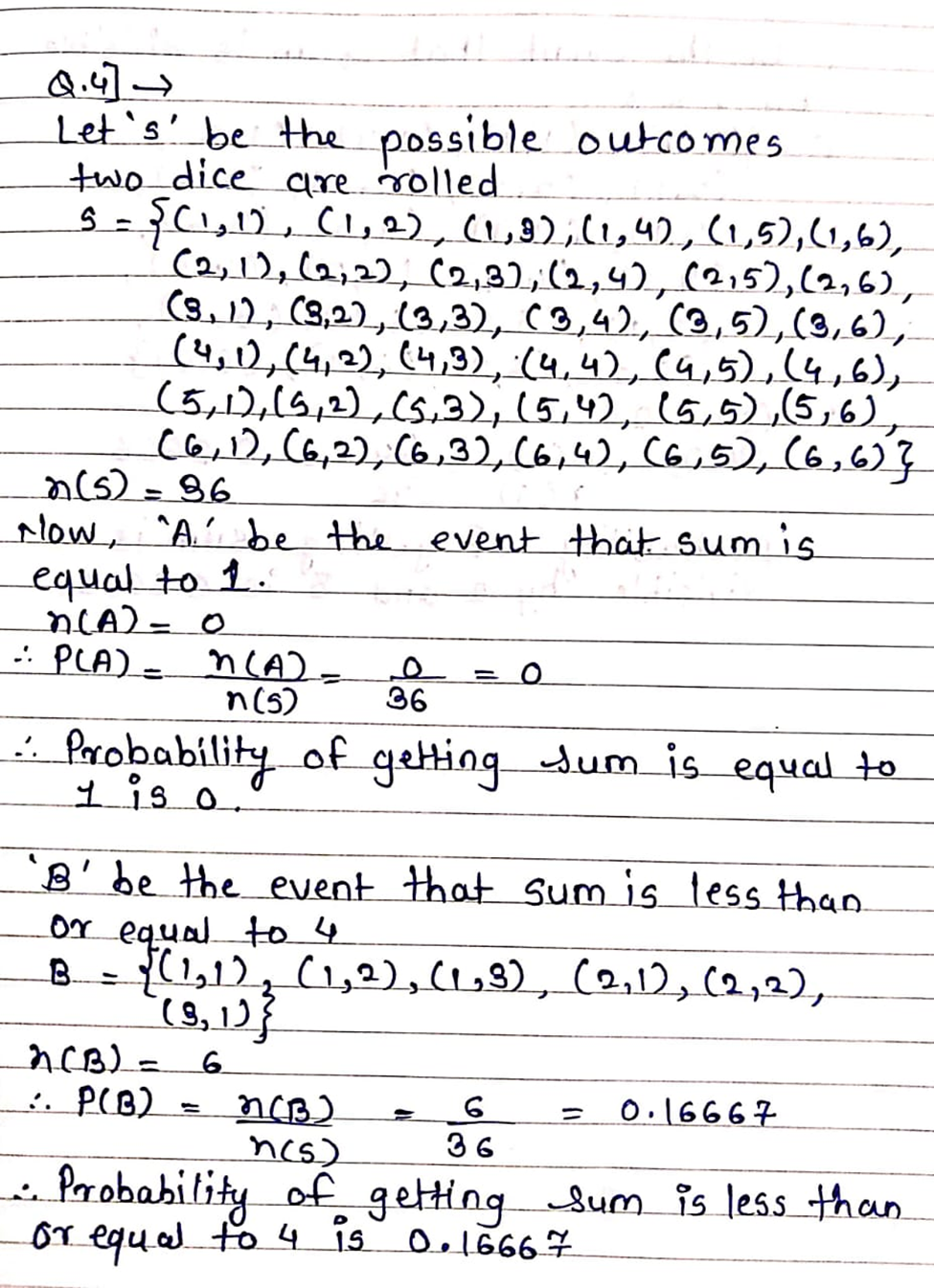
**ANSWER:**

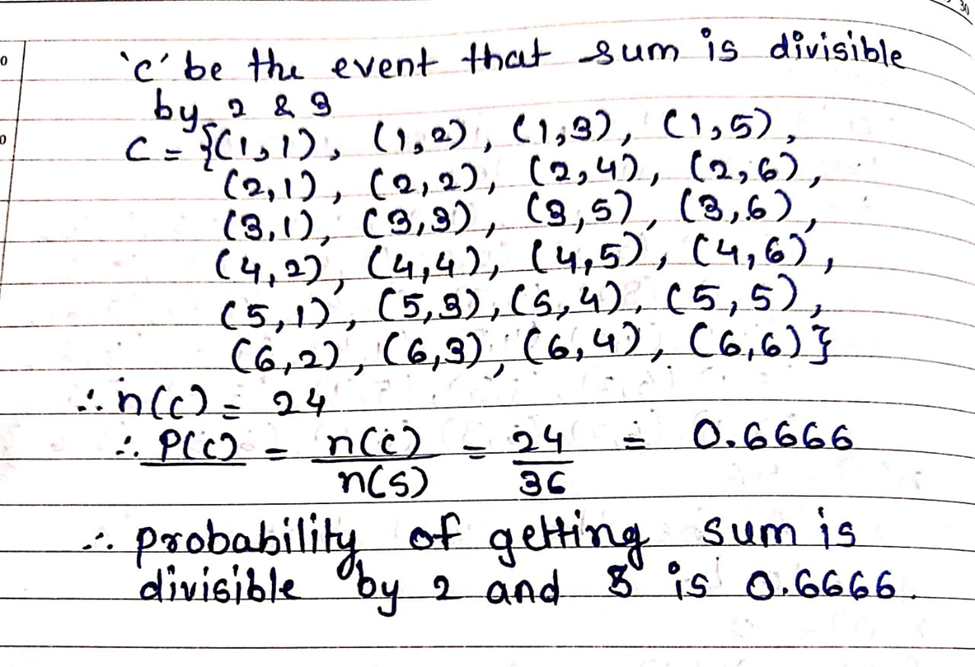
****

**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**

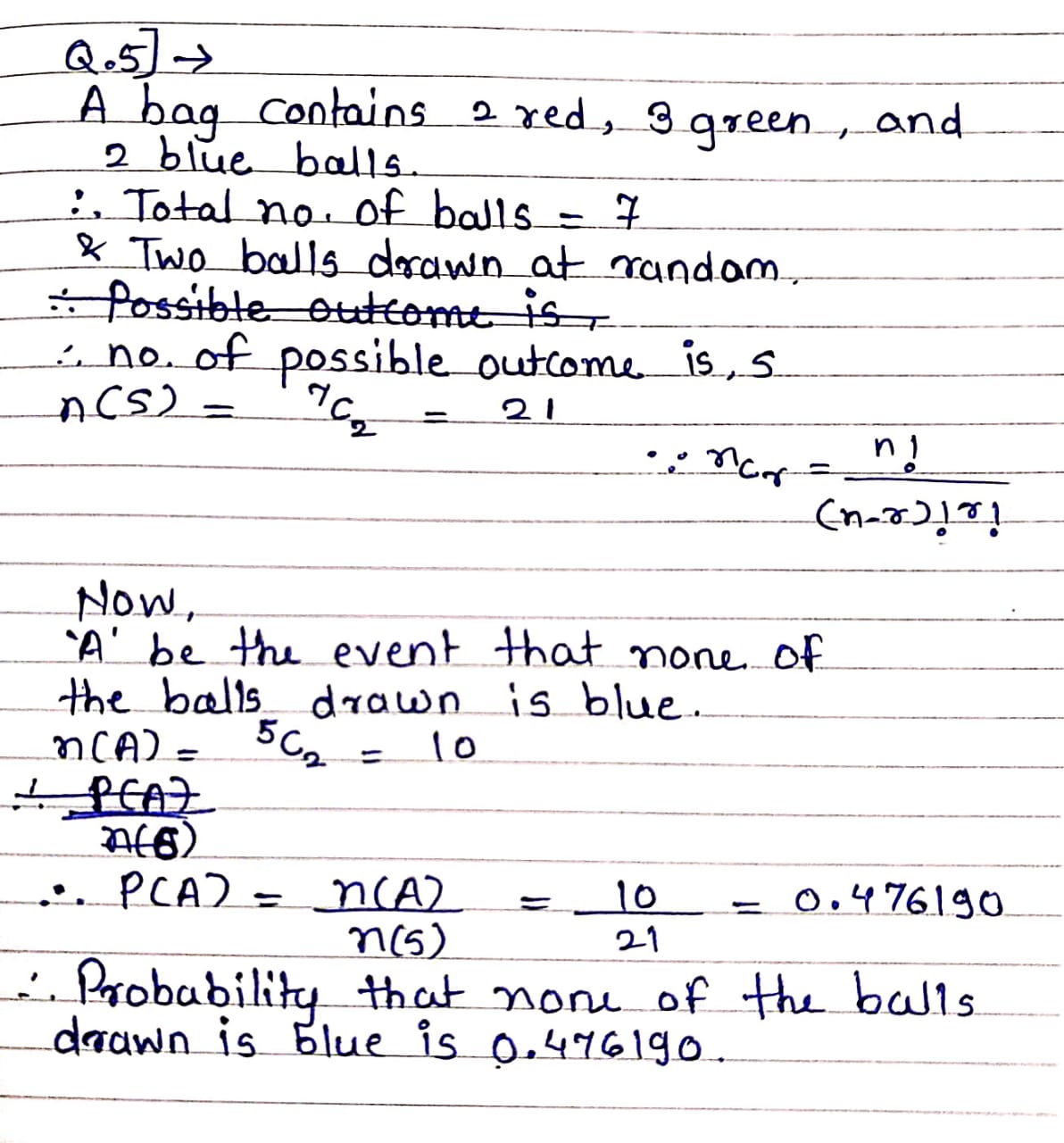
**ANSWER:**





**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?**

**ANSWER:**

****

**Q6) Calculate the Expected number of candies for a randomly selected child**

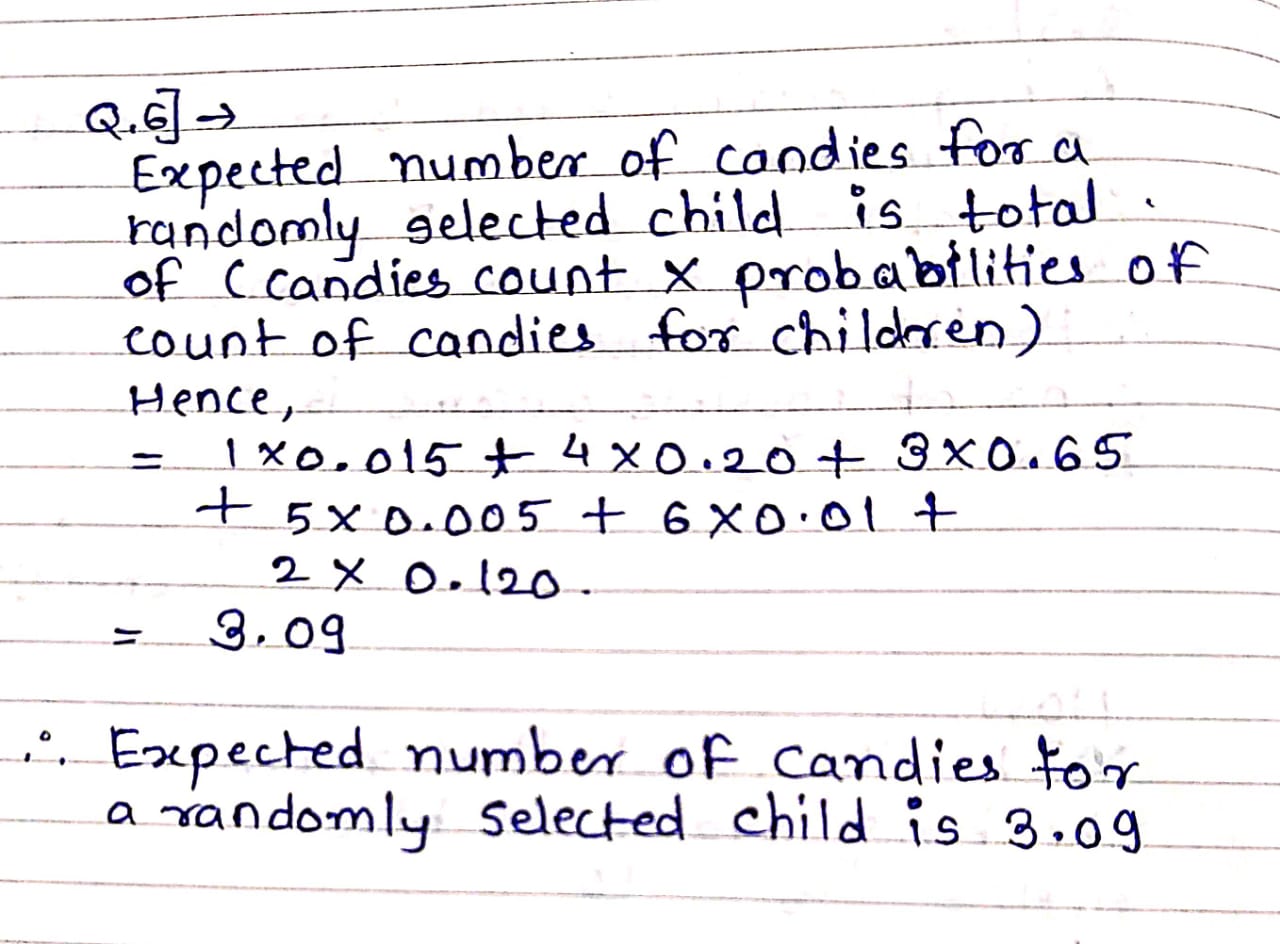
**Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)**

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

**ANSWER:**

****

**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**

**Answer:**

**Using R-Studio**

# Finding mean of Points, Score & Weigh columns of file name Q7

> mean(Q7$Points)

[1] 3.596563

> mean(Q7$Score)

[1] 3.21725

> mean(Q7$Weigh)

[1] 17.84875

**Comment:**

1. Average value of Points column is 3.596563
2. Average value of Score column is 3.21725
3. Average value of Weigh column is 17.84875

> # Finding median of Points, Score & Weigh columns of file name Q7

> median(Q7$Points)

[1] 3.695

> median(Q7$Score)

[1] 3.325

> median(Q7$Weigh)

[1] 17.71

**Comment:**

After arranging the data set in an ascending or descending order, we get median value which is an average of middle of two values of the data set.

1. For Points column, median value is 3.695
2. For Score column, median value is 3.325
3. For Weigh column, median value is 17.71

> # Finding mode of Points, Score & Weigh columns of file name Q7

> # For finding mode, installing package as "modeest"

> install.packages("modeest")

> library(modeest)

> mfv(Q7$Points)

[1] 3.07 3.92

> mfv(Q7$Score)

[1] 3.44

> mfv(Q7$Weigh)

[1] 17.02 18.90

**Comment:**

Mode is that value which has maximum frequency. (most frequently occurring value)

1)For Points column, mode is 3.07,3.92 (i.e bimodal data)

2) For Score, mode is 3.44 (i.e Unimodal data)

3) For Weigh, mode is 17.02, 18.90 (i.e Bimodal data)

># Finding varience of Points, Score & Weigh columns of file name Q7

> var(Q7$Points)

[1] 0.2858814

> var(Q7$Score)

[1] 0.957379

> var(Q7$Weigh)

[1] 3.193166

**Comment:**

Varience for Points, Score & Weigh columns is 0.2858814, 0.957379 & 3.193166 respectively.

From above these 3 colums, we have Weigh column which contains more variety of information.

> # Finding Standard deviation of Points, Score & Weigh columns of file name Q7

> sd(Q7$Points)

[1] 0.5346787

> sd(Q7$Score)

[1] 0.9784574

> sd(Q7$Weigh)

[1] 1.786943

**Comment:**

Standard deviation for Points, Score & Weigh columns is 0.5346787, 0.9784574 & 1.786943 resp.

> # Finding range of Points, Score & Weigh columns of file name Q7

> range(Q7$Points)

[1] 2.76 4.93

> range(Q7$Score)

[1] 1.513 5.424

> range(Q7$Weigh)

[1] 14.5 22.9

**Comment:**

1) In point column, maximum value is 4.93 & minimum value is 2.76.

2) In Score column, maximum value is 5.424 & minimum value is 1.513.

3) In Weigh column, maximum value is 22.9 & minimum value is 14.5.

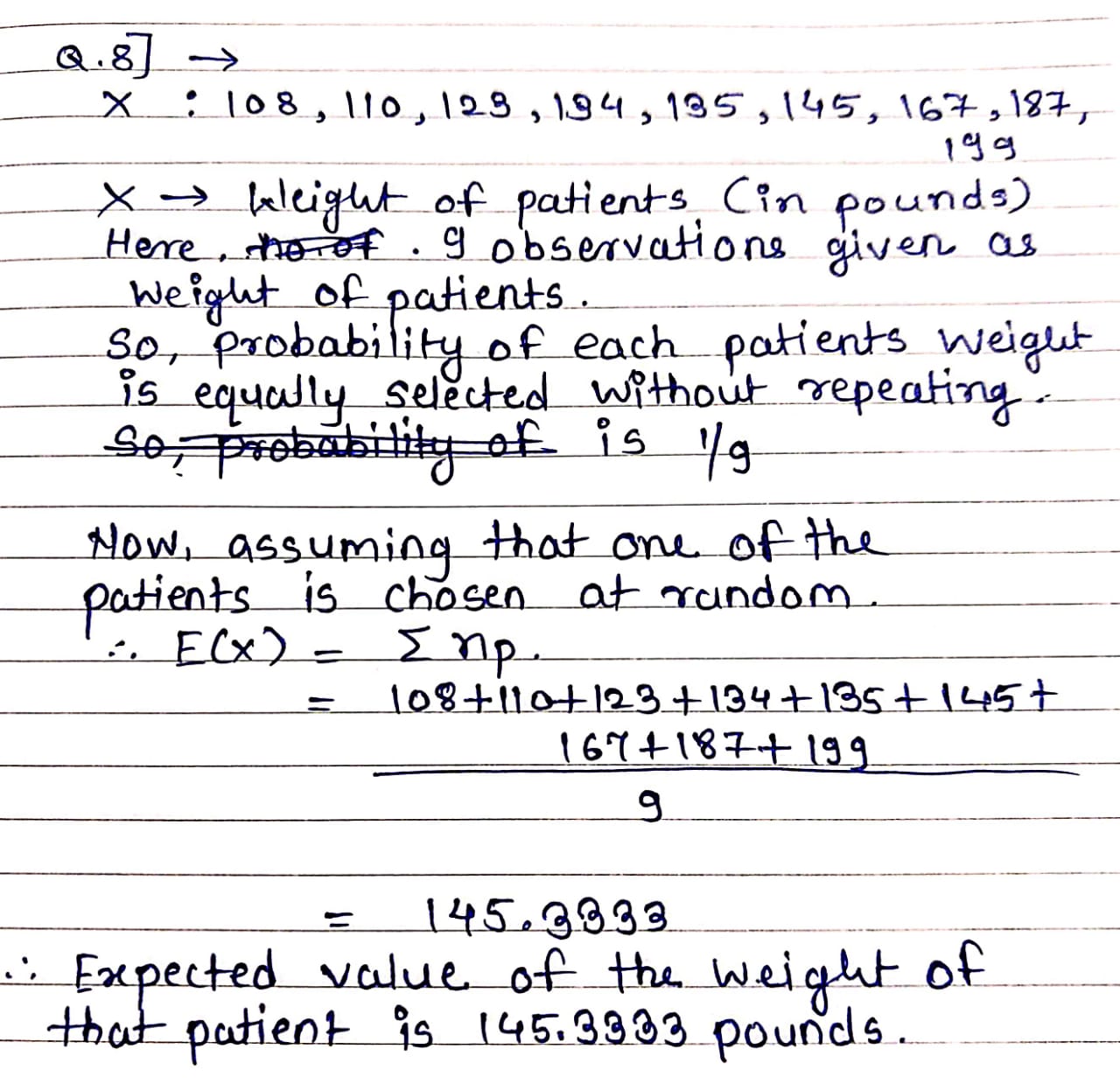
**Q8) Calculate Expected Value for the problem below**

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?**

**ANSWER:**

****

**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**

**Answer:**

# Importing data set

> Q9\_a <- read.csv("~/courses/data science/assignment/assignment1/Q9\_a.csv")

> View(Q9\_a)

> Q9\_b <- read.csv("~/courses/data science/assignment/assignment1/Q9\_b.csv")

> View(Q9\_b)

#Calculating skewness for car speed & distance of the data (file name:Q9\_a)

#For skewness installing package as "moments"

>install.packages("moments")

>library(moments)

>attach(Q9\_a)

> skewness(speed)

[1] -0.1139548

> skewness(dist)

[1] 0.7824835

># Calculating skewness for SP and Weight(WT) of the data (file name: Q9\_b)

>attach(Q9\_b)

> skewness(SP)

[1] 1.581454

> skewness(WT)

[1] -0.6033099

**Comment:**

Skewness for Car speed and WT is -0.1139548 & -0.6033099 respectively.

From this we can interpret that respective data is left skewed. (i.e skewness is not equal to zero)

Also, Skewness for car distance and SP is 0.7824835 & 1.581454 respectively.

From this we can interpret that respective data is right skewed. (i.e skewness is not equal to zero)

#Finding kurtosis values for the both the datasets

# “moments” package is also for kurtosis

>library(moments)

> kurtosis(speed)

[1] 2.422853

**Comment:**

Data is having a flat peak as value is 2.42 which is less than 3.

> kurtosis(dist)

[1] 3.248019

**Comment:**

Data is having a sharp peak as value is 3.248 which is greater than 3.

> kurtosis(SP)

[1] 5.723521

**Comment:**

Data is having a sharp peak as value is 5.72 which is greater than 3.

> kurtosis((WT))

[1] 3.819466

**Comment:**

Data is having a sharp peak as value is 3.81 which is greater than 3.

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



**Answer:**

Here, suppose we draw a line in a centre of the histogran then from that we got to know most of the observations are concentrated on the left hand side & a very few observations are concentrated on the right hand side of the histogram. Therefore, it is a right skewed histogram because it inclind towards right hand side of the histogram. For right skewed histogram skewness is positive and it can’t be equal to zero.

Upper whisker

Lower whisker



Lower limit

1st quartile

median

3rd quartile

Upper limit



**Answer:**

This is the boxplot used to detecting outliers. It is converted the data into four quartiles. It gives the information about median, 1st quartile, 3rd quartile, lower limit & upper limit. Middle portion is called as interquartile range or box and it contains the 50% of the observations. Also, lower and upper whisker contains 25% of the observations. In these boxplot, there is outliers present in the above of the upper limit.

**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?**

**Answer:**

**Using R-studio:**

>sample\_mean=200

> sample\_sd=30

> n=2000

> # Here population standard deviation is unknown, t-score is used

> # Finding CI

> # Interval estimation = Point estimation +- Margin of error

> # Point estimation = sample mean = 200

> # Margin of error = t-score \* sample sd / sqrt(n)

> # t-score= qt(confidence level, degrees of freedom(df))

> # 1) For 94%

> t\_score=round(qt(0.97,1999),4)

> margin\_error=t\_score\*30/sqrt(n)

> lower\_interval= sample\_mean-margin\_error

> lower\_interval

[1] 198.7376

> upper\_interval= sample\_mean+margin\_error

> upper\_interval

[1] 201.2624

> # 2)For 98%

> t\_score1=round(qt(0.99,1999),4)

> margin\_error1=t\_score1\*30/sqrt(n)

> lower\_interval1= sample\_mean-margin\_error1

> lower\_interval1

[1] 198.4382

> upper\_interval1= sample\_mean+margin\_error1

> upper\_interval1

[1] 201.5618

> # 3)For 96%

> t\_score2=round(qt(0.98,1999),4)

> margin\_error2=t\_score2\*30/sqrt(n)

> lower\_interval2= sample\_mean-margin\_error2

> lower\_interval2

[1] 198.6214

> upper\_interval2= sample\_mean+margin\_error2

> upper\_interval2

[1] 201.3786

**(ALSO, DONE IN PYTHON )**

**12) 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?**

**Answer:**

**R-studio:**

> scores=c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)

> mean(scores)

[1] 41

> median(scores)

[1] 40.5

> var(scores)

[1] 25.52941

> sd(scores)

[1] 5.052664

Conclusion:

The average score obtained by a student in tests is 41.

After arranging the scores in an ascending / decending order, getting median score is 40.5 which is a average of middle of two scores obtained by a student in a tests.

Variance and standard deviation of score obtained by student in a test is 25.52941 & 5.052664

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

**Answer:**

When mean, median of data are equal then nature of skewness of that data is perfectly symmetric. i.e skewness is equal to zero.

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

**Answer:**When mean is greater than median then nature of skewness of that data is right skewed. i.e data is positively skewed.

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

**Answer:**

When median is greater than mean then nature of skewness of that data is left skewed. i.e data is negatively skewed.

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

**Answer:**

If kurtosis value for data is positive, which indicates that the data is having a sharp peak.

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

**Answer:**

If kurtosis value for data is negative, which indicates that the data is having a flat peak.

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



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

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

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

**Answer:**

The distribution of the data is asymmetric or not normal distribution because here proportions around the median is not equal.

The nature of skewness of the data is left skewed (i.e negatively skewed) because median is closer to the top quartile (Q3).  
The IQR (Inter Quartile Range) of the data is,

Q3-Q1=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.**

**Answer:**

From both the boxplot, median is same.

**Q20) 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)**

**Answer:**

**Using R-studio**

> Cars <- read.csv("~/courses/data science/assignment/assignment1/Cars.csv")

> View(Cars)

> # Finding mean and variance of MPG

> mean(Cars$MPG)

[1] 34.42208

> sd(Cars$MPG)

[1] 9.131445

> # a. P(MPG>38)=?

> 1-pnorm(38,34.42208,9.131445)

[1] 0.3475941

> # b. P(MPG<40)=?

> pnorm(40,34.42208,9.131445)

[1] 0.7293497

> # c. P (20<MPG<50)=?

> pnorm(50,34.42208,9.131445)-pnorm(20,34.42208,9.131445)

[1] 0.8988689

**Q21) 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**

**Answer:**

(Done in python)

**Conclusion:**

Data follows normal distribution,

1. when we plot a distribution plot of data and getting curve is bell shaped
2. when proportion around the median is equal in boxplot
3. when skewness is zero
4. when kurtosis is 3
5. when mean = median = mode

here, MPG from Cars dataset & Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data setdoesn’t follow any of the above properties. So, data is not normally distributed.

**Q22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval**

**Answer:**

># zscore=qnorm(confidence level)

> # for 90% CI

> qnorm(0.95)

[1] 1.644854

> # for 94%

> qnorm(0.97)

[1] 1.880794

> # for 60%

> qnorm(0.8)

[1] 0.8416212

(Done in python also)

**Q23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25**

**Answer:**

>n=25

> # tscore=qt(confidence level, df=n-1)

> # for 95% CI

> qt(0.975,24)

[1] 2.063899

> # for 96% CI

> qt(0.98,24)

[1] 2.171545

> # for 99% CI

> qt(0.995,24)

[1] 2.79694

(also, done in python)

**Q24) 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**

**Answer:**

>population\_mean=270

> n=18

> sample\_mean=260

> sample\_sd=90

> # Null hypothesis: avg life of bulb is more than 260 days

> # Alternative hypothesis: avg life of bulb is not more than 260 days

> # here sample size is less than 30 also population standard deviation is not known so finding prob. using t score

> t\_score= (260-270)/(90/sqrt(18))

> t\_score

[1] -0.4714045

> #p(x>=260)=?

> pt(t\_score,17)

[1] 0.3216725

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.3216725 .

If assuming significance value is 0.05

i.e p-value > significance value

therefore, we do not reject the null hypothesis.

Hence, CEO's claim were false i.e average life of bulb is more than 260 days.