**Assignment 1:**

**Provide all the code and results from R and report the answers in a document along with your interpretation.**

1. **Data Camp: [40 marks]**

Complete the first three chapters to the Introduction to R in Data Camp:

1. **Vectors in R: [25 marks]**

**Regression The following are a sample of observations on incoming solar radiation at a greenhouse.**

**12.3 11.4 7.1 9.2 10.3 10.8 9.5 11.8**

1. **Assign the object to a vector called *solar.radiation***

*> #vector creation*

*> solar.radiation=c(12.3, 11.4, 7.1, 9.2, 10.3, 10.8, 9.5, 11.8)*

1. **Find the mean, median and variance.**

*> #Mean of solar.radiation*

*> mean\_solar.radiation = mean (solar.radiation)*

*> mean\_solar.radiation*

*[1] 10.3*

*>*

*> #Median of solar.radiation*

*> median\_solar.radiation = median (solar.radiation)*

*> median\_solar.radiation*

*[1] 10.55*

*>*

*> #variance of solar.radiation*

*> var\_solar.radiation = var(solar.radiation)*

*> var\_solar.radiation*

*[1] 2.828571*

1. **+ Add 10 to each observation and assign the resulting vector to *sr10*. Find the mean, median and variance of *sr10*. Which statistics change and by how much?**

> # +10 = sr10

> sr10= solar.radiation + 10

> sr10

[1] 22.3 21.4 17.1 19.2 20.3 20.8 19.5 21.8

>

> #Mean of sr10

> mean\_sr10= mean (sr10)

> mean\_sr10

[1] 20.3

>

> #Median of sr10

> median\_sr10 = median (sr10)

> median\_sr10

[1] 20.55

>

> #variance of sr10

> var\_sr10=var(sr10)

> var\_sr10

[1] 2.828571

> #different of Solar.radiation to sr10

> mean\_sr10 - mean\_solar.radiation

[1] 10

> median\_sr10 - median\_solar.radiation

[1] 10

(> var\_sr10 - var\_solar.radiation

[1] -2.220446e-15 ????)

> 2.828571 - 2.828571

[1] 0

1. **Multiple each observation by -2 and assign the resulting vector to *sr2*. Find the mean, median and variance of *sr2*. Which statistics change and by how much?**

> #solar.radiation x -2 (sr2)

> sr2= solar.radiation \* -2

> sr2

[1] -24.6 -22.8 -14.2 -18.4 -20.6 -21.6 -19.0 -23.6

>

> #Mean of sr2

> mean\_sr2 = mean(sr2)

> mean\_sr2

[1] -20.6

>

> #Median of sr2

> median\_sr2 = median(sr2)

> median\_sr2

[1] -21.1

>

> #Variance of sr2

> var\_sr2= var(sr2)

> var\_sr2

[1] 11.31429

>

> #different of Solar.radiation to sr2

> mean\_solar.radiation - mean\_sr2

[1] 30.9

> median\_solar.radiation - median\_sr2

[1] 31.65

> var\_solar.radiation - var\_sr2

[1] -8.485714

1. **Plot a histogram of each of *solar.radiation*, *sr10* and *sr2***.

*hist\_sola = hist(solar.radiation, main="Solar Radiation", col= c("blue"))*

Chart, histogram

Description automatically generated

*hist\_sr10 = hist(sr10, main="Solar Radiation + 10", col= c("grey"))*

Chart, histogram

Description automatically generated

*hist\_sr2 = hist(sr2, main="Solar Radiation x -2", col= c("yellow"))*

Chart, histogram

Description automatically generated

1. **Descriptive Statistics - Old faithful data: [35 marks]**

The Melanoma data frame has data on 205 patients in Denmark with malignant melanoma.The data is stored in R in the package MASS, in a dataset called ‘Melanoma’.

1. Load the package "MASS ". (hint: use library() commands to install package and then load it. Use install.packages() if you have never installed the package). Then load in the data "Melanoma" in R (hint: use the data() command and help() to understand the data).
2. Describe the data, referring to the number of variables and sample size?

The Melanoma data are data collected from 205 melanoma patients which includes:

* Age in year
* Sex: male or female
* Time: how long they have survived for in days
* Status: If they are dead or alive and if dead did, they did from the melanoma
* Year: when the operation was done
* Thickness: the thickness of the tumour in mm
* Ulcer: where or not if they had ulcer.

1. What type of variables are present in the data? Refer to each variable.

* time: Numerical (discrete)
* status: Categorical (ordinal)
* sex: Categorical
* age: Numerical (continuous)
* year: Numerical (continuous)
* thickness: Numerical (continuous)
* ulcer: Categorical

**D) Make suitable univariate plots for each of the variables in the data set and interpret the results.**

**Length of survival (Time)**

*#plot of time*

*boxplot(Melanoma$time)*

*print(boxplot(Melanoma$time))*

*boxplot(Melanoma$time,ylab="Lenght of Survival (Days)", main= "Boxplot Of Length Of Survival")*

Chart, box and whisker chart

Description automatically generated

Most of the patient falls above the survival mean of 2005 days, with the lowest length of survival being 10 days and the longest being 5565 represented by the outlier dot.

**Patients Status**

*#plot of Status*

*status.percent=round(100\*status.freq/sum(status.freq),0)*

*status.pie.label=paste(status.percent,"%",sep="")*

*pie(status.freq, main="Patients Status", labels=c(status.pie.label), col=rainbow(length(status.freq)))*

*legend("bottomright", c("Died From Melanoma", "Alive", "Died From Other Causes"),fill=rainbow(length(status.freq)))*

Chart, pie chart

Description automatically generated

Quick visual representation of the patient’s status distribution.

**Patient’s sex**

*#plot of sex*

*sex.percent= round(100\*sex.freq/sum(sex.freq),0)*

*sex.pie.label=paste(sex.percent,"%",sep="" )*

*pie(sex.freq, main="Pie Chart of Sex", labels=c(sex.pie.label), col=rainbow(length(sex.freq)))*

*legend("bottomright", c("Female", "Male"), fill=rainbow(length(sex.freq)))*

Chart, pie chart

Description automatically generated

Quick visual representation of the patient’s distribution.

**Patient Age in Years**

*#plot of age*

*boxplot(Melanoma$age)*

*print(boxplot(Melanoma$age))*

*boxplot(Melanoma$age,ylab="Age in years", main= "Boxplot of Age Distrubution")*

Chart, box and whisker chart

Description automatically generated

The boxplot shows that the most sample group were in around the median age of 54.

**Years Of Operations**

*#plot of year*

*boxplot(Melanoma$year)*

*print(boxplot(Melanoma$year))*

*boxplot(Melanoma$year,ylab="Year of Operation", main= "Boxplot of Operation Year")*

Chart, box and whisker chart

Description automatically generated

The evenly distributed plot shows that their sample year of operation have a normal behaviour to them.

**Thickness of Patient Tumour**

*#plot of thickness*

*boxplot(Melanoma$thickness)*

*print(boxplot(Melanoma$thickness))*

*boxplot(Melanoma$thickness,ylab="Tumour Thickness(mm)", main= "Boxplot of Tumour Thickness")*

Diagram

Description automatically generated with medium confidence

The tumour thickness blot plot shows there’s not a particular pattern in its tumour thickness in relation to the sample size.

**Patients With Ulcer**

*#plot of Ulcer*

*ulcer.percent=round(100\*ulcer.freq/sum(ulcer.freq),0)*

*ulcer.percent*

*ulcer.pie.label=paste(ulcer.percent,"%",sep="")*

*pie(ulcer.freq, main="Patients With Ulcer", labels=c(ulcer.pie.label), col=rainbow(length(ulcer.freq)))*

*legend("bottomright", c("Absence", "Present"),fill=rainbow(length(ulcer.freq)))*

Chart, pie chart

Description automatically generated

Quick visual representation of the % of patient’s that had an ulcer with the melanoma.

**E. For all numerical variables, report appropriate descriptive statistics for each of to describe the centrality and spread of the data. Explain the choice of statistic used.**

**1) (Melanoma$time centrality) Median was used to consider the outliers that would affect the centrality result if mean was use.**

*> median(Melanoma$time)*

**2005**

**2) (Melanoma$age centrality) Median was used to consider the outliers that would affect the centrality result if mean was use.**

*> median(Melanoma$age)*

**54**

**3) (Melanoma$year centrality) Mean was used to consider the normal distribution of the date. As the data type is in whole number the mean can be rounded up to better represent the sample. Which is also the same s the median showing how centralise the data was.**

*> mean(Melanoma$year)*

**1970**

*> median(Melanoma$year)*

**1970**

**4) (Melanoma$thickness centrality) Mean was used to consider the distribution of the date and to include the outliers which would have been ignored by the median affecting the result dramatically.**

**> #Stat of Melanoma$thickness**

*> mean(Melanoma$thickness)*

**2.919854**