STAT 46700 /CS 59000 **Topics in Data Science** **Spring 2025**

Lab 2

[Vaishak Balachandra]

Q.N. 1) Journal of Statistics Education, Volume 4, Number 2 (July 1996) include an article What's Normal? -- Temperature, Gender, and Heart Rate by A. Shoemaker. The dataset used in the article are provided in http://www.amstat.org/publications/jse/datasets/normtemp.dat.txt. The description of the data can be accessed in the link below.

http://jse.amstat.org/datasets/normtemp.dat.txt

a) How many variables are included in the study?

b) Print first five observations of the data.

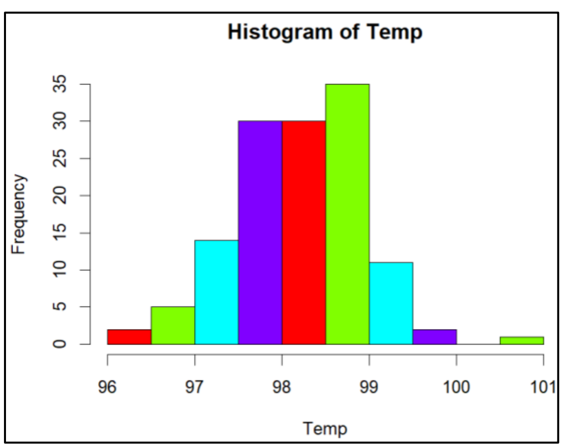
c) Is the distribution of body temperatures normal?

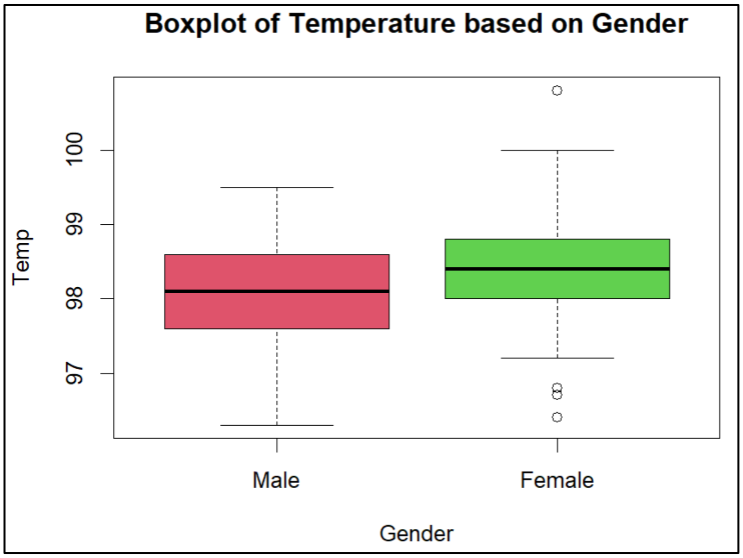
d) Is the true population mean really 98.6 degrees F?

e) Is there a significant difference on average temperature of males and females?

```
> ################################################### Q1
>
> # a
> q1 <- read.table("https://jse.amstat.org/datasets/normtemp.dat.txt")
> head(q1)
    V1 V2 V3
1 96.3  1 70
2 96.7  1 71
3 96.9  1 74
4 97.0  1 80
5 97.1  1 73
6 97.1  1 75
> names(q1) = c("Temp","Gender","Hrate")
> colnames(q1)
[1] "Temp"   "Gender" "Hrate"
> attach(q1)
> length(q1)
[1] 3
> # dim(q1)
> cat("There are 3 variables in the given dataset.")
There are 3 variables in the given dataset.
>
> # b
> head(q1,5)
  Temp Gender Hrate
1 96.3      1    70
2 96.7      1    71
3 96.9      1    74
4 97.0      1    80
5 97.1      1    73
>
> # c
> hist(Temp, col = rainbow(4))
> cat("Can't say using the histogram.")
Can't say using the histogram.
> shapiro.test(Temp)

	Shapiro-Wilk normality test

data:  Temp
W = 0.98658, p-value = 0.2332
```

```
> cat("Here, the pvalue = 0.2332, which is > 0.05. Thus, we fail to reject the null hypothesis that
conclude that the data comes from normally distributed as we don't have enough evidence to prove th
e data is not normally distributed.")
Here, the pvalue = 0.2332, which is > 0.05. Thus, we fail to reject the null hypothesis that conclu
de that the data comes from normally distributed as we don't have enough evidence to prove the data
is not normally distributed.
>
> # d
> t.test(Temp, mu = 98.6) # not mean, it has to be mu

	One Sample t-test

data:  Temp
t = -5.4548, df = 129, p-value = 2.411e-07
alternative hypothesis: true mean is not equal to 98.6
95 percent confidence interval:
 98.12200 98.37646
sample estimates:
mean of x
 98.24923

> cat("Here,the pvalue < 0.05. Thus, we reject the null hypothesis, and conclude that true populati
on mean is not really 98.6 degrees F")
Here,the pvalue < 0.05. Thus, we reject the null hypothesis, and conclude that true population mean
is not really 98.6 degrees F
>
> # e
> table(q1$Gender)

 1  2
65 65
> boxplot(Temp~Gender, col = c(2,3), main = "Boxplot of Temperature based on Gender", names = c("Ma
le","Female"))
> t.test(Temp~Gender)

	Welch Two Sample t-test

data:  Temp by Gender
t = -2.2854, df = 127.51, p-value = 0.02394
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean in group 1 mean in group 2
       98.10462        98.39385

> cat("Here, the pvalue = 0.02394, which is less than 0.05. Thus, we reject the null hypothesis and
conclude that average temperature of males and females are not equal.")
Here, the pvalue = 0.02394, which is less than 0.05. Thus, we reject the null hypothesis and conclu
de that average temperature of males and females are not equal.
```





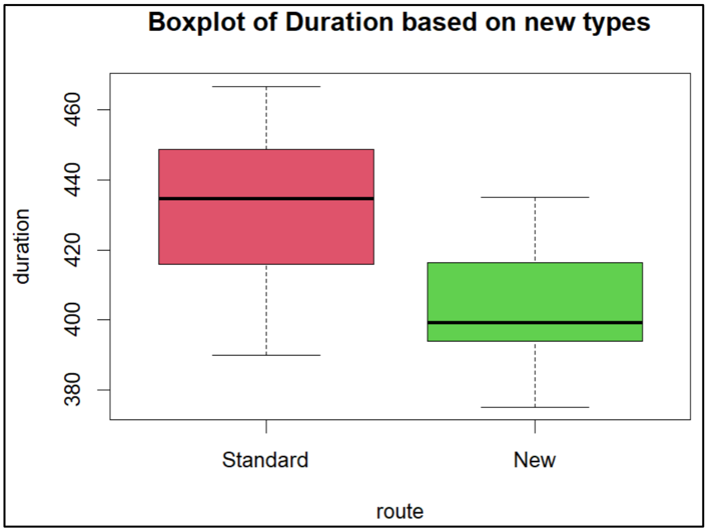
Q.N. 2) A company is investigating how long it takes its drivers to deliver goods from its factory to a port for export. Records from two different routes are provided in the link below.
http://media.pearsoncmg.com/cmg/pmmg_mml_shared/mathstatsresources/Akritas/DriveDurat.txt

Note that the routes has been coded as: 1- standard route, 2- new route

Is this sufficient evidence for the company to conclude, at $\alpha = 0.05$, that the new route is faster than the standard one?

```
> ################################################## Q2
>
> q2 <- read.table("https://media.pearsoncmg.com/cmg/pmmg_mml_shared/mathstatsresources/Akritas/DriveDurat.txt",header = T)
> head(q2)
  duration route
1    407.5     1
2    466.7     1
3    435.8     1
4    399.6     1
5    447.3     1
6    466.4     1
> attach(q2)
> dim(q2)
[1] 82  2
> names(q2)
[1] "duration" "route"
> boxplot(duration~route, col = c(2,3), names = c("Standard","New"), main = "Boxplot of Duration based on new types")
> table(q2$route)

 1  2
48 34
> t.test(duration~route, alt = "greater")

	Welch Two Sample t-test

data:  duration by route
t = 7.3387, df = 79.455, p-value = 8.13e-11
alternative hypothesis: true difference in means between group 1 and group 2 is greater than 0
95 percent confidence interval:
 22.57971      Inf
sample estimates:
```

```
mean in group 1 mean in group 2
       432.7021        403.5000
> # Lesser the duration, the faster it is. That's why, since we are dealing with duration though 1(
standard) comes before 2(New), we put 'greater', instead of 'lesser'
> cat("Here, the pvalue <0.05. Thus, we can conclude that New route is faster than standard route")
Here, the pvalue <0.05. Thus, we can conclude that New route is faster than standard route
```



Q.N. 3) Body measurements for a sample of 198 children are provided in the dataset accompanying this lab.

a) Import the data in R

b) Print the name of the variables

c) Draw a boxplot to display the height distribution based on gender (Note Sex: 0 – male, 1-female)

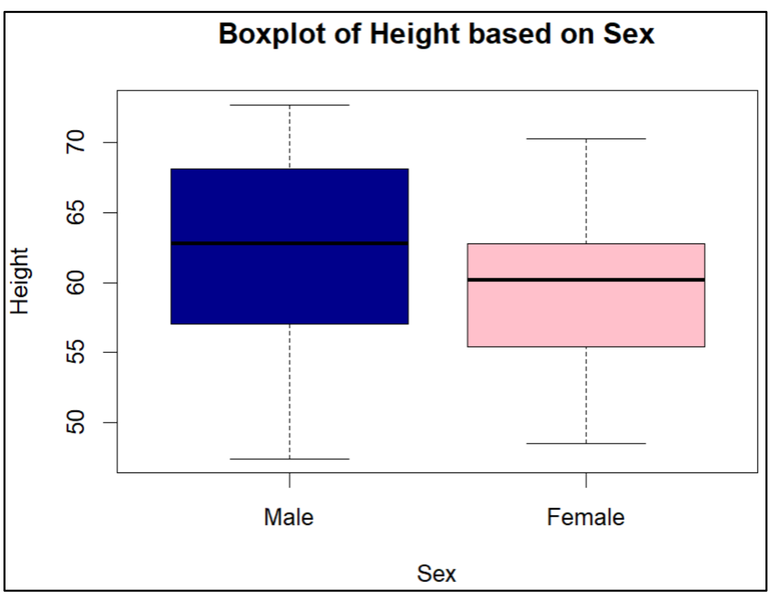
d) Draw a boxplot to display the weight distribution based on the gender

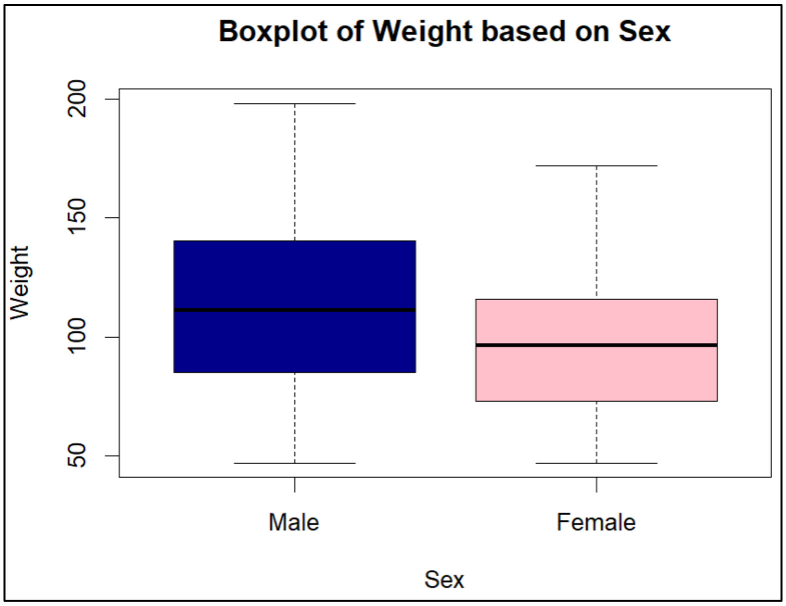
e) Is there a significant difference in average weight between male and female?

f) Are male taller than females ?

```
> ################################################## Q3
>
> # a
> q3 <- read.csv("C:/Users/PNW_checkout/Downloads/sem2/0. Coursework/Data science/Lab/Lab 2/Lab2 Da
ta.csv")
> head(q3)
  X Height Weight Age Sex Race
1 1   67.8    166 210   0    1
2 2   63.0     93 144   1    0
3 3   50.1     54 119   0    0
4 4   55.7     69 130   1    0
5 5   63.2    115 157   0    0
6 6   48.8     52 102   0    0
>
> # b
> names(q3)
[1] "X"      "Height" "Weight" "Age"    "Sex"    "Race"
>
> # c
> attach(q3)
```

```
> boxplot(Height~Sex, names = c("Male","Female"), col = c("darkblue","pink"), main = "Boxplot of He
ight based on Sex")
>
> # d
> boxplot(Weight~Sex, names = c("Male","Female"), col = c("darkblue","pink"), main = "Boxplot of We
ight based on Sex")
>
> # e
> t.test(Weight~Sex)

	Welch Two Sample t-test

data:  Weight by Sex
t = 3.5197, df = 180.01, p-value = 0.0005471
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
  7.116825 25.277783
sample estimates:
mean in group 0 mean in group 1 
      112.35417        96.15686 

> cat("Here, the pvalue =0.0005471, which is less than 0.05. Thus, we reject the null hypothesis an
d conclude that there is a significant difference in the average weight between males and females")
Here, the pvalue =0.0005471, which is less than 0.05. Thus, we reject the null hypothesis and concl
ude that there is a significant difference in the average weight between males and females
>
>
> # f
> t.test(Height~Sex, alt = "greater")

	Welch Two Sample t-test

data:  Height by Sex
t = 3.7132, df = 179.95, p-value = 0.0001364
alternative hypothesis: true difference in means between group 0 and group 1 is greater than 0
95 percent confidence interval:
 1.729423      Inf
sample estimates:
mean in group 0 mean in group 1 
       62.29896        59.18137 

> cat("Here, the pvalue = 0.0001364, which is less than 0.05. Thus, we reject the null hypothesis a
nd conclude that the height of male is significantly higher than that of females.")
Here, the pvalue = 0.0001364, which is less than 0.05. Thus, we reject the null hypothesis and conc
lude that the height of male is significantly higher than that of females.
```





Q.N. 4) Data of the manuscript 'Analysis of data with censored initiating and terminating times: a missing-data approach' by Xin Tu are provided in the link below
http://lib.stat.cmu.edu/jcgs/tu

a. Import the data in R without saving in your computer and determine its dimension.
b. The last column TRT indicates which treatment group an individual belongs to. Determine how many individuals received treatment 2.

```
> ################################################### Q4
>
> # a
> q4 <- read.table("https://lib.stat.cmu.edu/jcgs/tu", skip = 4, header = T)
> head(q4)
  XL XR ZL ZR AGE MULT TRT
1 15 24  1 24   1   13   1
2 15 24  1 24   2    6   1
3 16 24  1 24   1   15   1
4 16 24  1 24   2   16   1
5 17 24  1 24   1    9   1
6 17 24  1 24   2    1   1
> dim(q4)
[1] 136   7
> cat("There are 136 rows and 7 columns.")
There are 136 rows and 7 columns.
>
> # b
> attach(q4)
> sum(q4$TRT == 2)
[1] 69
> # or subset method, or table method
> cat("Thus, there are 69 individuals received treatment 2.")
Thus, there are 69 individuals received treatment 2.
```