Name-Surname : **Email** :

No : **Signature** :

HOME WORK 1 (Return by 5.12.2017)

BLM3590 – Statistical Data Analysis

Q1(15)	Q2(15)	Q3(20)	Q4(20)	Q5(15)	Q6(15)					Total(100)

Q1: We have measured the height (in inches) and weight (in pounds) for five newborn babies (Table).
Manually calculate the mean and standard deviation of height and weight; show all the steps.

Table: Height (in inches) and weight (in pounds) for five newborn babies

Observation	Height	Weight
1	18	7.8
2	21	9.1
3	17	8.2
4	16	6.4
5	19	8.8

$$Mean\ of\ Height = \bar{x}_H = \frac{x_{H1} + x_{H2} + \cdots + x_{Hn}}{\#\ of\ observations} = \frac{18 + 21 + 17 + 16 + 19}{5} = 18.2$$

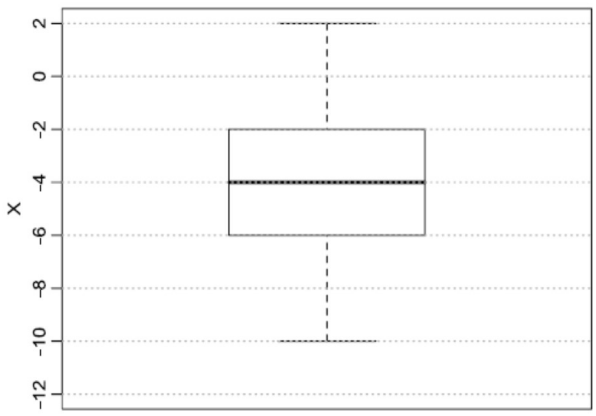
$$Mean\ of\ Weight = \bar{x}_W = \frac{x_{W1} + x_{W2} + \cdots + x_{Wn}}{\#\ of\ observations} = \frac{7.8 + 9.1 + 8.2 + 6.4 + 8.8}{5} = 8.06$$

$$Standard\ deviation\ of\ Height = \sqrt{\sum_{i=1}^{n} \frac{(x_{Hi} - \bar{x}_H)^2}{n-1}} = \sqrt{\frac{(18 - 18.2)^2 + \cdots + (19 - 18.2)^2}{5-1}} = 1.92$$

$$Standard\ deviation\ of\ Weight = \sqrt{\sum_{i=1}^{n} \frac{(x_{Wi} - \bar{x}_W)^2}{n-1}} = \sqrt{\frac{(7.8 - 8.06)^2 + \cdots + (8.8 - 8.06)^2}{5-1}} = 1.06$$

Q2: Based on the following boxplot, write down the five-number data summary, range and IQR of variable *X*.

Boxplot of variable *X*



Five number summary = (-10, -6, -4, -2, 2)
Range = 2 - (-10) = 12
IQR = -2 - (-6) = 4

Q3: Download the "BodyTemperature.txt" from the book website (http://extras.springer.com), and find the five-number data summary for all numerical variables. For numerical variables, provide the histograms and boxplots. Comment on the central and the form of the histograms. Are there any outliers in the data?

Remember to import new dataset "BodyTemperature.txt". To find five number summaries for all numerical variables at the same time, go to *Statistics* → *Summaries*→ *Numerical Summaries*, highlight all numerical variables "Age", "HeartRate", and "Temperature", then click "OK". Here are the results:
Numerical Summaries:

	Min.	Q_1	Q_2	Mean	Q_3	Max.
Age	21.00	33.75	37.00	37.62	42.00	50.00
HeartRate	61.00	69.00	73.00	73.66	78.00	87.00
Temperature	96.20	97.70	98.30	98.33	98.90	101.30

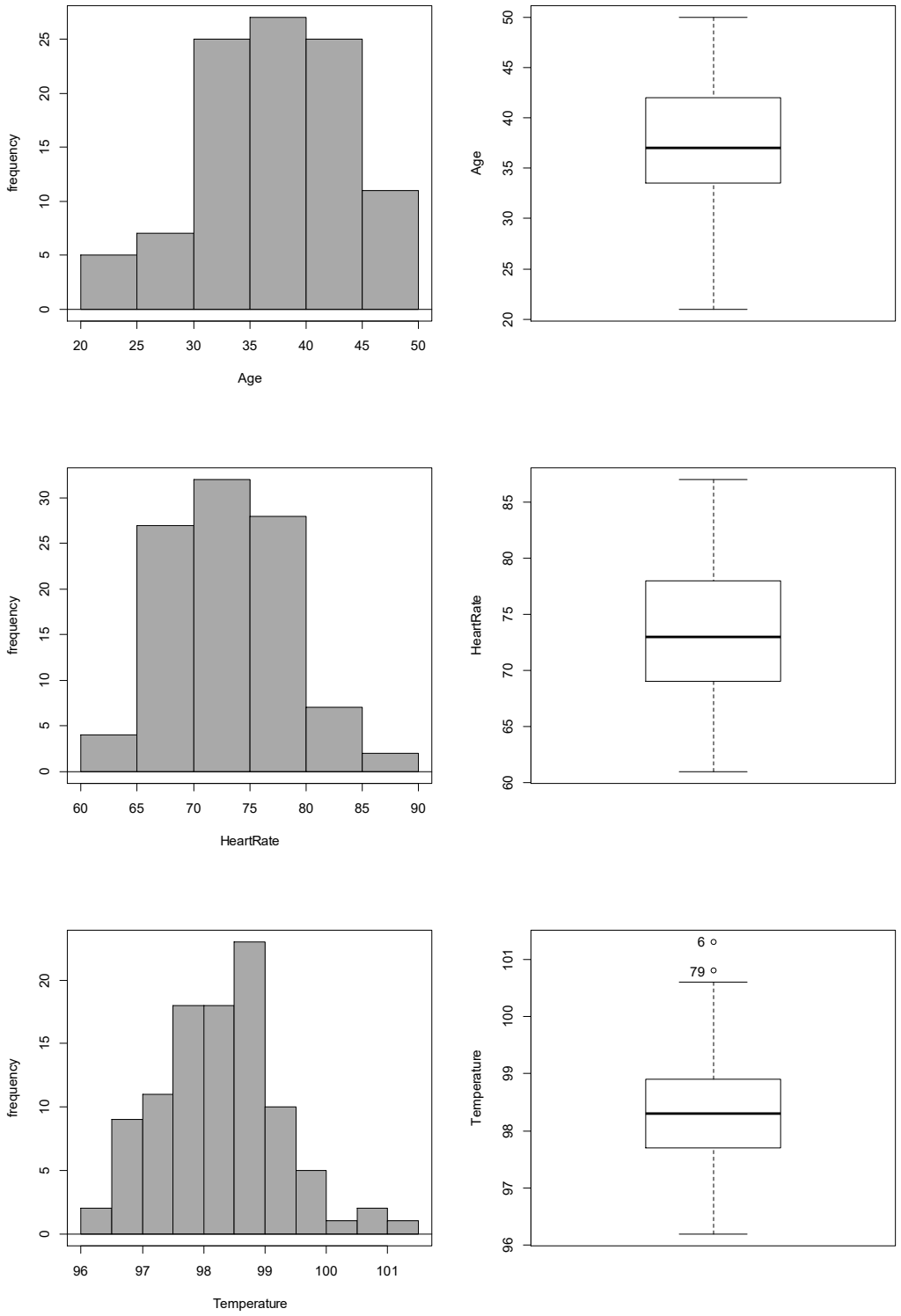


Figure: Histograms and box plots of variables *Age*, *HeartRate*, and *Temperature*

Study the figure for histograms and boxplots for "Age", "HeartRate", and "Temperature".

For "Age", the histogram is almost symmetric, there is no outlier, central tendency is around 35-40.

For "HeartRate", the histogram is almost symmetric, there is no outlier, central tendency is around 70-75.

For "Temperature", there seems to be a second mode after 100.
The sample might have included a group of individuals who had mild fever even though the target population was healthy individuals.
On the other hand, because there are only few (4) individuals with body temperature above 100, they might be simply outliers. The boxplot shows that two of them can be in fact considered as outlier (denoted with dots in boxplot).
The central tendency is around 98-99.
(You can use the sample mean and median to provide a more precise values for the central tendency.)

==========

Q4: Using the "BodyTemperature.txt" data set, create the scatterplot for body temperature by heart rate. Describe the pattern and comment on possible relationship between the two variables. Find the correlation coefficient between body temperature and heart rate. Finally, create boxplots of body temperature for men and women separately. Which one tends to be higher? Which one has higher dispersion?

Five-number data summary
Gender: 49 F, 46 M

	Min.	Q_1	Q_2	Mean	Q_3	Max.
Age	21.00	33.75	37.00	37.62	42.00	50.00
HeartRate	61.00	69.00	73.00	73.66	78.00	87.00
Temperature	96.20	97.70	98.30	98.33	98.90	101.30

After you upload "BodyTemperature" into R-Commander, to create the scatterplot, click *Graphs* →*scatterplot*, select "HeartRate" for *x*-variable and "Temperature" for *y*-variable. To make a scatterplot with just the least-squares line (i.e., trend line), you should unmark other options, such as "Smooth line", "Show spread", and "Marginal boxplots", then click OK.
The scatterplot between body temperature and heart rate is shown in the left panel of the following figure.

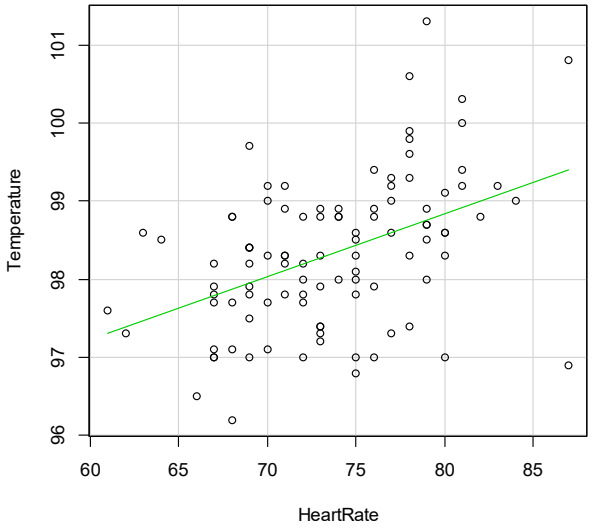
The plot suggests that the increase in heart rate tends to coincide with the increase in body temperature. The two variables seem to have a positive linear relationship.

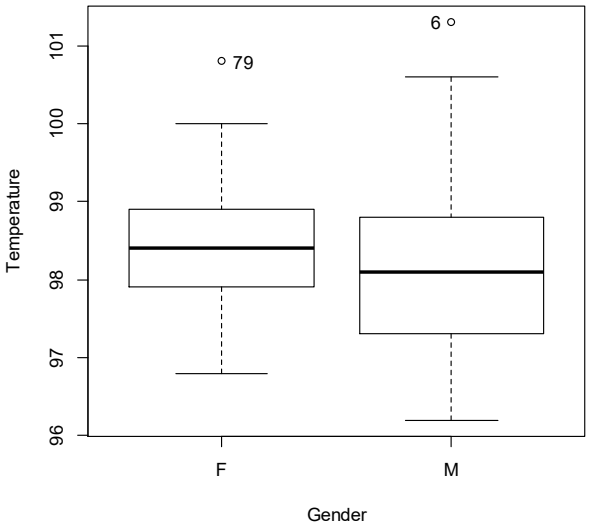
To find correlation coefficient between body temperature and heart rate, go to *Statistics* → *Summaries* → *Correlation matrix*..., select "Temperature" and "HeartRate", then click OK.

You should get correlation = 0.448

This correlation coefficient is in accordance to what we found from examining the scatterplot.

To create boxplots, point to *Graphs* → *boxplot*, highlight "Temperature", click on "Plot by groups" to select Gender, then click OK. This will create boxplots of temperature separately for men and women. Boxplots for temperature by gender is shown in the right panel of following figure.
Men's body temperature tends to be slightly lower. Further, body temperature for men seems to be more dispersed compared to that of women.





==========

Q5: We assume that the probability distribution of blood pressure, X, is $N(\mu,\sigma^2)$distribution. Suppose we know that $\sigma = 6$. To estimate μ, we randomly selected 9 people and measured their blood pressure. The sample mean is $\overline{x} = 110$.

a. Write down the sampling distribution of the sample mean $\overline{X}$ and find its standard deviation.
b. Find the 80% confidence interval estimation for μ.

---

a. $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$, $\quad \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{9}} = \frac{6}{3} = 2$
b. With the point estimate $\bar{x}$, the confidence interval for the population mean at c confidence level is

$$\left[\bar{x} - z_{crit} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{crit} \times \frac{\sigma}{\sqrt{n}}\right] = \left[110 - z_{crit} \times \frac{\sigma}{\sqrt{n}}, 110 + z_{crit} \times \frac{\sigma}{\sqrt{n}}\right]$$

The upper-tail probability of $z = (1 - 0.8) / 2 = 0.1$
Using R-Commander, $z_{crit} = 1.28$
80% confidence interval estimation for μ is

$$[110 - 1.28 \times 2, 110 + 1.28 \times 2] = [107.44, 112.56]$$

==========

Q6: For the question 5, suppose that we did not know σ and estimated it using the sample standard deviation $s = 6$.

a. Find the standard error for the sample mean as the estimator of the population mean.
b. Find the 80% confidence interval estimation for μ based on this sample.

---

a. $SE = \frac{s}{\sqrt{n}} = \frac{6}{\sqrt{9}} = \frac{6}{3} = 2$
b. The confidence interval for the population mean at c confidence level is

$$\left[\bar{x} - t_{crit} \times \frac{s}{\sqrt{n}}, \bar{x} + t_{crit} \times \frac{s}{\sqrt{n}}\right] = [110 - t_{crit} \times 2, 110 + t_{crit} \times 2]$$

We use the t-distribution with $9 - 1 = 8$ degrees of freedom .
t_{crit} : its upper tail probability is $(1 - 0.8) / 2 = 0.1$. (0.8 confidence)
Using R-Commander, $t_{crit} = 1.40$
The 80% confidence interval estimation for μ based on this sample:

$$[110 - 1.40 \times 2, 110 + 1.40 \times 2] = [107.2, 112.8]$$