AMCS 210 - Applied Statistics and Data Analysis
Hernando Catequista Ombao
Applied Mathematics and Computational Sciences Program
Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

MIDTERM I
Take-Home Part

Henrique Aparecido Laureano

Fall Semester 2017

Contents

A clinician is studying the impact of reduction in fasting blood glucose level on subjects with early signs of diabetes. In this study, $N = 200$ participants with *similar* clinical data are randomly assigned into four treatment groups with each treatment described below. There are 50 participants in each group.

- **The first treatment consists of a regular diet and a low intensity exercise program.**
- **The second treatment consists of a regular diet and a high intensity exercise program.**
- **The third treatment consists of a strict low sugar diet and a low intensity exercise program.**
- **The fourth treatment consists of a strict low sugar diet and a high intensity exercise program.**

After 6 months of treatment, the reduction in blood glucose level is measured for each participant. The key questions being investigated are: (a.) Is there a difference in the mean blood reduction level between those with high intensity vs low intensity exercise? (b.) Is there a difference between the two types of diet? (c.) Is there a difference among these four groups. Answer the following questions carefully and completely.

First you need to download the dataset "BloodSugar4" from the class dropbox folder BOOK-DATA. Choose your working directory to be the folder you stored your dataset. Then load the data into `R` as follows

```
dataY = matrix(scan("BloodSugar4"), byrow = T, ncol = 4);
```

```
# <code r> ============================================================================ #
path <- "~/Dropbox/CLASS-DROPBOX/BOOK-DATA/"
dataY <- matrix(scan(paste0(path, "BloodSugar4")), byrow = TRUE, ncol = 4)
# </code r> ============================================================================ #
```

Check that there should be 4 columns and that the k-th column corresponds to the k-th treatment.

```
# <code r> ============================================================================ #
ncol(dataY)
# </code r> ============================================================================ #
```

```
[1] 4
```

Moreover, you can combine columns 1 and 2 as consisting of all participants who were given a regular diet; columns 1 and 3 as all participants who were enrolled in a low-intensity exercise and so on.

```
# <code r> ============================================================================ #
colnames(dataY) <- c("RD-LI", "RD-HI", "LD-LI", "LD-HI")
# </code r> ============================================================================ #
```

Question 1

Compare the boxplots and histograms of reduction data for the two exercise groups. Comment on the boxplots (keeping in mind the main goals of this study).

Solution:

```
# <code r> ============================================================================ #
par(mfrow = c(2, 2))
par(mar = c(1, 4, 3, 2))

boxplot(c(dataY[ , c(1, 3)]), las = 1, pch = 16, ylim = c(5, 25)
        , main = "Low intensity", ylab = "Reduction in BGL")
boxplot(c(dataY[ , c(2, 4)]), las = 1, ylim = c(5, 25), main = "High intensity"
        , ylab = "Reduction in BGL")

par(mar = c(4, 4, .75, 2))

hist(c(dataY[ , c(1, 3)]), las = 1, xlim = c(0, 30), main = NA
     , xlab = "Reduction in BLG")
hist(c(dataY[ , c(2, 4)]), las = 1, xlim = c(0, 30), ylim = c(0, 30), main = NA
     , xlab = "Reduction in BLG")
# </code r> ============================================================================ #
```

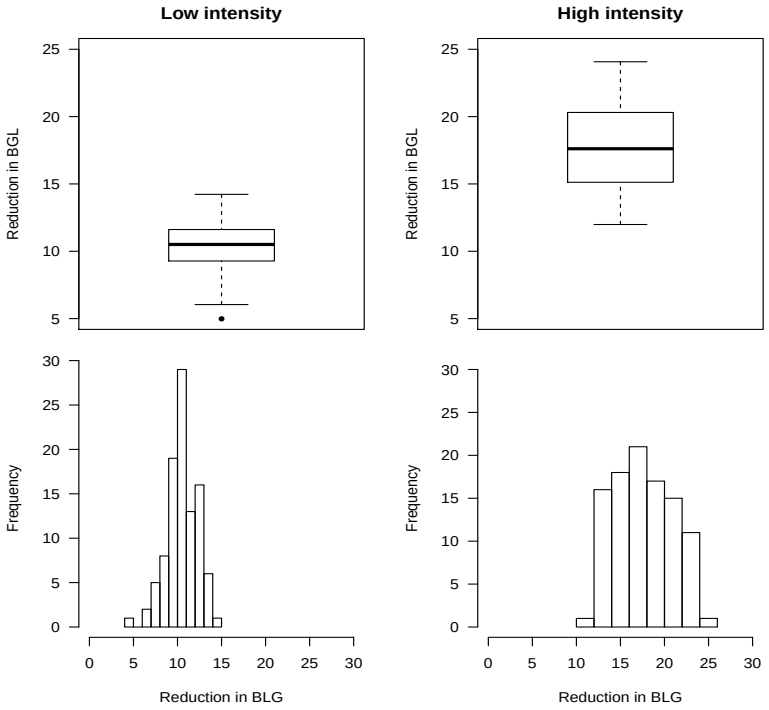

Figure 1: Boxplots and histograms of the reduction in Blood Glucose Level (BGL) for the two exercise groups. Low intensity group in the left, high intensity in the right.

We can see in the Figure 1 that the variability and the median are bigger in the high intensity group. The median in the high intensity (HI) group is around 50% bigger than in the low intensity (LI) group. The variability is around 1/3 bigger in the HI than in the LI group. Looking for this values we can imagine that perhaps exist a difference between the groups.

□

Question 2

Compare the boxplots and histograms of reduction data for the two diet groups. Again, comment on the boxplots (keeping in mind the main goals of this study).

Solution:

```
# <code r> ============================================================================ #
par(mfrow = c(2, 2))
par(mar = c(1, 4, 3, 2))

boxplot(c(dataY[ , 1:2]), las = 1, ylim = c(5, 25), main = "Regular diet"
        , ylab = "Reduction in BGL")
boxplot(c(dataY[ , 3:4]), las = 1, ylim = c(5, 25), main = "Strict low sugar diet"
        , ylab = "Reduction in BGL")

par(mar = c(4, 4, .75, 2))

hist(c(dataY[ , 1:2]), las = 1, xlim = c(0, 30), main = NA
     , xlab = "Reduction in BLG")
hist(c(dataY[ , 3:4]), las = 1, xlim = c(0, 30), ylim = c(0, 25), main = NA
     , xlab = "Reduction in BLG")
# </code r> ============================================================================ #
```

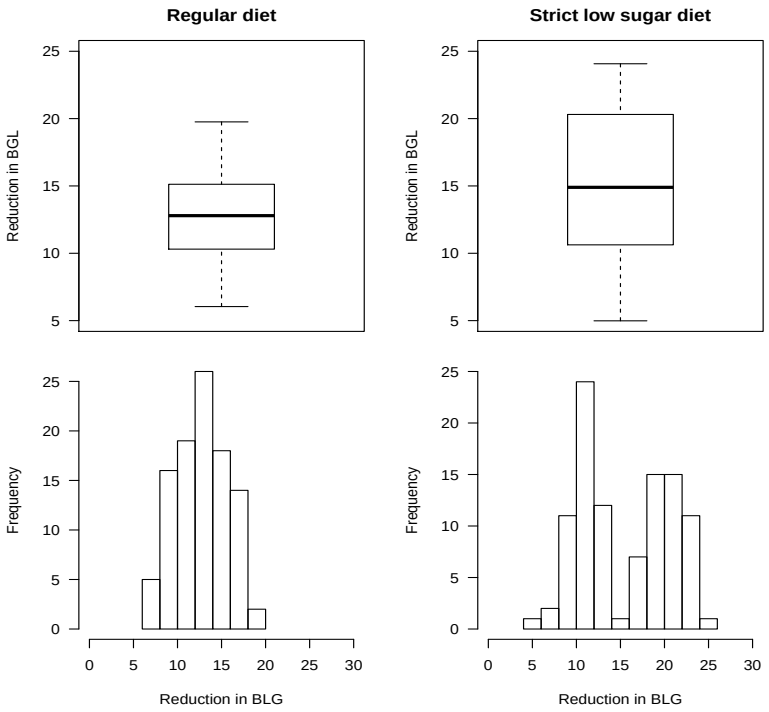

Figure 2: Boxplots and histograms of the reduction in Blood Glucose Level (BGL) for the two diet groups. Regular diet group in the left, strict low sugar diet in the right.

We can see in the Figure 2 that the variability and the median are bigger in the strict low sugar diet. The median in the strict low sugar diet (LD) is less than five units bigger than in the regular diet (RD) group. The variability is around 25% bigger in the LD than in the RD group. Looking for this values we can see that exist a difference, and that this difference is smaller than the difference observed between the two exercise groups.

□

Question 3

Compare the boxplots and histograms of reduction data of the four treatment groups. Again, comment on the boxplots (keeping in mind the main goals of this study).

Solution:

```
# <code r> ============================================================================ #
layout(matrix(c(rep(1, 4), 2, 3, 4, 5), 2, 4, byrow = TRUE))
par(mar = c(3, 4, .1, 2))
boxplot(dataY, las = 1, pch = 16, ylab = "Reduction in BLG", ylim = c(5, 25))
par(mar = c(4, 4, 2, 2))
hist(dataY[ , 1], las = 1, xlab = "Reduction in BLG", main = "RD-LI"
     , xlim = c(5, 25), ylim = c(0, 20))
hist(dataY[ , 2], las = 1, xlab = "Reduction in BLG", main = "RD-HI"
     , xlim = c(5, 25), ylim = c(0, 20))
hist(dataY[ , 3], las = 1, xlab = "Reduction in BLG", main = "LD-LI"
     , xlim = c(5, 25), ylim = c(0, 20))
hist(dataY[ , 4], las = 1, xlab = "Reduction in BLG", main = "LD-HI"
     , xlim = c(5, 25), ylim = c(0, 20))
# </code r> ============================================================================ #
```

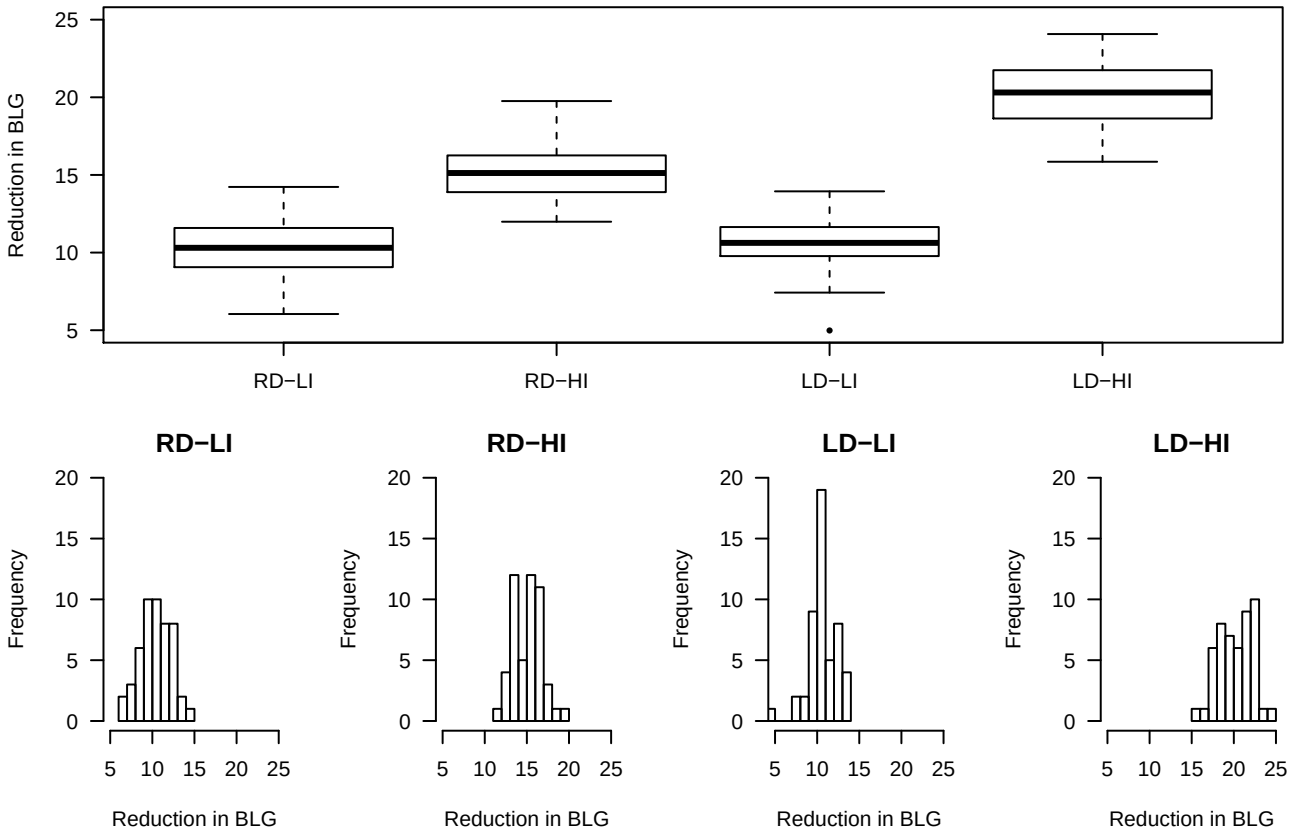

Figure 3: Boxplots and histograms of the reduction in Blood Glucose Level (BGL) for all the four treatment groups. RD: Regular Diet; LD: Low sugar Diet ; LI: Low Intensity; HI: High Intensity.

We can see in Figure 3 that for low intensity exercise program the diet type didn't show very important, with a median and variance very similar for both. For high intensity exercise program group the medians are considerably bigger, with the biggest values of reduction of blood glucose level obtained in the combination with a strict low sugar diet. In general, all the variances are similar. With this graphs we can imagine that exist a difference between the combination of treatments, with a greater emphasis in LD-HI.

Question 4

Conduct a formal test comparing the mean reduction for the four treatment groups.

Solution:

- Hyphotesis:

$H_0 = \mu_{\text{RD-LI}} = \mu_{\text{RD-HI}} = \mu_{\text{LD-LI}} = \mu_{\text{LD-HI}}$
$H_a =$ At least one mean reduction is different from the others.

- Test Statistic:

F Test.

$$F = \frac{S_4/(4-1)}{(200-4)\cdot S_p^2/(200-4)} \overset{H_0}{\sim} F_{4-1,200-4}.$$

```
# <code r> ============================================================================ #
bx_rd.li <- mean(dataY[ , 1]) ; bx_rd.hi <- mean(dataY[ , 2])
bx_ld.li <- mean(dataY[ , 3]) ; bx_ld.hi <- mean(dataY[ , 4])
bx <- mean(dataY)
# </code r> ============================================================================ #
```

$$\bar{X}_{\text{RD-LI}} = 10.3353, \quad \bar{X}_{\text{RD-HI}} = 15.0993, \quad \bar{X}_{\text{LD-LI}} = 10.6322, \quad \bar{X}_{\text{LD-HI}} = 20.227$$
$$\bar{X} = 14.0735$$

```
# <code r> ============================================================================ #
s_4 <-
  50*(bx_rd.li - bx)**2 + 50*(bx_rd.hi - bx)**2 +
  50*(bx_ld.li - bx)**2 + 50*(bx_ld.hi - bx)**2
# </code r> ============================================================================ #
```

$$\begin{aligned} S_4 &= 50\cdot(\bar{X}_{\text{RD-LI}} - \bar{X})^2 + 50\cdot(\bar{X}_{\text{RD-HI}} - \bar{X})^2 + 50\cdot(\bar{X}_{\text{LD-LI}} - \bar{X})^2 + 50\cdot(\bar{X}_{\text{LD-HI}} - \bar{X})^2 \\ &= 3236.7377 \end{aligned}$$

$$\eta_{\text{RD}-\text{LI}} = \eta_{\text{RD}-\text{HI}} = \eta_{\text{LD}-\text{LI}} = \eta_{\text{LD}-\text{HI}} = \frac{50-1}{200-4} = 0.25$$

```
# <code r> ================================================================== #
s_p.2 <-
  .25*sum((dataY[ , 1] - bx_rd.li)**2)/49 +
  .25*sum((dataY[ , 2] - bx_rd.hi)**2)/49 +
  .25*sum((dataY[ , 3] - bx_ld.li)**2)/49 +
  .25*sum((dataY[ , 4] - bx_ld.hi)**2)/49
# </code r> ================================================================= #
```

$$S^2_{\text{RD}-\text{LI}} = \sum_{i=1}^{50} \frac{(X_i^{\text{RD}-\text{LI}} - \bar{X}_{\text{RD}-\text{LI}})^2}{50-1} = 3.5005 \qquad S^2_{\text{RD}-\text{HI}} = \sum_{i=1}^{50} \frac{(X_i^{\text{RD}-\text{HI}} - \bar{X}_{\text{RD}-\text{HI}})^2}{50-1} = 2.7987$$

$$S^2_{\text{LD}-\text{LI}} = \sum_{i=1}^{50} \frac{(X_i^{\text{LD}-\text{LI}} - \bar{X}_{\text{LD}-\text{LI}})^2}{50-1} = 2.8971 \qquad S^2_{\text{LD}-\text{HI}} = \sum_{i=1}^{50} \frac{(X_i^{\text{LD}-\text{HI}} - \bar{X}_{\text{LD}-\text{HI}})^2}{50-1} = 3.7239$$

$$\begin{aligned} S_p^2 &= 0.25 \cdot S^2_{\text{RD}-\text{LI}} + 0.25 \cdot S^2_{\text{RD}-\text{HI}} 0.25 \cdot S^2_{\text{LD}-\text{LI}} + 0.25 \cdot S^2_{\text{LD}-\text{HI}} \\ &= 3.23 \end{aligned}$$

```
# <code r> ================================================================== #
f <- (s_4/3)/s_p.2
# </code r> ================================================================= #
```

$$\boxed{F = \frac{S_4/3}{S_p^2} = 334.0259.}$$

- Critical value:

Considering a $\boxed{\mathbb{P}(\text{Error type I}) = \alpha = 0.05.}$

```
# <code r> ================================================================== #
cv <- qf(.95, 3, 196)
# </code r> ================================================================= #
```

$$\boxed{F = 334.0259 \overset{H_0}{\sim} F_{3,196;0.95} = 2.6507.}$$

- Decision-making:

The test statistic value is <u>much</u> bigger than the critical value, $334.0259 > 2.6507$. Therefore we don't accept the null hyphotesis, H_0. We have stronger statistical evidence that at least one mean reduction is different from the others (at least one treatment group is different from the others).

Question 5

The clinician believes that the mean reduction among those engaged in high intensity exercise is greater than those engaged in low intensity exercise. Conduct a formal test of hypothesis.

Solution:

- Hyphotesis:

$$\boxed{H_0 = \mu_{\text{HI}} = \mu_{\text{LI}} \qquad H_a = \mu_{\text{HI}} > \mu_{\text{LI}}}$$

- Test Statistic:

$\boxed{t \text{ Test.}}$

$$t = \frac{\bar{X}_{\text{HI}} - \bar{X}_{\text{LI}}}{\sqrt{S_p^2\left(\frac{1}{100} + \frac{1}{100}\right)}} \overset{H_0}{\sim} t_{99+99}.$$

```
# <code r> ================================================================== #
bx_hi <- mean(dataY[ , c(2, 4)]) ; bx_li <- mean(dataY[ , c(1, 3)])
# </code r> ================================================================= #
```

$$\bar{X}_{\text{HI}} = 17.6632, \qquad \bar{X}_{\text{LI}} = 10.4838$$
$$W_{\text{HI}} = W_{\text{LI}} = \frac{99}{99+99} = 0.5$$

```
# <code r> ================================================================== #
s_p.2 <-
  .5*sum((c(dataY[ , c(2, 4)]) - bx_hi)**2)/99 +
  .5*sum((c(dataY[ , c(1, 3)]) - bx_li)**2)/99
# </code r> ================================================================= #
```

$$S^2_{\text{HI}} = \sum_{i=1}^{100} \frac{(X_i^{\text{HI}} - \bar{X}_{\text{HI}})^2}{100-1} = 9.868 \qquad S^2_{\text{LI}} = \sum_{i=1}^{100} \frac{(X_i^{\text{LI}} - \bar{X}_{\text{LI}})^2}{100-1} = 3.1887$$

$$\begin{aligned} S_p^2 &= 0.5 \cdot S^2_{\text{HI}} + 0.5 \cdot S^2_{\text{LI}} \\ &= 6.5284 \end{aligned}$$

```
# <code r> ================================================================== #
t <- (bx_hi - bx_li)/sqrt(s_p.2*(2/100))
# </code r> ================================================================= #
```

$$\boxed{t = \frac{\bar{X}_{\text{HI}} - \bar{X}_{\text{LI}}}{\sqrt{S_p^2\left(\frac{2}{100}\right)}} = 19.8688.}$$

• Critical value:

Considering a $\boxed{\mathbb{P}(\text{Error type I}) = \alpha = 0.05.}$

```
# <code r> ================================================================== #
cv <- qt(.95, 198)
# </code r> ================================================================= #
```

$$\boxed{t = 19.8688 \overset{H_0}{\sim} t_{198;0.95} = 1.6526. \quad \text{(one-tail)}}$$

• Decision-making:

The test statistic value is much bigger than the critical value, $19.8688 > 1.6526$. Therefore we don't accept the null hyphotesis, H_0. We have strong statistical evidence that the mean reduction among those engaged in high intensity exercise is greater than those engaged in low intensity exercise.

Question 6

The clinician believes that the mean reduction among those given a strict low sugar diet is greater than those given only regular diet. Conduct a formal test of hypothesis.

Solution:

• Hyphotesis:

$$\boxed{H_0 = \mu_{\text{LD}} = \mu_{\text{RD}} \qquad H_a = \mu_{\text{LD}} > \mu_{\text{RD}}}$$

• Test Statistic:

$\boxed{t \text{ Test.}}$

$$t = \frac{\bar{X}_{\text{LD}} - \bar{X}_{\text{RD}}}{\sqrt{S_p^2\left(\frac{1}{100} + \frac{1}{100}\right)}} \overset{H_0}{\sim} t_{99+99}.$$

```
# <code r> ================================================================== #
bx_ld <- mean(dataY[ , 3:4]) ; bx_rd <- mean(dataY[ , 1:2])
# </code r> ================================================================= #
```

$$\bar{X}_{\text{LD}} = 15.4296, \qquad \bar{X}_{\text{RD}} = 12.7173$$
$$W_{\text{LD}} = W_{\text{RD}} = \frac{99}{99+99} = 0.5$$

```
# <code r> ============================================================================ #
s_p.2 <-
  .5*sum((c(dataY[ , 3:4]) - bx_ld)**2)/99 +
  .5*sum((c(dataY[ , 1:2]) - bx_rd)**2)/99
# </code r> ============================================================================ #
```

$$S^2_{\text{LD}} = \sum_{i=1}^{100} \frac{(X_i^{\text{LD}} - \bar{X}_{\text{LD}})^2}{100-1} = 26.5247 \qquad S^2_{\text{RD}} = \sum_{i=1}^{100} \frac{(X_i^{\text{RD}} - \bar{X}_{\text{RD}})^2}{100-1} = 8.8491$$

$$\begin{aligned} S_p^2 &= 0.5 \cdot S^2_{\text{LD}} + 0.5 \cdot S^2_{\text{RD}} \\ &= 17.6869 \end{aligned}$$

```
# <code r> ============================================================================ #
t <- (bx_ld - bx_rd)/sqrt(s_p.2*(2/100))
# </code r> ============================================================================ #
```

$$\boxed{t = \frac{\bar{X}_{\text{LD}} - \bar{X}_{\text{RD}}}{\sqrt{S_p^2\left(\frac{2}{100}\right)}} = 4.5603.}$$

- Critical value:

Considering a $\boxed{\mathbb{P}(\text{Error type I}) = \alpha = 0.05.}$

```
# <code r> ============================================================================ #
cv <- qt(.95, 198)
# </code r> ============================================================================ #
```

$$\boxed{t = 4.5603 \overset{H_0}{\sim} t_{198;0.95} = 1.6526. \quad \text{(one-tail)}}$$

- Decision-making:

The test statistic value is bigger than the critical value, $4.5603 > 1.6526$. Therefore we don't accept the null hyphotesis, H_0. We have statistical evidence that the mean reduction among those given a strict low sugardiet is greater than those given only regular diet.

■