**ENV 503: Statistics for Bioinformatics**

**Homework Set #4**

**Due: September 26, 2018**

*Instructions:*

*Use R to complete this assignment. Assignment is to be submitted via Blackboard.*

Use the R dataset **Pima.tr** to answer questions 1 and 2, and the R dataset **immer** to answer question 3. Both of these datasets can be found in the MASS library, as well as downloadable in CSV format from the website https://vincentarelbundock.github.io/Rdatasets/datasets.html

1. The dataset **Pima.tr** contains data on 200 women of Pima Indian heritage living near Phoenix, Arizona. All women were tested for diabetes according to World Health Organization criteria. The dataset contains whether or not they met WHO criteria for diabetes (variable type=Yes for diabetes, No otherwise), their plasma glucose concentration (variable glu), as well as other variables. Subset the data to create two separate datasets: one for women who met criteria for diabetes and another dataset for women who didn’t. Perform the following one-sample t tests:
   1. Does the mean blood glucose level in the women considered diabetic differ from the cutoff of 200 mg/dL that is indicative of diabetes?

> t.test(diabetic.yes$glu, mu = 200,alternative = "greater")

One Sample t-test

data: diabetic.yes$glu

t = -15.041, df = 67, p-value = 1

alternative hypothesis: true mean is greater than 200

95 percent confidence interval:

138.9665 Inf

sample estimates:

mean of x

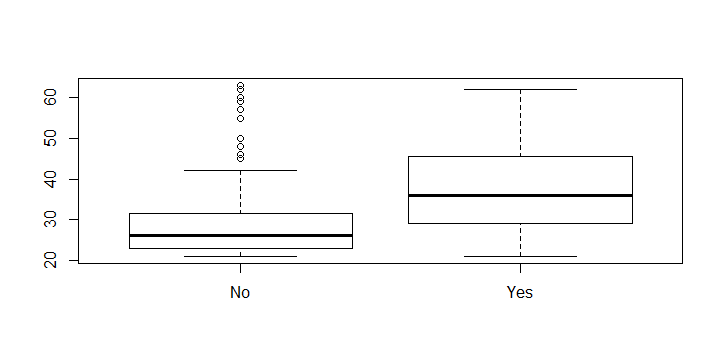
**Summarization**: From the output, we can see that the mean blood glucose level for the women considered diabetic is 145.588. Since, the p-value =1 is not less than the significance level of 0.05, we cannot reject the null hypothesis that the mean blood glucose level is equal to 200 ml.

* 1. Does the mean blood glucose level in the women considered not diabetic differ from the normal blood glucose value of 140 mg/dL?

|  |
| --- |
| > t.test(diabetic.no$glu, mu = 140,alternative = "greater")  One Sample t-test  data: diabetic.no$glu  t = -11.6, df = 131, p-value = 1  alternative hypothesis: true mean is greater than 140  95 percent confidence interval:  109.2653 Inf  sample estimates:  mean of x  113.1061  **Summarization**: From the output, we can see that the mean blood glucose level for the women considered not diabetic is 113.1061. Since, the p-value =1 is not less than the significance level of 0.05, we cannot reject the null hypothesis that the mean blood glucose level is equal to 140 ml. |
|  |
| |  | | --- | |  | |  | |  | | |  | | --- | |  | | |

1. Perform two-sample t tests to determine if the following variables in **Pima.tr** are associated with meeting criteria for diabetes: age, diastolic blood pressure, body mass index, and triceps skin fold thickness. For each of these variables, make a box plot to visualize the association between it and diabetes.

Boxplot (age ~ type, Pima.tr)



These box plots suggest that 50 and 60 age group have more diabetics.

t.test(age ~ type, Pima.tr)

Welch Two Sample t-test

data: age by type

t = -5.2162, df = 115.7, p-value = 8.106e-07

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.667372 -5.245284

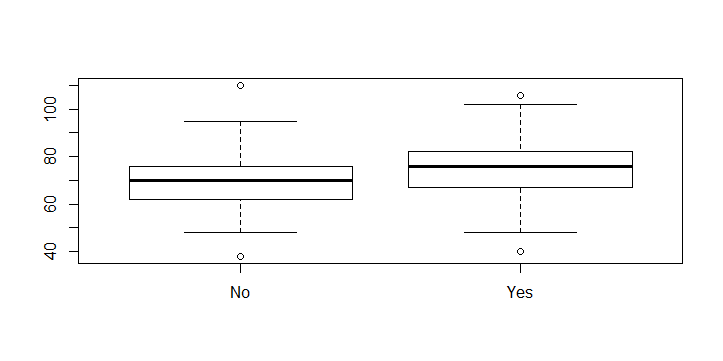
sample estimates:

mean in group No mean in group Yes

29.23485 37.69118

From the output, we can see that the difference between group no and group is -8.4563.P value 8.106e-07 proved that this data are significantly heteroscedastic, and thus the assumption of homoscedasticity in the regression residuals is violated. In this case the data violate the assumption of homoscedasticity.

boxplot(bp ~ type, Pima.tr)



These box plots suggest that 80 to 100 bp group has more diabetics.

t.test(bp ~ type, Pima.tr)

Welch Two Sample t-test

data: bp by type

t = -2.9592, df = 130.28, p-value = 0.003665

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-8.414080 -1.671482

sample estimates:

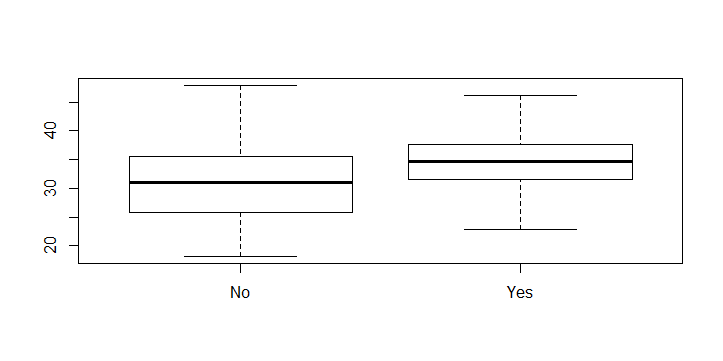
mean in group No mean in group Yes

69.54545 74.58824

From the output, we can see that the difference between group no and group is -8.P value 0.0036 is less than the significant value 0.05 which proved that the null

hypothesis is rejected.

boxplot(bmi ~ type, Pima.tr)



These box plots suggest that morethan 35 group has more diabetics.

> t.test(bmi ~ type, Pima.tr)

Welch Two Sample t-test

data: bmi by type

t = -4.512, df = 171.46, p-value = 1.188e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-5.224615 -2.044547

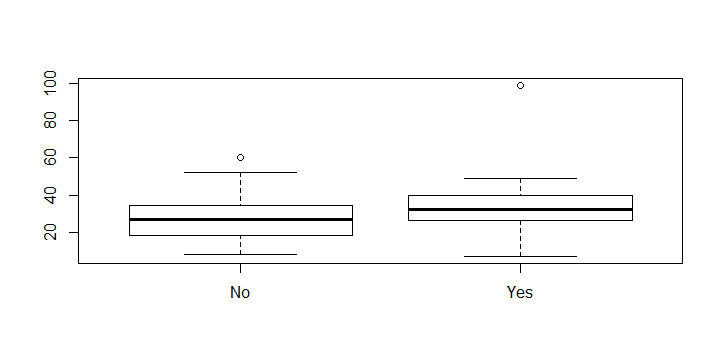
sample estimates:

mean in group No mean in group Yes

31.07424 34.70882

From the output, we can see that the difference between group no and group yes is -3.63 .P value 1.188e-05 that this data are significantly heteroscedastic, and thus the assumption of homoscedasticity in the regression residuals is violated. In this case the data violate the assumption of homoscedasticity.

boxplot(skin ~ type, Pima.tr)



These box plots suggest that more than 40 group has more diabetics.

> t.test(skin ~ type, Pima.tr)

Welch Two Sample t-test

data: skin by type

t = -3.3421, df = 122.23, p-value = 0.001104

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-9.415489 -2.410715

sample estimates:

mean in group No mean in group Yes

27.20455 33.11765

From the output, we can see that the difference between group no and group yes is -5.91 .P value 0.001104 is less than the significant value 0.05 which proved that thenunull hypothesis is rejected.

1. The dataset **immer** contains yields from an agricultural field trial in which six varieties of barley were grown in six locations in 1931 and 1932. Perform a paired t test to determine if the barley yield was the same for these two years.

t.test(immer$Y1, immer$Y2, paired=TRUE)

Paired t-test

data: immer$Y1 and immer$Y2

t = 3.324, df = 29, p-value = 0.002413

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

6.121954 25.704713

sample estimates:

mean of the differences

15.91333

From the output, we can see that the mean of the differences is 15.91333 .P value 0.002413 is less than the significant value 0.05 which proved that the null hypothesis is rejected.

**For each test performed, write a sentence or two summarizing the results of the test. Full credit for these problems will not be given for R output alone. Please put all components of your answers (R output, figures, results summaries) into a single Word document. Including the R code is not necessary. Thanks ☺**