assignment4

2022-09-27

R Markdown

The data for this question are stored in the file hw4q1 and correspond to an experiment to measure the effect of a new drug in the memory of patients in a nursing home. The patients were tested for memory before the treatment started and again after one month taking the drug. The dataset has two variables, mem, the score in the test, and type with two values, before for the initial score and after for the final score (a) Load the dataset and check whether type is stored as a factor. If it is not, transform it into a factor.

```
dataset = read.table("hw4q1")
str(dataset)
```

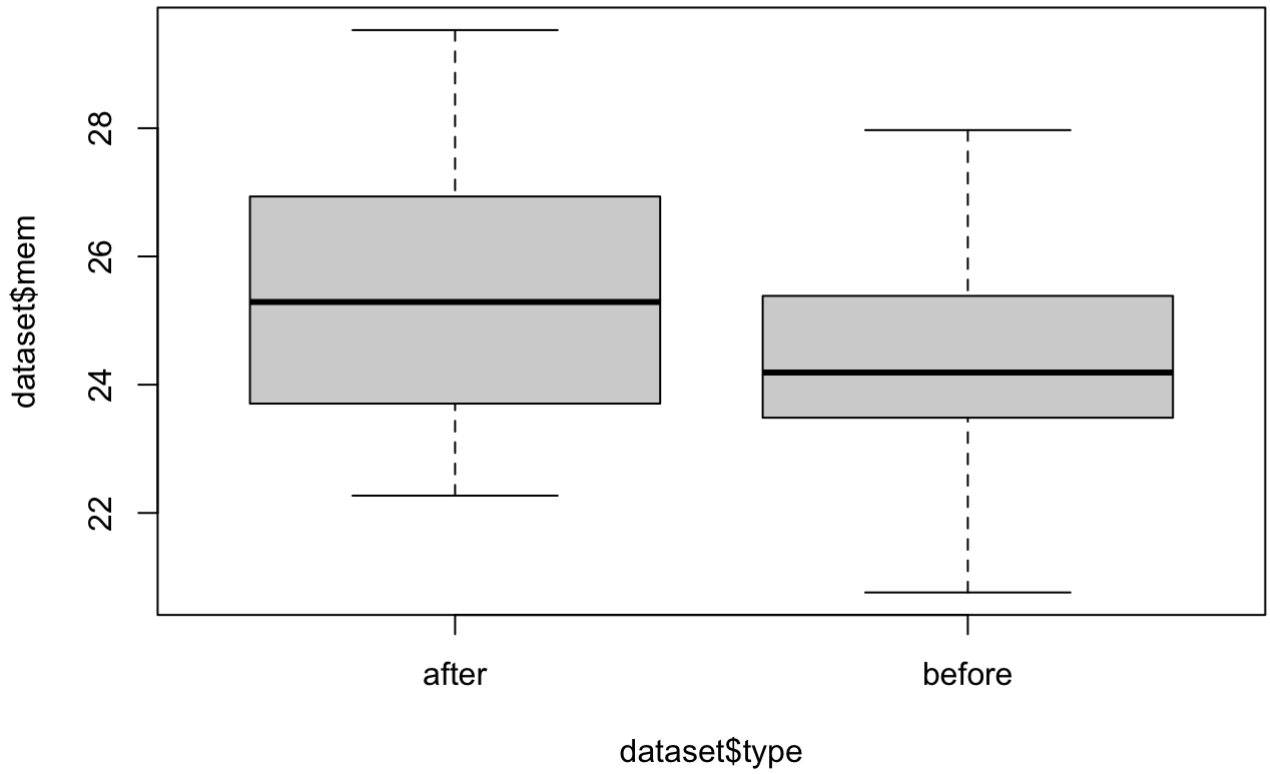
```
## 'data.frame':    30 obs. of  2 variables:
##  $ type: chr  "before" "before" "before" "before" ...
##  $ mem : num  24.2 25.6 20.8 22.2 26.3 ...
```

```
dataset$type = as.factor(dataset$type)
str(dataset)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ type: Factor w/ 2 levels "after","before": 2 2 2 2 2 2 2 2 2 2 ...
##  $ mem : num  24.2 25.6 20.8 22.2 26.3 ...
```

b. Draw boxplots for mem according to type and comment.

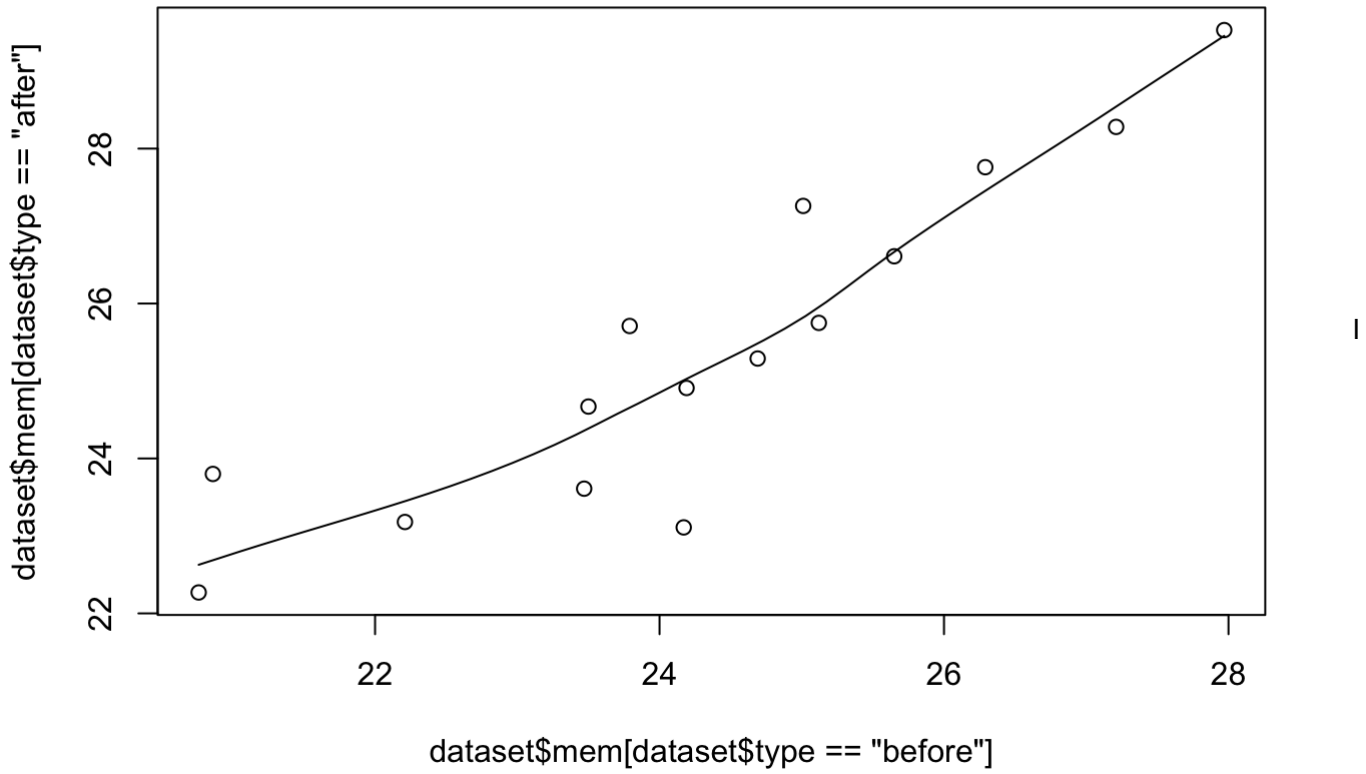
```
boxplot(dataset$mem ~dataset$type)
```



after one month taking the drug, the patients obviously have better score in test.

c. Draw a scatterplot of the memory score after versus the memory score before and comment on what you observe. Do you think the two scores are independent?

```
scatter.smooth(dataset$mem[dataset$type=="before"],dataset$mem[dataset$type=="after"
])
```



think the two scores are dependent. Because the polt nearly a straight line.

d. We want to determine whether the treatment had an impact on the memory score of the patients. State clearly the statistical hypothesis that you want to test. What test or tests would you consider adequate in this situation and why? What are the assumptions? Are they satisfied in this case? Carry out all appropriate tests for this problem and comment on your results.

```
t.test(dataset$mem[dataset$type=="before"],dataset$mem[dataset$type=="after"], var.eq
ual = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dataset$mem[dataset$type == "before"] and dataset$mem[dataset$type == "afte
r"]
## t = -1.4742, df = 28, p-value = 0.1516
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.6842158  0.4375492
## sample estimates:
## mean of x mean of y
##  24.32600  25.44933
```

H0: mean1 == mean2 H1: mean1 != mean2 We can use t test. Because n <30, std is unknown. We can reject H0, so the treatment actually had an impact on the memory score of the patients.

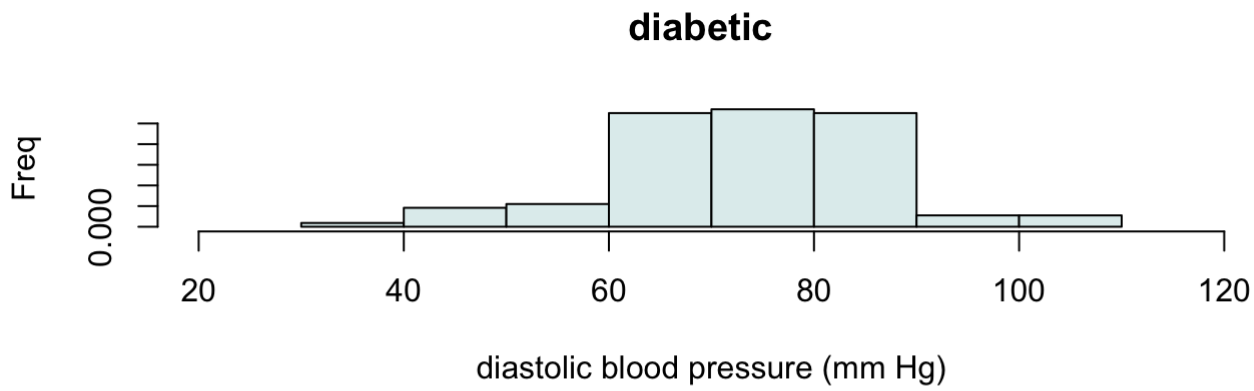
Question 2

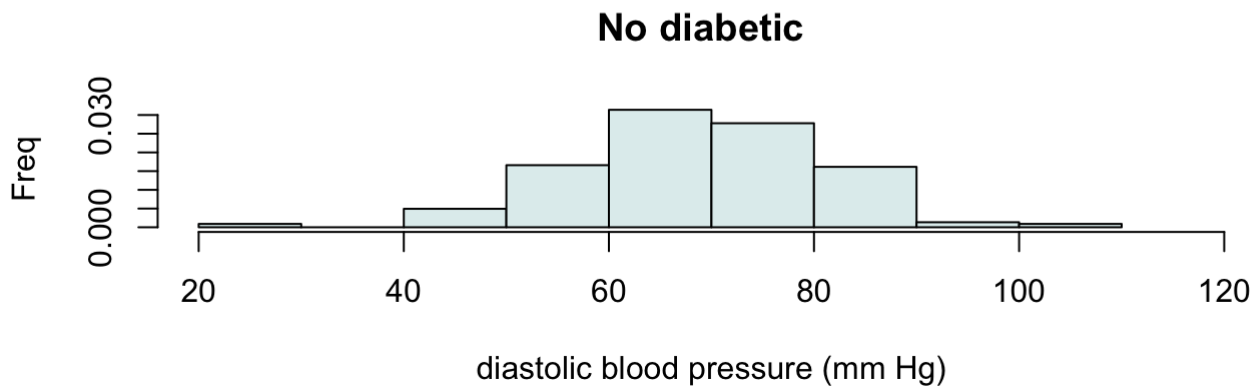
We will use the data set Pima.te in the MASS package for this question. Open the help file for this data set and get acquainted with it. We are going to focus on two variables, bp and type. (a) Divide the plotting window into two regions, one single column with two rows, and plot histograms for bp for types Yes and No. Since you want to use these graphs for comparing the two populations, use the same scales in both cases. Use reasonable labels for the axes and a title indicating the corresponding type. Make sure that the area for the figure is large enough so that the histograms are clearly seen. Compare the two graphs and comment on similarities and differences.

```
library(MASS)
pim = MASS::Pima.te
str(pim)
```

```
## 'data.frame':    332 obs. of  8 variables:
##  $ npreg: int  6 1 1 3 2 5 0 1 3 9 ...
##  $ glu  : int  148 85 89 78 197 166 118 103 126 119 ...
##  $ bp   : int  72 66 66 50 70 72 84 30 88 80 ...
##  $ skin : int  35 29 23 32 45 19 47 38 41 35 ...
##  $ bmi  : num  33.6 26.6 28.1 31 30.5 25.8 45.8 43.3 39.3 29 ...
##  $ ped  : num  0.627 0.351 0.167 0.248 0.158 0.587 0.551 0.183 0.704 0.263 ...
##  $ age  : int  50 31 21 26 53 51 31 33 27 29 ...
##  $ type : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 2 1 1 2 ...
```

```
par(mfcol=c(2,1))
hist(pim$bp[pim$type=="Yes"],col = 'azure2',xlab='diastolic blood pressure (mm Hg)',x
lim=c(20,120),  ylab='Freq',main = "diabetic  ", prob= TRUE)
hist(pim$bp[pim$type=="No"],col = 'azure2',xlab='diastolic blood pressure (mm Hg)',xl
im=c(20,120),  ylab='Freq',main = "No diabetic  ", prob= TRUE)
```

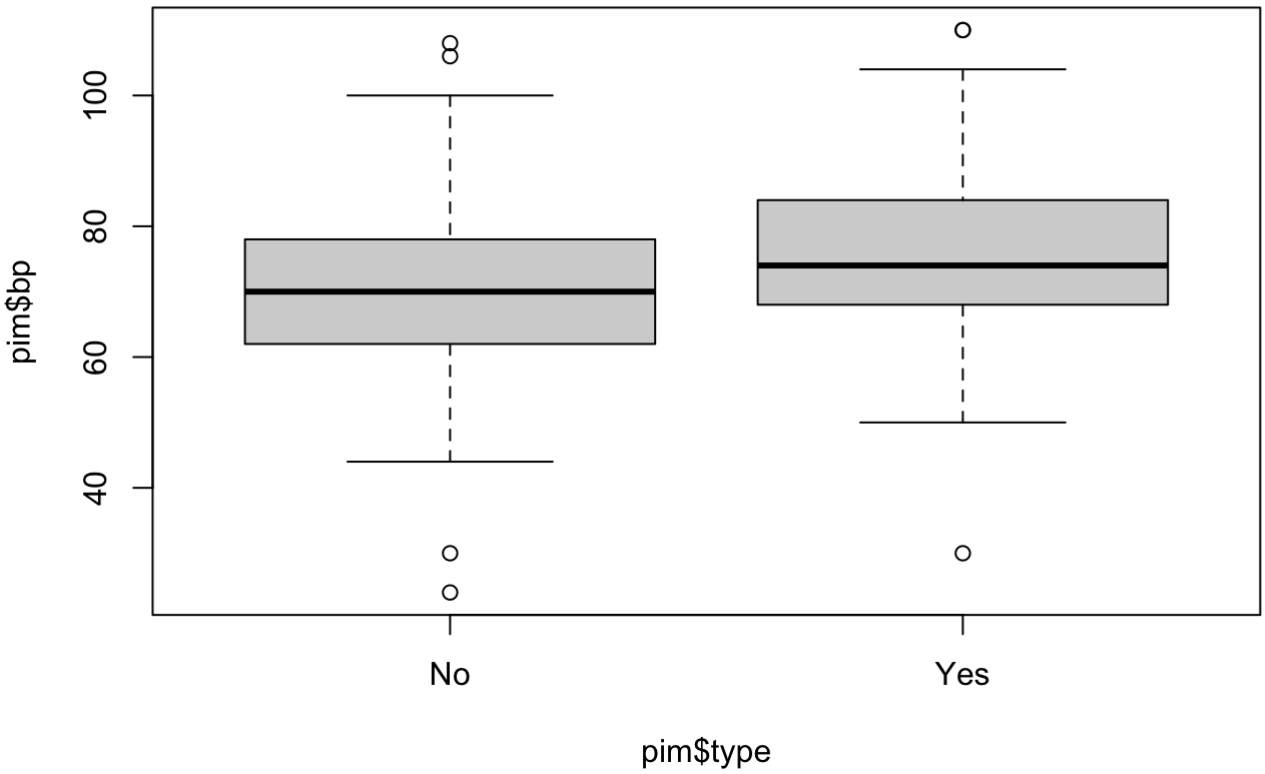




Similarties: blood pressure mainly in 60-80. Differences: Diabetic patients have more people over 80 blood pressure.

b. Boxplot blood pressure as a function of type and comment on the graph. Make sure you have a single plotting window with both boxplot.

```
boxplot(pim$bp~pim$type)
```



Diabetic people have higher mean blood pressure.

c. Calculate mean and standard deviation for both types and find how many subjects of each type are there in the dataset.

```
mean(pim$bp[pim$type=="Yes"])
```

```
## [1] 74.77064
```

```
sd(pim$bp[pim$type=="Yes"])
```

```
## [1] 13.12803
```

```
#count(pim$bp[pim$type=="Yes"])
mean(pim$bp[pim$type=="No"])
```

```
## [1] 70.13004
```

```
sd(pim$bp[pim$type=="No"])
```

```
## [1] 12.38192
```

```
#count(pim$bp[pim$type=="No"])
table(pim$type)
```

```
##
##  No Yes
## 223 109
```

d. We want to determine if the pulse rate for diabetic women is significantly different from a reference value of 70 mm Hg. What (parametric) statistical test do you think is appropriate in this case? Carry this test out and discuss your results. Describe the assumptions you need for this test to be valid and check whether they are satisfied by the data set.

```
(tn <- (mean(pim$bp[pim$type=="Yes"])-70)/(sd(pim$bp[pim$type=="Yes"])/sqrt(108)))
```

```
## [1] 3.776498
```

```
t.test(pim$bp[pim$type=="Yes"], mu=70)
```

```
##
##  One Sample t-test
##
## data:  pim$bp[pim$type == "Yes"]
## t = 3.7939, df = 108, p-value = 0.0002449
## alternative hypothesis: true mean is not equal to 70
## 95 percent confidence interval:
##  72.27818 77.26310
## sample estimates:
## mean of x
##  74.77064
```

H0: meanbp==70 H1: meanbp!=70

The p value is just below 0.05 and the decision depends on our choice for α. If we choose 0.05,0.02 or 0.01,the null hypothesis is rejected. Thus, the pulse rate for diabetic women is significantly different from a reference value of 70 mm Hg.

e. We now want to compare the two populations (Pima women with and without diabetes) to see if there is a difference in the average blood pressure. What (parametric) test would you perform in this case? What assumptions are needed? Do they look reasonable in this case? Carry out this test and discuss your results.

```
t.test(pim$bp[pim$type=="Yes"],pim$bp[pim$type=="No"], var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  pim$bp[pim$type == "Yes"] and pim$bp[pim$type == "No"]
## t = 3.1437, df = 330, p-value = 0.00182
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.736687 7.544508
## sample estimates:
## mean of x mean of y
##  74.77064  70.13004
```

H0: mean1 == mean2 H1: mean1 != mean2 We can use t test. Because std is unknown. We can reject H0, so the treatment actually had an impact on the memory score of the patients.

f. What non-parametric tests would be adequate for parts (d) and (e)? Carry this test out and compare your results with the tests in (d) and (e).

```
wilcox.test(pim$bp[pim$type=="Yes"], mu=70)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pim$bp[pim$type == "Yes"]
## V = 3532.5, p-value = 0.0002201
## alternative hypothesis: true location is not equal to 70
```

```
wilcox.test(pim$bp[pim$type=="Yes"],pim$bp[pim$type=="No"], alternative = 'less')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pim$bp[pim$type == "Yes"] and pim$bp[pim$type == "No"]
## W = 14822, p-value = 0.9994
## alternative hypothesis: true location shift is less than 0
```

The result is same with (d) But the result is not same with (e),so the treatment doesn’t had an impact on the memory score of the patients.