STAT 210
Applied Statistics and Data Analysis:
Homework 4 - Solution

Due on Oct. 2/2022

Question 1

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.

```
data2 <- read.table('hw4q1')
str(data2)
```

```
## '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 ...
```

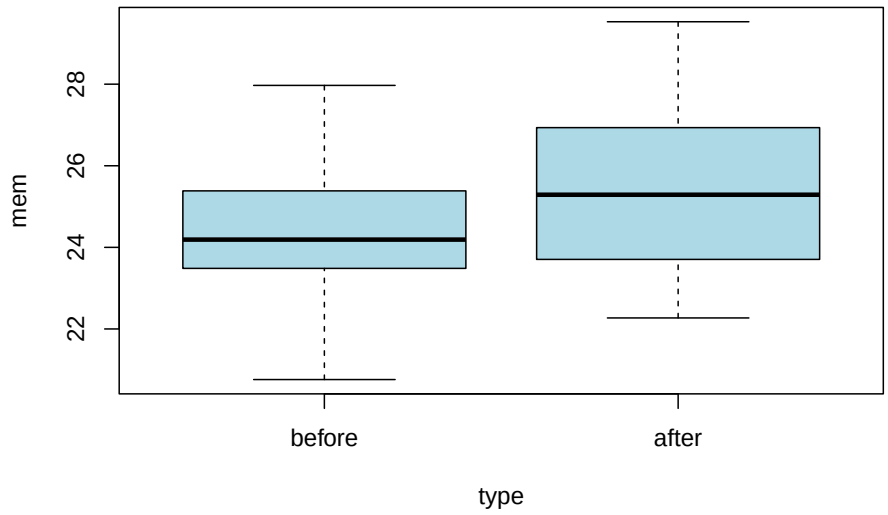
The variable `type` is stored in character mode. We transform it into a factor

```
data2$type <- factor(data2$type, levels = c('before','after'))
str(data2)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ type: Factor w/ 2 levels "before","after": 1 1 1 1 1 1 1 1 1 1 ...
##  $ mem : num  24.2 25.6 20.8 22.2 26.3 ...
```

(b) Draw boxplots for `mem` according to `type` and comment.

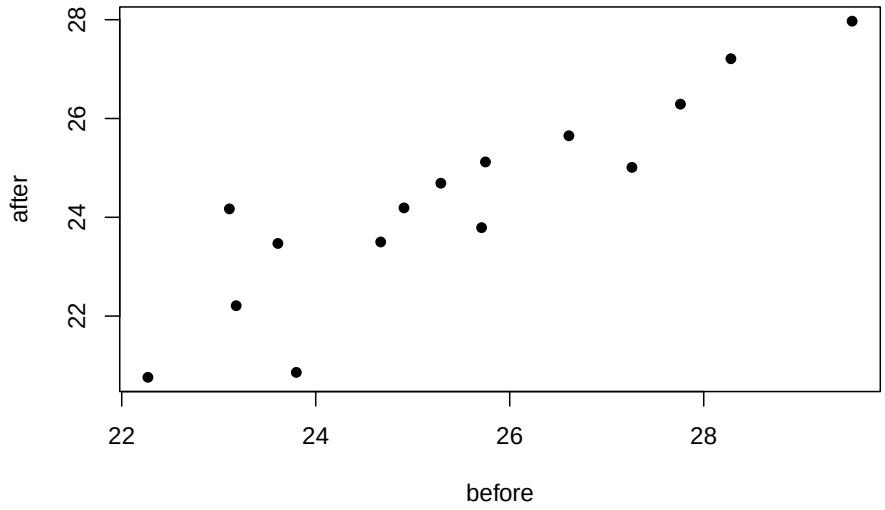
```
boxplot(mem ~ type, data = data2, col = 'lightblue')
```



There seems to be an increase of the memory score after the treatment. Both datasets look symmetric and have similar dispersion.

(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?

```
plot(data2$mem[data2$type == 'after'], data2$mem[data2$type == 'before'],
     xlab = 'before', ylab = 'after', pch = 16)
```



We see that there seems to be a linear relation between the scores before and after treatment. This indicates that these variables are correlated.

(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.

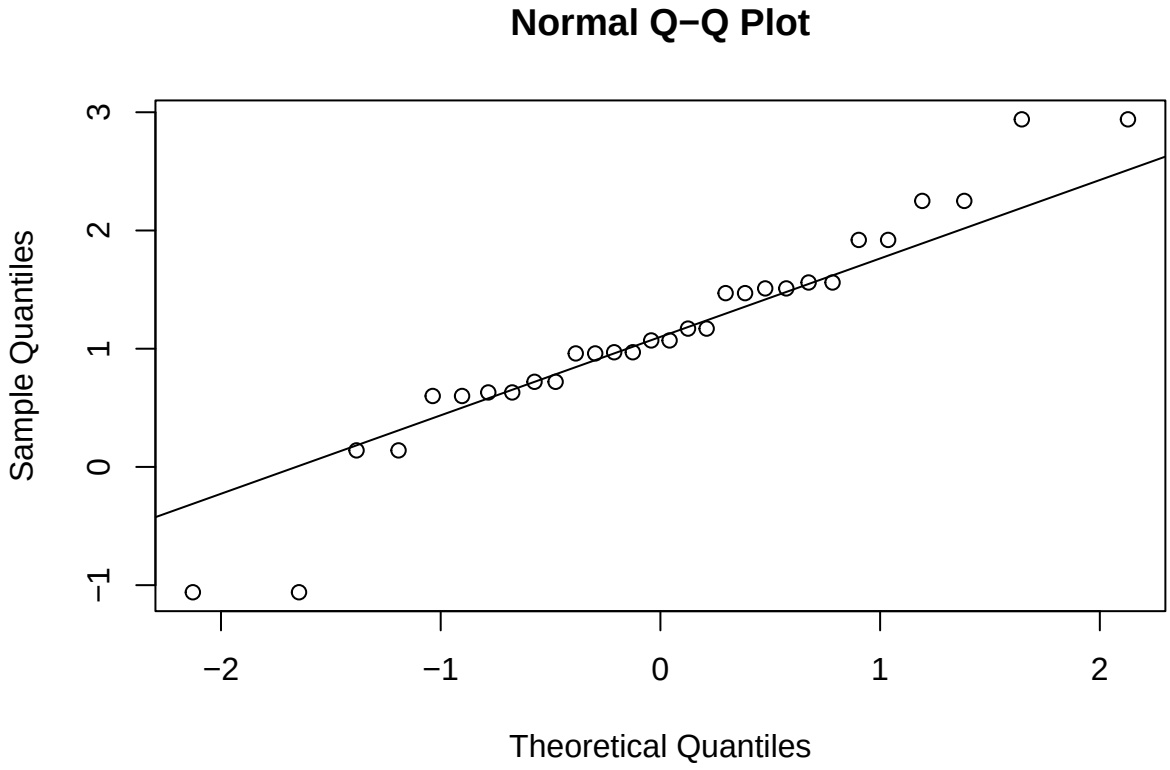
Since we have measurements before and after treatment for the same subjects we should do a paired test. There are two options, the paired t-test and the paired Wilcoxon test.

If m_i is the memory score before treatment and w_i the score after treatment for subject $i, i = 1, 2, \ldots, 15$, define $d_i = w_i - m_i$. Let μ_d be the (population) average for this difference. We want to test

$$H_0 : \mu_d = 0 \quad vs. \quad H_A : \mu_d \neq 0$$

The paired t-test assumes that the differences follow a normal distribution. To check this we can use a quantile plot. We create a vector with the difference values:

```
data2$dif <- data2$mem[data2$type == 'after'] -
  data2$mem[data2$type == 'before']
qqnorm(data2$dif)
qqline(data2$dif)
```




```
shapiro.test(data2$dif)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data2$dif
## W = 0.94688, p-value = 0.1394
```

We have no grounds to reject the hypothesis of normality. We now do a paired t-test

```
t.test(data2$mem[data2$type == 'after'],
       data2$mem[data2$type == 'before'],
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data2$mem[data2$type == "after"] and data2$mem[data2$type == "before"]
## t = 4.6733, df = 14, p-value = 0.0003589
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6077823 1.6388844
## sample estimates:
## mean of the differences
##                1.123333
```

The p-value is small, so we reject the null hypothesis of equal means, which is equivalent to the treatment having no effect.

The other test we have available is Wilcoxon's. This test assumes that the distribution is continuous and symmetric with respect to the mean, which seems to be true in this case (since the hypothesis of normality was not rejected).

```
wilcox.test(data2$mem[data2$type == 'after'],
       data2$mem[data2$type == 'before'],
       paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data2$mem[data2$type == "after"] and data2$mem[data2$type == "before"]
## V = 113, p-value = 0.00116
## alternative hypothesis: true location shift is not equal to 0
```

Again, the p-value is small and we have the same conclusion as before.

Observe that if you do not do a paired test the conclusion is different:

```
t.test(data2$mem[data2$type == 'after'],
       data2$mem[data2$type == 'before'])
```

```
##
##  Welch Two Sample t-test
##
## data:  data2$mem[data2$type == "after"] and data2$mem[data2$type == "before"]
## t = 1.4742, df = 27.981, p-value = 0.1516
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4375963  2.6842629
## sample estimates:
## mean of x mean of y
##  25.44933  24.32600
```

The p value now is large and we would not reject the null hypothesis.

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`.

```
library(MASS)
str(Pima.te)
```

```
## '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 ...
```

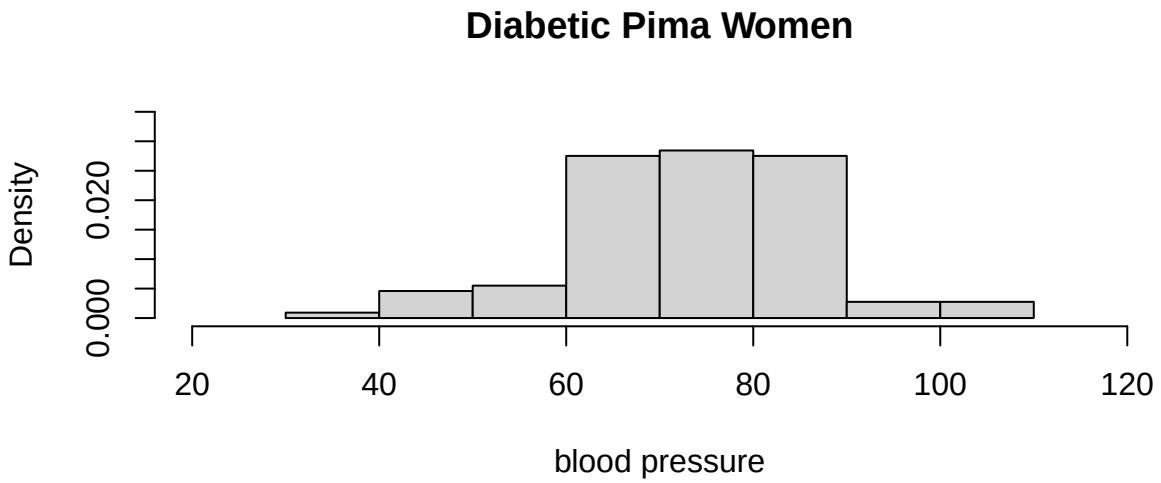
(a) Divide the plotting window into two regions, one single column with two rows, and plot histograms for `bp` for types `Yes` and `No`. Since you want to use these graphs for comparing the two populations, use the same scales in both cases. Use reasonable labels for the axes and a title indicating the 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.

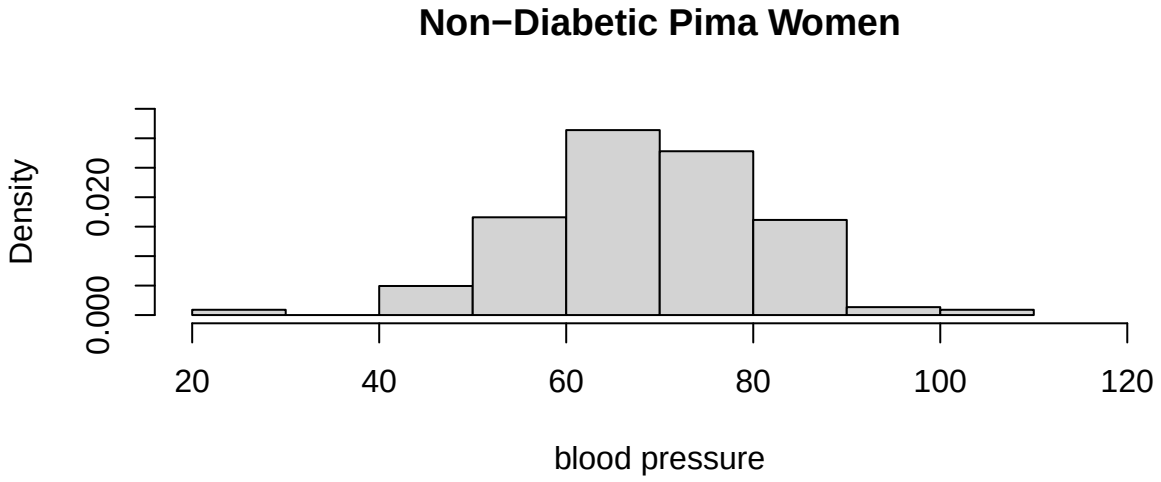
```
par(mfrow = c(2,1))
range(Pima.te$bp)
```

```
## [1]  24 110
```

The range of values goes from 24 to 110. We use as limits for the x-axis 20 and 120. For convenience we create two vectors with the values for `No` and `Yes` and plot histograms of relative frequencies. We set a common scale for the y axis by trial and error and use a large plotting window to guarantee that there is enough space for both histograms.

```
bp.yes <- Pima.te$bp[Pima.te$type == 'Yes']
bp.no <- Pima.te$bp[Pima.te$type == 'No']
par(mfrow=c(2,1))
hist(bp.yes, xlim = c(20,120), ylim = c(0, 0.035), freq = F,
     xlab = 'blood pressure', main = 'Diabetic Pima Women')
hist(bp.no, xlim = c(20,120), ylim = c(0, 0.035), freq = F,
     xlab = 'blood pressure', main = 'Non-Diabetic Pima Women')
```

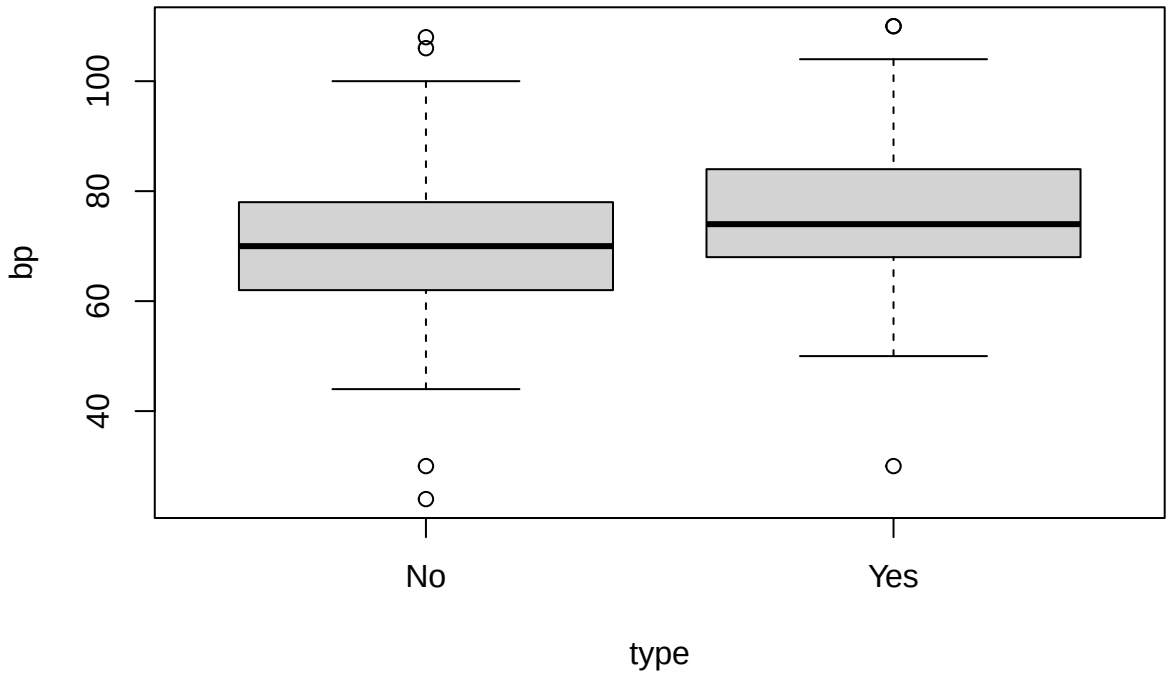




We see that diabetic women tend to have higher blood pressure than non-diabetic women. The dispersion of the data seems to be similar for both populations.

(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.

```
plot(bp ~ type, data = Pima.te)
```



As before, we see that values of blood pressure for diabetic women tend to be higher. The height of the boxes (the inter-quartile range) is similar for both populations.

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

```
tapply(Pima.te$bp, Pima.te$type, mean)
```

```
##       No      Yes
## 70.13004 74.77064
```

```
tapply(Pima.te$bp, Pima.te$type, sd)
```

```
##       No      Yes
## 12.38192 13.12803
```

```
sum(Pima.te$type == 'Yes')
```

```
## [1] 109
```

```
sum(Pima.te$type == 'No')
```

```
## [1] 223
```

(d) We want to determine if the pulse rate for diabetic women 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.

The one-sample t-test is appropriate here. The test assumes that the sampling distribution for the (sample) mean is approximately Gaussian. Since the sample size is large, 109 points, the Central Limit Theorem says that this a reasonable assumption.

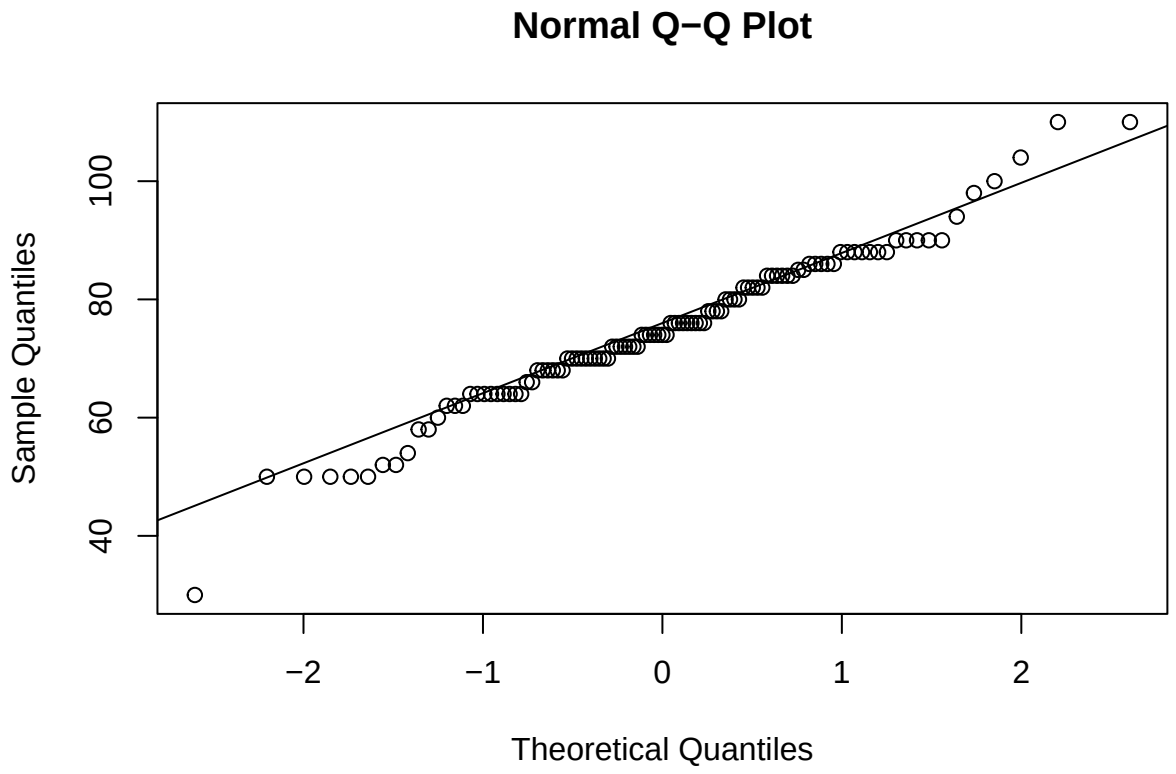
```
t.test(bp.yes, mu = 70)
```

```
##
##  One Sample t-test
##
## data:  bp.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
```

The p-value is very small, so we would reject the null hypothesis that the true mean for the population of diabetic Pima women is 70.

We can test normality by doing a quantile plot:

```
qqnorm(bp.yes); qqline(bp.yes)
```



The plot looks reasonable, although there are many repeated values in the observations (this is due to the fact that blood pressure is integer-value) and the graph looks like a step function. To confirm that the assumption of Gaussianity is reasonable, we can carry out a Shapiro-Wilk test for normality

```
shapiro.test(bp.yes)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bp.yes
## W = 0.97961, p-value = 0.09311
```

We would not reject the hypothesis of normality with this p-value at the usual levels.

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

In this case we can use a two-sample t-test, comparing the blood pressure for diabetic and non-diabetic Pima women. The test assumes that the sampling distribution for the means is approximately Gaussian. Since the sample sizes are large, 109 and 223 points, the Central Limit Theorem says that this a reasonable assumption.

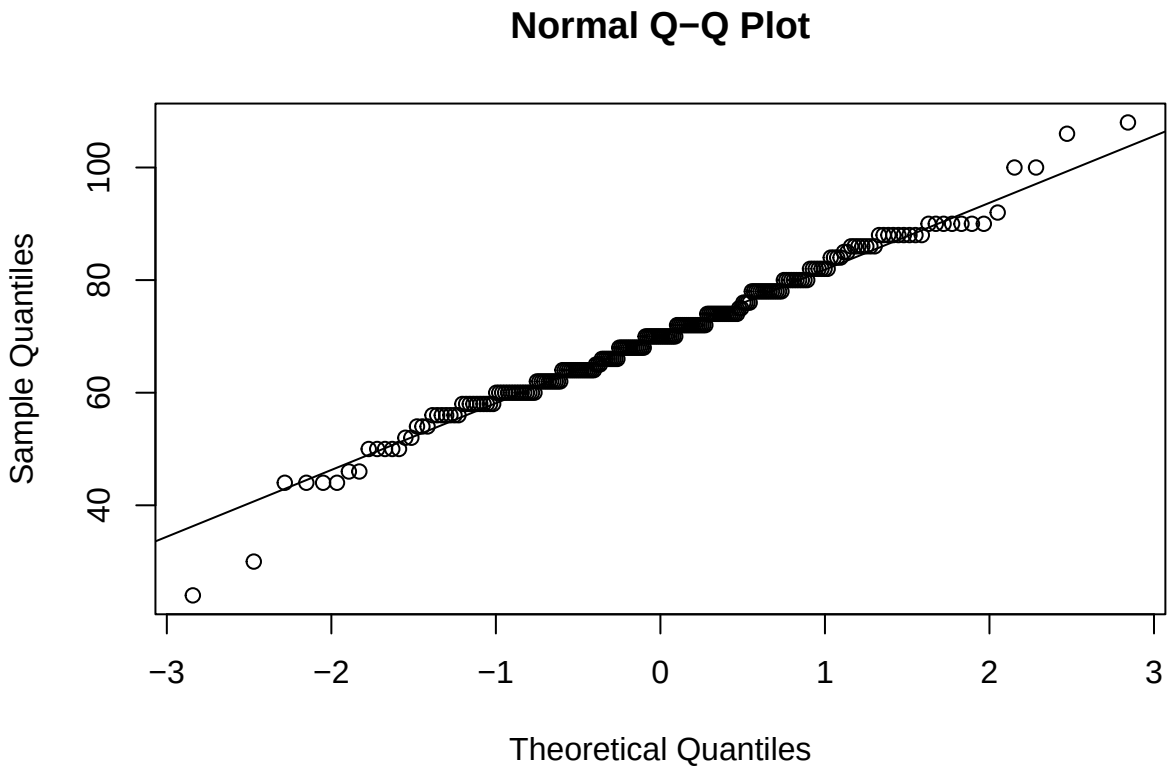
```
t.test(bp ~ type, data = Pima.te)
```

```
##
##  Welch Two Sample t-test
##
## data:  bp by type
## t = -3.081, df = 203.61, p-value = 0.002348
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -7.610351 -1.670844
## sample estimates:
##  mean in group No mean in group Yes
##          70.13004          74.77064
```

The small p-value implies that we reject the null hypothesis of equal means for the blood pressure.

We have already looked at the assumption of normality for diabetic women. Let us look at the non-diabetics

```
qqnorm(bp.no); qqline(bp.no)
```




```
shapiro.test(bp.no)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  bp.no
## W = 0.9863, p-value = 0.03045
```

In this case the p-value for the Shapiro-Wilk test is small and at the 5% level we would reject the hypothesis of normality for the data. However, the test is on the mean values, and even if the population is not normal, by the Central Limit Theorem we know that for large sample the sampling distribution of the mean is approximately Gaussian, as is the case here since the sample size is 223.

(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).

Wilcoxon's test is appropriate here. For (d):

```
wilcox.test(Pima.te$bp, mu = 70)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Pima.te$bp
## V = 26999, p-value = 0.01726
## alternative hypothesis: true location is not equal to 70
```

The p-value in this case is 0.0173, which is higher than before. At the 5% or 2% level we would reach the same conclusion as in (d), but at the 1% level it would be different.

For (e):

```
wilcox.test(bp ~ type, data = Pima.te)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bp by type
## W = 9485.5, p-value = 0.001144
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is small and the conclusion coincides with the t-test.