Exercise 1 - solution

Paul Blanche

Question 1

We load the data and have a look a the first lines.

```r
load(url("http://paulblanche.com/files/SCD.rda"))
d <- SCD # shorter name, just for convenience
head(d)
```

```
##   age sex SCD Pdias Psys height weight pulse  HCT Creat  Hb
## 1  26   2   1    62  120    164   57.0    68 21.5    37 7.7
## 2  33   2   1    60  115    165   58.0    64 21.6    55 7.8
## 3  43   2   1    66  117    152   57.2    67 29.3    68 9.2
## 4  23   1   1    67  129    177   61.0    62 25.0    66 8.9
## 5  37   1   1    47  106    167   50.0    65 23.1    66 8.0
## 6  46   2   1    65  134    163   68.0    67 21.0    69 7.7
```

We get summary statistics for all variables.

```r
summary(d)
```

```
##       age             sex             SCD          Pdias            Psys      
##  Min.   :16.00   Min.   :1.000   Min.   :0.0   Min.   :43.00   Min.   : 97.0  
##  1st Qu.:23.00   1st Qu.:1.000   1st Qu.:0.0   1st Qu.:59.00   1st Qu.:113.0  
##  Median :29.00   Median :2.000   Median :0.5   Median :64.50   Median :121.0  
##  Mean   :30.85   Mean   :1.534   Mean   :0.5   Mean   :65.95   Mean   :122.6  
##  3rd Qu.:37.00   3rd Qu.:2.000   3rd Qu.:1.0   3rd Qu.:71.00   3rd Qu.:131.0  
##  Max.   :66.00   Max.   :2.000   Max.   :1.0   Max.   :95.00   Max.   :160.0  
##                                                                               
##      height          weight            pulse              HCT       
##  Min.   :145.0   Min.   : 41.00   Min.   : 46.00   Min.   :14.10  
##  1st Qu.:162.0   1st Qu.: 52.75   1st Qu.: 64.00   1st Qu.:24.18  
##  Median :169.0   Median : 60.00   Median : 72.00   Median :32.00  
##  Mean   :169.0   Mean   : 62.14   Mean   : 73.19   Mean   :32.58  
##  3rd Qu.:176.2   3rd Qu.: 68.25   3rd Qu.: 80.25   3rd Qu.:39.65  
##  Max.   :193.0   Max.   :123.00   Max.   :114.00   Max.   :88.70  
##                                                    NA's   :2      
##      Creat              Hb        
##  Min.   : 16.00   Min.   : 4.90
##  1st Qu.: 56.00   1st Qu.: 8.60
##  Median : 68.00   Median :10.60
##  Mean   : 69.38   Mean   :10.90
##  3rd Qu.: 81.00   3rd Qu.:13.07
##  Max.   :161.00   Max.   :17.20
##  NA's   :3        NA's   :2
```

The variables HCT, Creat and Hb have missing values.

Question 2

```
table(d$sex)
```

```
##
##  1  2
## 82 94
```

There are 82 women and 94 men.

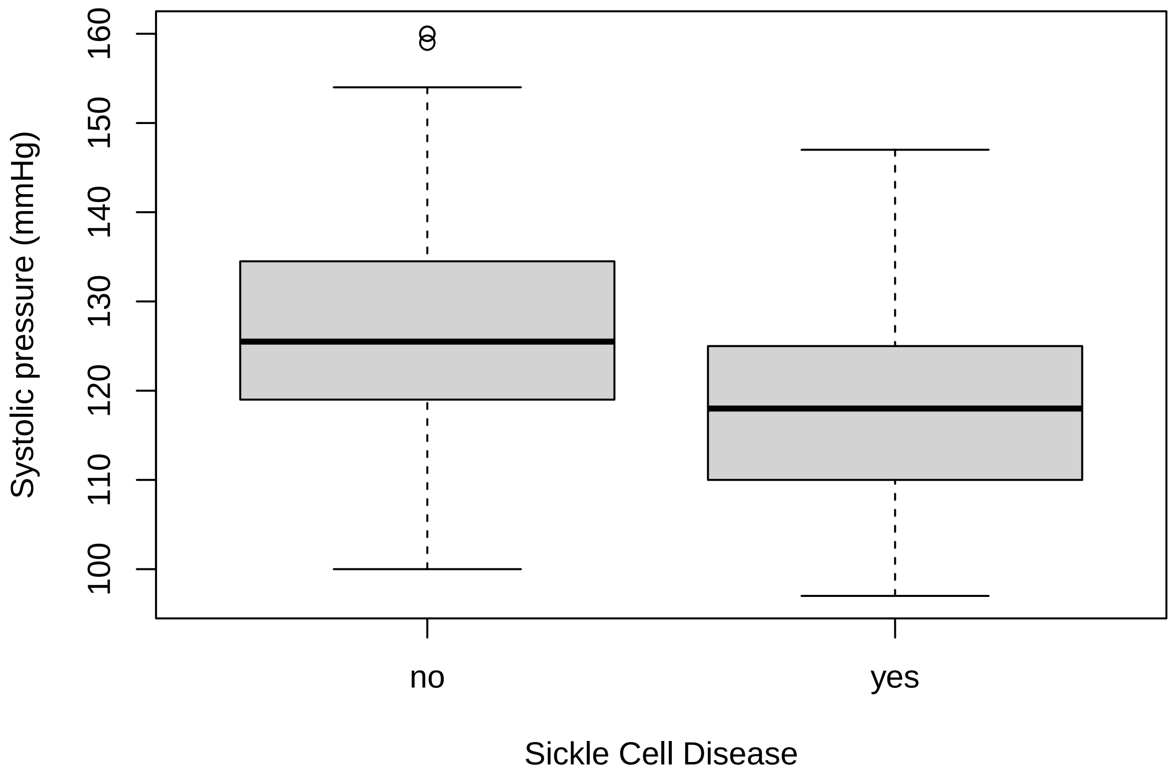
```
table(d$SCD)
```

```
##
##  0  1
## 88 88
```

There are 88 subjects with and 88 without Sickle Cell Disease.

Question 3

```
boxplot(d$Psys~factor(d$SCD,levels=c(0,1),labels=c("no","yes")),
                      xlab="Sickle Cell Disease",
                      ylab="Systolic pressure (mmHg)")
```



Interpretation:

- Box:
- middle line: median Q2 (50%)
- bottom: first quartile Q1 (25%)
- top: third quartile Q3 (75%)
- Whiskers: size is 1.5 times the height of the box, but not exceeding minimum and maximum.
- Dots: observations beyond whiskers

The overall interpretation seems to make sense: lower values are observed for subjects with SCD (median for SCD patients is approximately Q1 for subjects without SCD).

Note: some **experienced** students might prefer to use an alternative code, using the packages **dplyr** and **ggplot2**. See appendix for such a code.

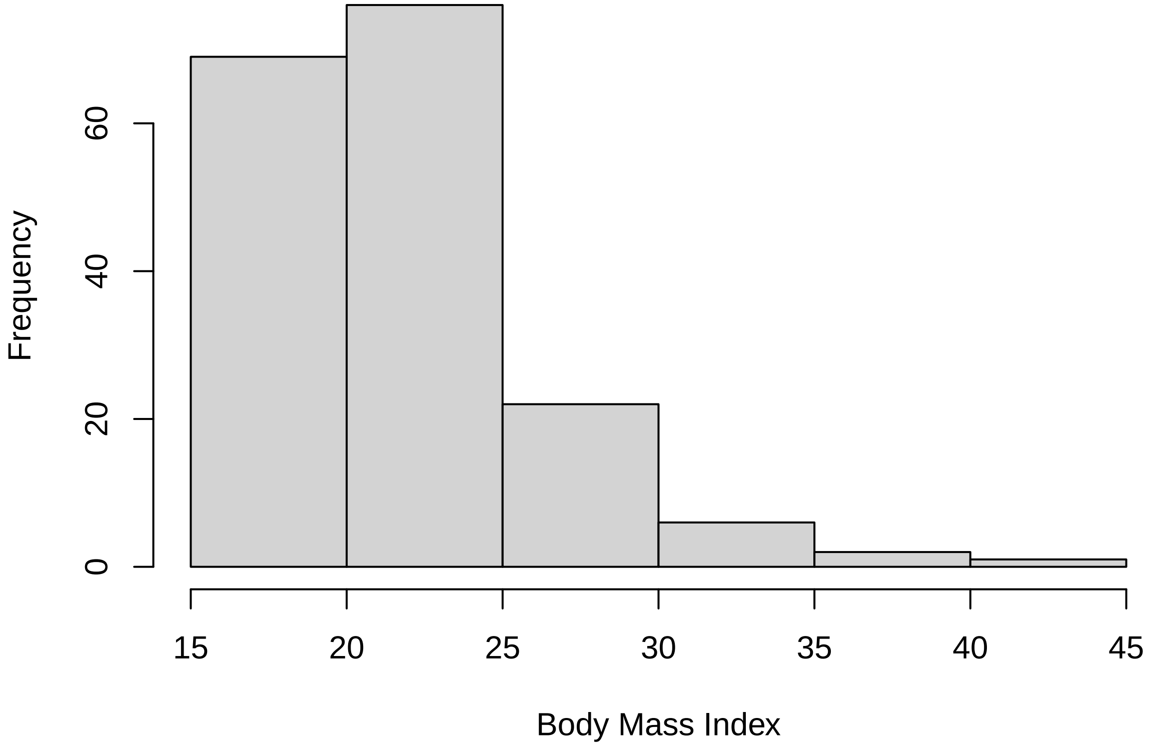
Question 4

First we create the BMI variable and add it to the data.

```
d$BMI <- d$weight/(d$height/100)^2
```

Note that we divide the height by 100 because we want the height unit to be meters, not centimeters.

```
hist(d$BMI,xlab="Body Mass Index",main="")
```



BMI does not look normally distributed because the distribution does not seem symmetric around the mean/median.

Mean of BMI:

```
mean(d$BMI)
```

```
## [1] 21.75127
```

Sdandard deviation of BMI:

```
sd(d$BMI)
```

```
## [1] 4.207904
```

Minimum of BMI:

```
min(d$BMI)
```

```
## [1] 16.10588
```

Maximum of BMI:

```
max(d$BMI)
```

```
## [1] 41.57653
```

First and third quartiles of BMI (25% and 75%):

```
quantile(d$BMI)
```

```
##        0%      25%      50%      75%     100%
## 16.10588 18.90320 20.79603 23.88694 41.57653
```

or:

```
quantile(d$BMI, 0.25)
```

```
##     25%
## 18.9032
```

```
quantile(d$BMI, 0.75)
```

```
##      75%
## 23.88694
```

Median of BMI has already been computed (50% above). Alternatively:

```
median(d$BMI)
```

```
## [1] 20.79603
```

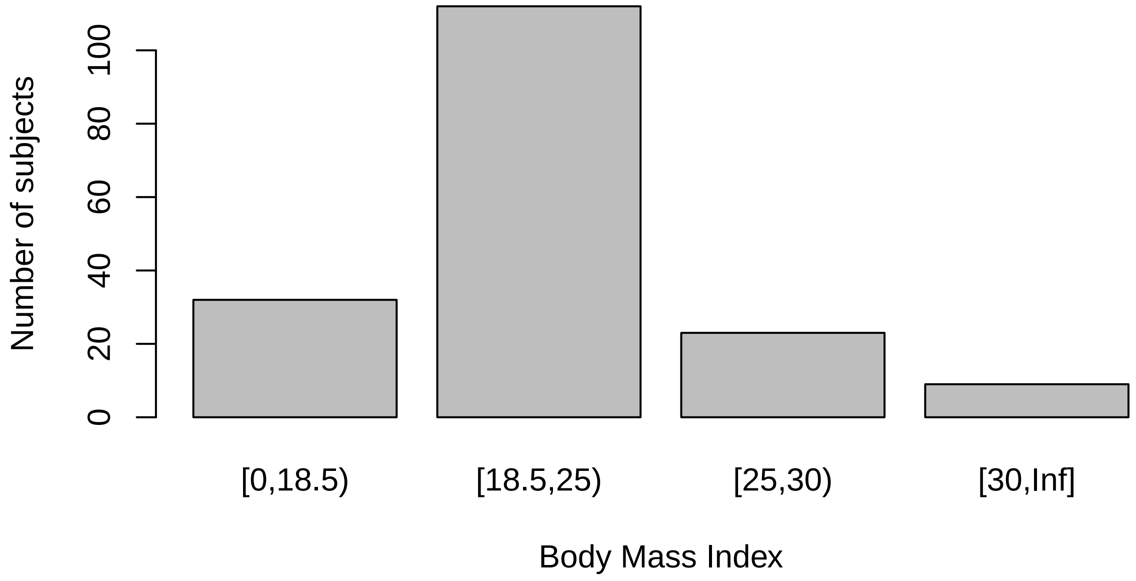
Because the variable is far from being normally distributed, it is generally more informative to report the median and the first and third quartiles rather than mean and sd. We now compute the frequencies for each BMI group.

```
d$BMIgroup <- cut(d$BMI,
                  breaks=c(0,18.5,25,30,Inf),
                  include.lowest=TRUE,right=FALSE)
table(d$BMIgroup)
```

```
##
##  [0,18.5) [18.5,25)   [25,30)  [30,Inf]
##        32       112        23         9
```

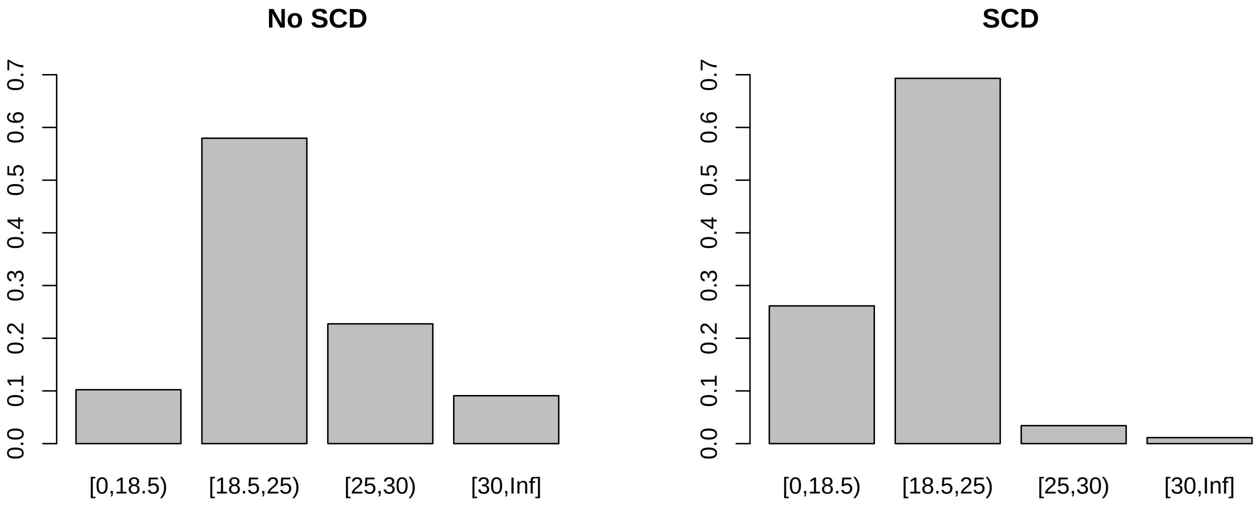
We observe 32 subjects “underweight”, 112 “normal”, 23 “overweight” and 9 “obese”. Graphically:

```
barplot(table(d$BMIgroup),xlab="Body Mass Index",ylab="Number of subjects")
```



We now do the same, but separately for subjects with and without SCD. However, we now report proportions instead of counts, to facilitate the comparison between the two groups. We make sure that the y-axis is the same for both plots to facilitate the comparison.

```
par(mfrow=c(1,2))
barplot(table(d$BMIgroup[d$SCD==0])/sum(table(d$BMIgroup[d$SCD==0])),
        main="No SCD",
        ylim=c(0,0.7))
barplot(table(d$BMIgroup[d$SCD==1])/sum(table(d$BMIgroup[d$SCD==1])),
        main="SCD",
        ylim=c(0,0.7))
```



We observe somewhat lower BMI for subjects with SCD (less overweight and less obese subjects).

Question 5

We compute the mean and sd for subjects with (SCD=1) and without (SCD=0) SCD.

```
mean(d$Pdias[d$SCD==1])
```

```
## [1] 61.80682
```

```
mean(d$Pdias[d$SCD==0])
```

```
## [1] 70.09091
```

```
sd(d$Pdias[d$SCD==1])
```

```
## [1] 6.969322
```

```
sd(d$Pdias[d$SCD==0])
```

```
## [1] 10.82848
```

Next, we use the t.test function to compute 95% confidence interval for the mean diastolic pressure in the two populations of subjects with and without SCD.

```
t.test(d$Pdias[d$SCD==1])
```

```
##
##  One Sample t-test
##
## data:  d$Pdias[d$SCD == 1]
## t = 83.193, df = 87, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  60.33016 63.28348
## sample estimates:
## mean of x
##  61.80682
```

```
t.test(d$Pdias[d$SCD==0])
```

```
##
##  One Sample t-test
##
## data:  d$Pdias[d$SCD == 0]
## t = 60.721, df = 87, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  67.79657 72.38525
## sample estimates:
## mean of x
##  70.09091
```

The confidence intervals are [60.3;63.3] and [67.8;72.4] for subjects with and without SCD, respectively. They do not overlap, hence we can conclude to a significant difference in mean diastolic pressure between the two groups (at the usual 5% level for the type-I error control).

We now compute the 95% **prediction** interval for the diastolic pressure in the two populations of subjects with and without SCD.

```
predict(lm(d$Pdias[d$SCD==0]~1),interval="prediction")[1,]
```

```
## Warning in predict.lm(lm(d$Pdias[d$SCD == 0] ~ 1), interval = "prediction"): predicti
```

```
##      fit      lwr      upr
## 70.09091 48.44619 91.73563
```

```
predict(lm(d$Pdias[d$SCD==1]~1),interval="prediction")[1,]
```

```
## Warning in predict.lm(lm(d$Pdias[d$SCD == 1] ~ 1), interval = "prediction"): predicti
```

```
##      fit      lwr      upr
## 61.80682 47.87605 75.73758
```

Under the assumption that the diastolic pressure is (approximately) normally distributed in each population (which should be checked using e.g. a QQplot!) then we can expect that (close to) 95% of the subjects without SCD (and similar to that of the study) have a diastolic pressure between 48.4 and 91.7. For subjects with SCD, between 47.9 and 75.7.

Question 6

```
t.test(d$Psys[d$SCD==1])
```

```
##
##  One Sample t-test
##
## data:  d$Psys[d$SCD == 1]
## t = 95.985, df = 87, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  115.8792 120.7799
## sample estimates:
## mean of x
##  118.3295
```

```
t.test(d$Psys[d$SCD==0])
```

```
##
##  One Sample t-test
##
## data:  d$Psys[d$SCD == 0]
## t = 89.55, df = 87, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  124.0701 129.7027
## sample estimates:
## mean of x
##  126.8864
```

The confidence intervals are [115.9;120.8] and [124.1;129.7]. They do not overlap, hence a significant difference.

```
t.test(d$Psys[d$SCD==1 & d$sex==2])
```

```
##
##  One Sample t-test
##
## data:  d$Psys[d$SCD == 1 & d$sex == 2]
## t = 78.014, df = 46, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  113.8983 119.9315
## sample estimates:
## mean of x
##  116.9149
```

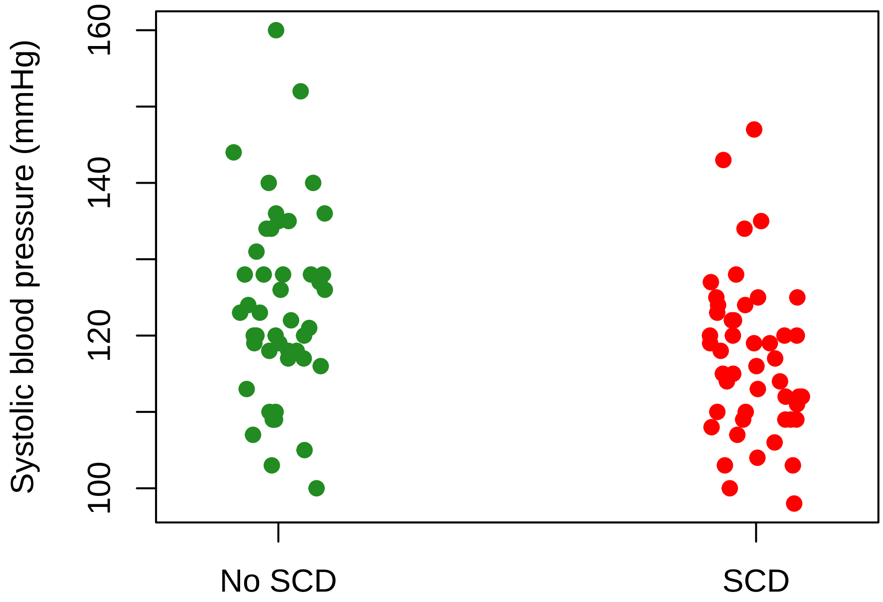
```
t.test(d$Psys[d$SCD==0 & d$sex==2])
```

```
##
##  One Sample t-test
##
## data:  d$Psys[d$SCD == 0 & d$sex == 2]
## t = 68.111, df = 46, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  119.8812 127.1826
## sample estimates:
## mean of x
##  123.5319
```

The confidence intervals are [113.9;119.9] and [119.9;127.2]. They do overlap (although very little), hence **we cannot rigorously say whether there is a significant difference**. We need to perform a two-sample t-test to rigorously conclude (see course Day 2).

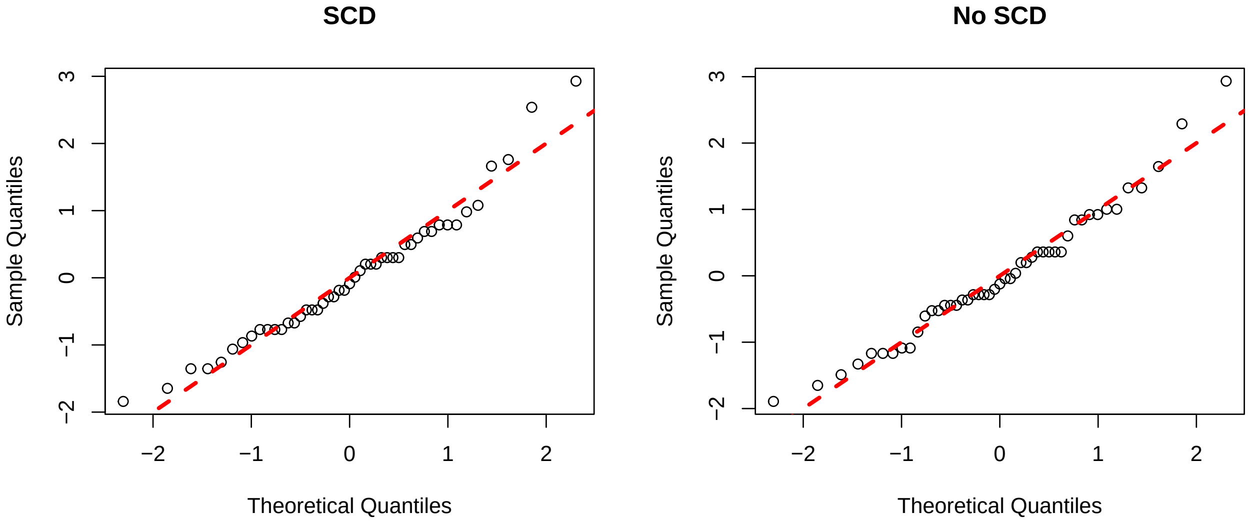
We now produce a dotplot to display the individual observations.

```
stripchart(d$Psys[d$sex==2]~factor(d$SCD[d$sex==2],
           levels=c(0,1),labels=c("No SCD","SCD")),
           vertical=TRUE,
           method="jitter",
           xlab="",
           ylab="Systolic blood pressure (mmHg)",
           pch=19,
           col=c("forestgreen","red"))
```



Now we produce QQplots for each group.

```
par(mfrow=c(1,2))
qqnorm(scale(d$Psys[d$SCD==1 & d$sex==2]),main="SCD")
abline(0,1,col="red",lty=2,lwd=3)
qqnorm(scale(d$Psys[d$SCD==0 & d$sex==2]),main="No SCD")
abline(0,1,col="red",lty=2,lwd=3)
```



We count the number of women with and without SCD.

```r
table(d$SCD[d$sex==2])
```

```
##
##  0  1
## 47 47
```

Because the QQplot looks good (observations close to the diagonal, no major deviation) and because the sample size of each group is "large enough" (n=47) we can trust the computation of the 95% confidence intervals. Hence, we can interpret the results quite confidently, without limitations about the appropriateness of the statistical method.

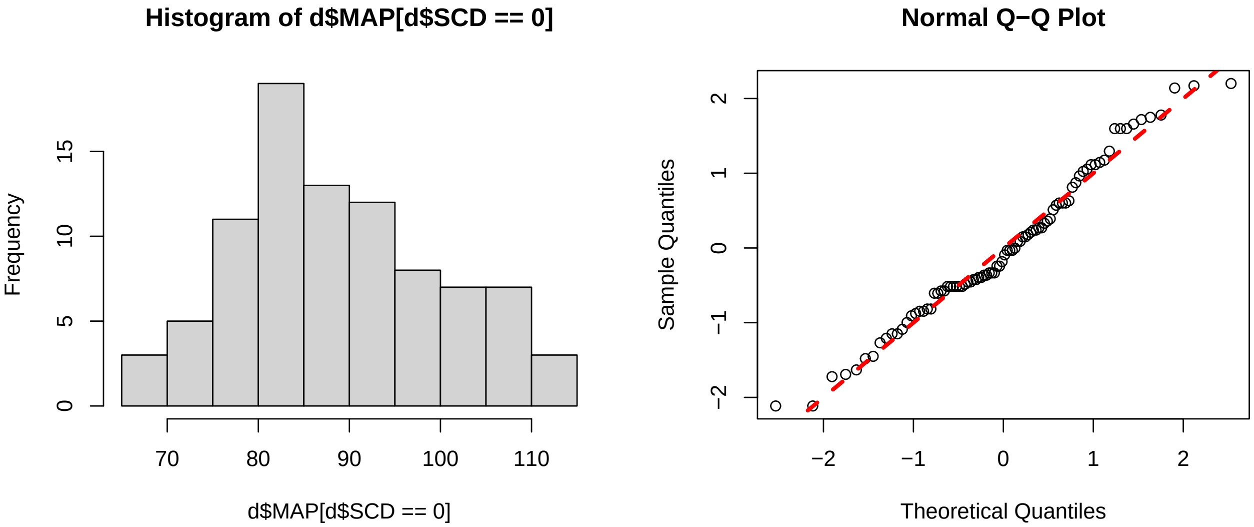
Question 7

We first create (and add to the data) a new variable "MAP" to define the Mean Arterial Pressure.

```r
d$MAP <- d$Pdias + (1/3)*(d$Psys-d$Pdias)
```

We make a histogram and a QQplot to visually assess whether the distribution of the Mean Arterial Pressure looks normally distributed, among subjects without SCD

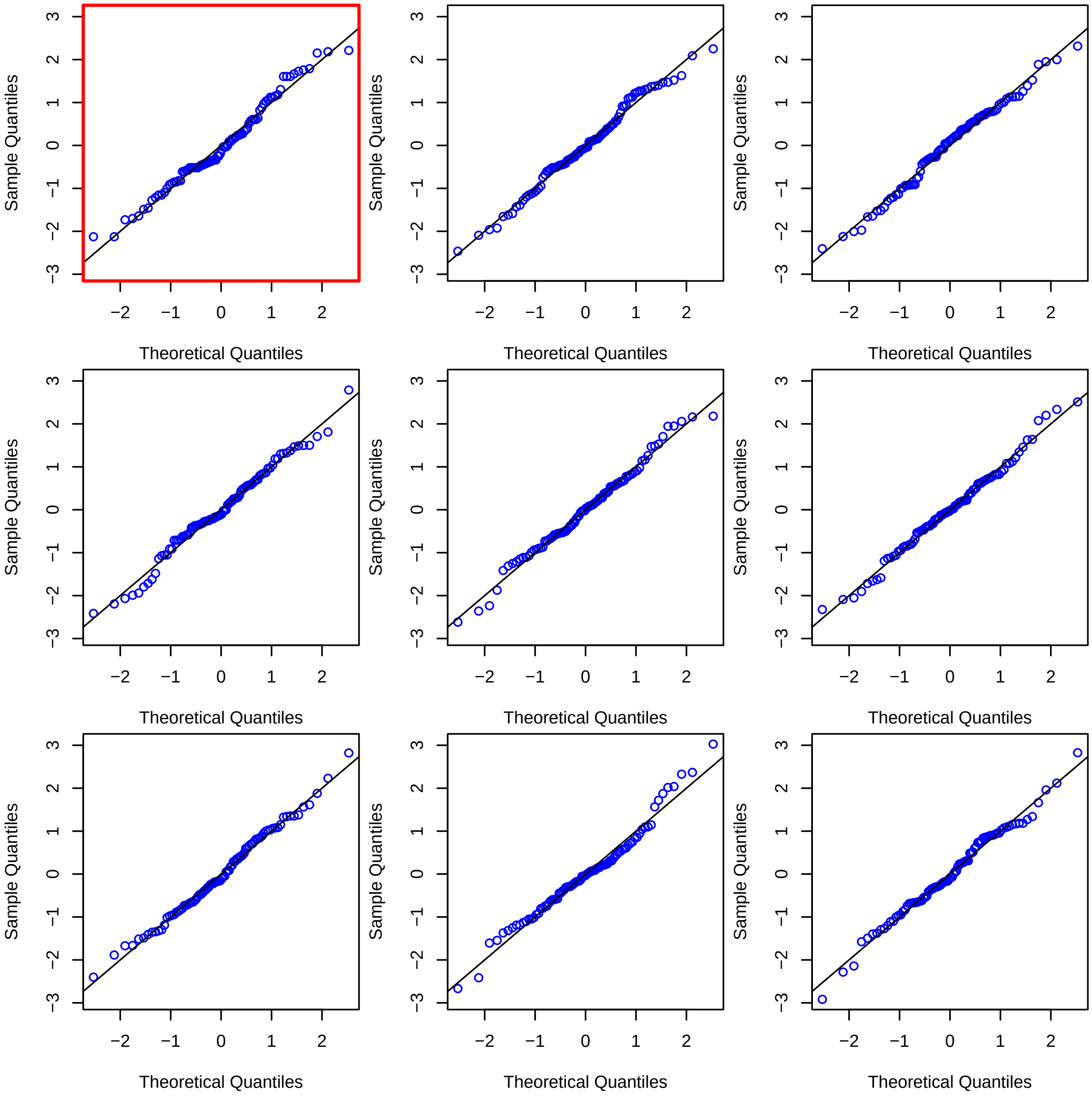
```r
par(mfrow=c(1,2))
hist(d$MAP[d$SCD==0])
qqnorm(scale(d$MAP[d$SCD==0]))
abline(0,1,col="red",lty=2,lwd=3)
```



It looks "fine", but it is, as always, difficult to say without "references" to compare to. Small sample random variation is difficult to understand. Hence the "Wally plot" can help.

```r
library(MESS)
lm0 <- lm(MAP~1,data=d[d$SCD==0,])
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...) ; abline(a=0, b=1) }
```

```
wallyplot(lm0, FUN=qqnorm.wally, main="",hide=FALSE,col="blue")
```



The “Wally plot” helps us to conclude that the QQplot of our data does not look bad: it does not look very different from the QQplot of random samples (of same sample size) drawn from a normal distribution.

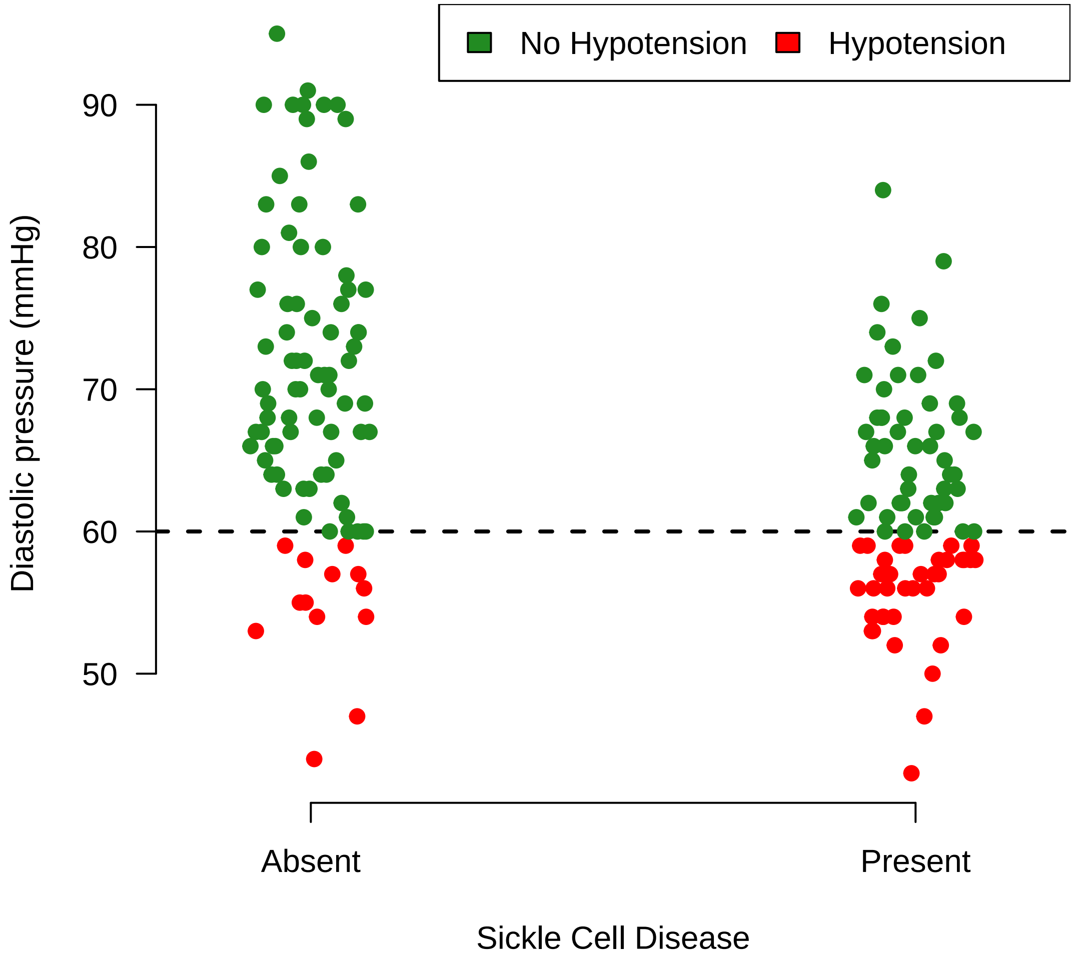
```
predict(lm(d$MAP[d$SCD==0]~1),interval="prediction")[1,]
```

```
## Warning in predict.lm(lm(d$MAP[d$SCD == 0] ~ 1), interval = "prediction"): prediction

##       fit       lwr       upr
##  89.02273  66.94477 111.10069
```

We estimate the “normal range” of MAP for patients without SCD as [66.9;111.1], which is very close to what is reported on wikipedia.

Additional challenge

```r
# First, create a binary variable for hypotension
d$Hypo <- cut(d$Pdias,breaks=c(0,60,Inf),
              include.lowest=TRUE,right=FALSE)
# create two data sets, with subjects with and without hypotension
dHypo <- d[d$Hypo=="[0,60)",]
dNoHypo <- d[d$Hypo=="[60,Inf]",]
# set seed for reproducibility (jitter)
set.seed(123)
# First plot observations with hypotension
stripchart(dHypo$Pdias~dHypo$SCD,
           vertical=TRUE,method="jitter",
           col="red",
           pch=19,
           ylim=c(min(d$Pdias),max(d$Pdias)),
           xlab="Sickle Cell Disease",
           ylab="Diastolic pressure (mmHg)",
           axes=FALSE)
# add axes
axis(1,at=1:2,c("Absent","Present"))
axis(2,las=2)
# add horizontal line at the threshold defining hypotension
abline(h=60,lwd=2,lty=2)
# Plot observations without hypotension on top
stripchart(dNoHypo$Pdias~dNoHypo$SCD,
           vertical=TRUE,
           method="jitter",
           col="forestgreen",pch=19,add=TRUE)
# Add a legend
legend("topright",
       fill=c("forestgreen","red"),
       legend=c("No Hypotension","Hypotension"),
       ncol=2)
```

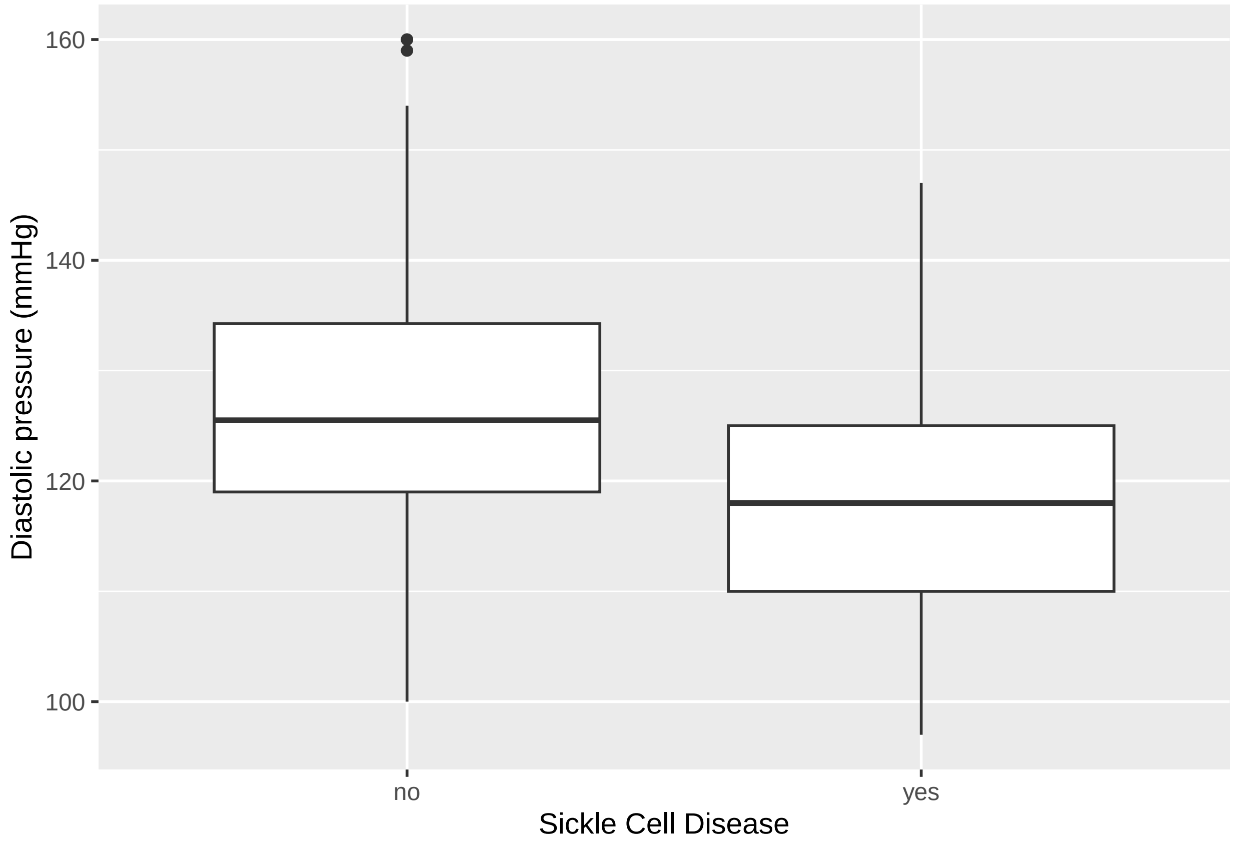


Note: some **experienced** students might prefer to use an alternative code, using the packages **dplyr** and **ggplot2**. See appendix for such a code.

Appendix

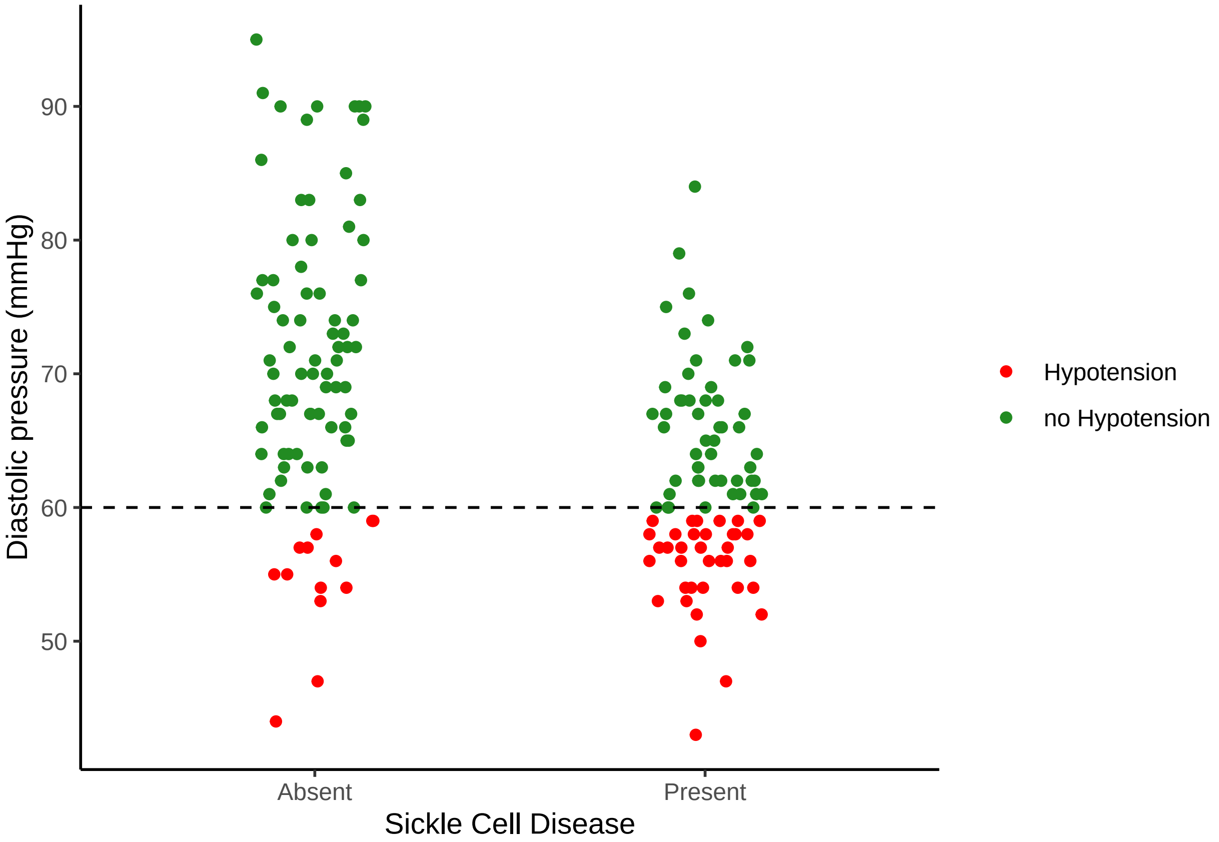
Alternative code for Question 3

```
library(dplyr)
library(ggplot2)

d %>%
        mutate(SCD = factor(SCD, levels=c(0,1), labels=c("no","yes"))) %>%
        ggplot(., aes(y = Psys, x = SCD)) +
                geom_boxplot() +
                xlab("Sickle Cell Disease") +
                ylab("Diastolic pressure (mmHg)")
```



Alternative code for the additional challenge

```
set.seed(123)

d %>%
mutate(Hypo = case_when(Pdias < 60 ~ "Hypotension",
                        Pdias >= 60 ~ "no Hypotension"),
       SCD = case_when(SCD == 0 ~ "Absent",
                       SCD == 1 ~ "Present")) %>%
ggplot(., aes(x=SCD, y=Pdias, color = Hypo)) +
        geom_jitter(width = 0.15, height = 0) + #only horizontal jitter
        xlab("Sickle Cell Disease") + #x-axis label
        ylab("Diastolic pressure (mmHg)") + #y-axis label
        scale_color_manual(values = c("red", "forestgreen")) + #set colour
        theme_classic() + #remove grid + gray background
        theme(legend.title = element_blank()) + #remove legend title
        geom_hline(yintercept = 60, linetype = "dashed") #horizontal line
```


Diastolic pressure (mmHg)
90
80
70
60
50
Absent
Present
Sickle Cell Disease
Hypotension
no Hypotension