Ngoc Ha

STAT 453

Lab 4 ¶

In [2]:
```
library(readxl)
library(fitdistrplus)
library(boot)
```

1. pgatour2006.xlsx analysis

(1a) Summarize

In [48]:
```
pga <- read_excel('pgatour2006.xlsx')
```

In [4]:
```
str(pga)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':	196 obs. of  11 variables:
 $ Name              : chr  "Aaron Baddeley" "Adam Scott" "Alex Aragon" "Alex
Cejka" ...
 $ PrizeMoney        : num  60661 262045 3635 17516 16683 ...
 $ AveDrivingDistance: num  288 301 303 289 288 ...
 $ DrivingAccuracy   : num  60.7 62 51.1 66.4 63.2 ...
 $ GIR               : num  58.3 69.1 59.1 67.7 64 ...
 $ PuttingAverage    : num  1.75 1.77 1.79 1.78 1.76 ...
 $ BirdieConversion  : num  31.4 30.4 29.9 29.3 29.3 ...
 $ SandSaves         : num  54.8 53.6 37.9 45.1 52.4 ...
 $ Scrambling        : num  59.4 57.9 50.8 54.8 57.1 ...
 $ BounceBack        : num  19.3 19.4 16.8 17.1 18.2 ...
 $ PuttsPerRound     : num  28 29.3 29.2 29.5 28.9 ...
```

10% trimmed mean

In [5]:
```
sapply(pga[,2:11], mean, trim = 0.1)
```

PrizeMoney	40027.2151898734
AveDrivingDistance	289.308860759494
DrivingAccuracy	63.3127848101266
GIR	65.2693670886076
PuttingAverage	1.77932911392405
BirdieConversion	29.0146202531646
SandSaves	48.9799367088608
Scrambling	57.5008860759494
BounceBack	19.5676582278481
PuttsPerRound	29.190253164557

Standard deviation

In [6]:
```
sapply(pga[,2:11], sd)
```

PrizeMoney	63902.9534175403
AveDrivingDistance	8.7305092931058
DrivingAccuracy	5.41302275758653
GIR	2.7223638625283
PuttingAverage	0.024728132964627
BirdieConversion	2.20655581118645
SandSaves	5.82831251040945
Scrambling	3.16225742012002
BounceBack	2.80611274288793
PuttsPerRound	0.441702272391593

Summary

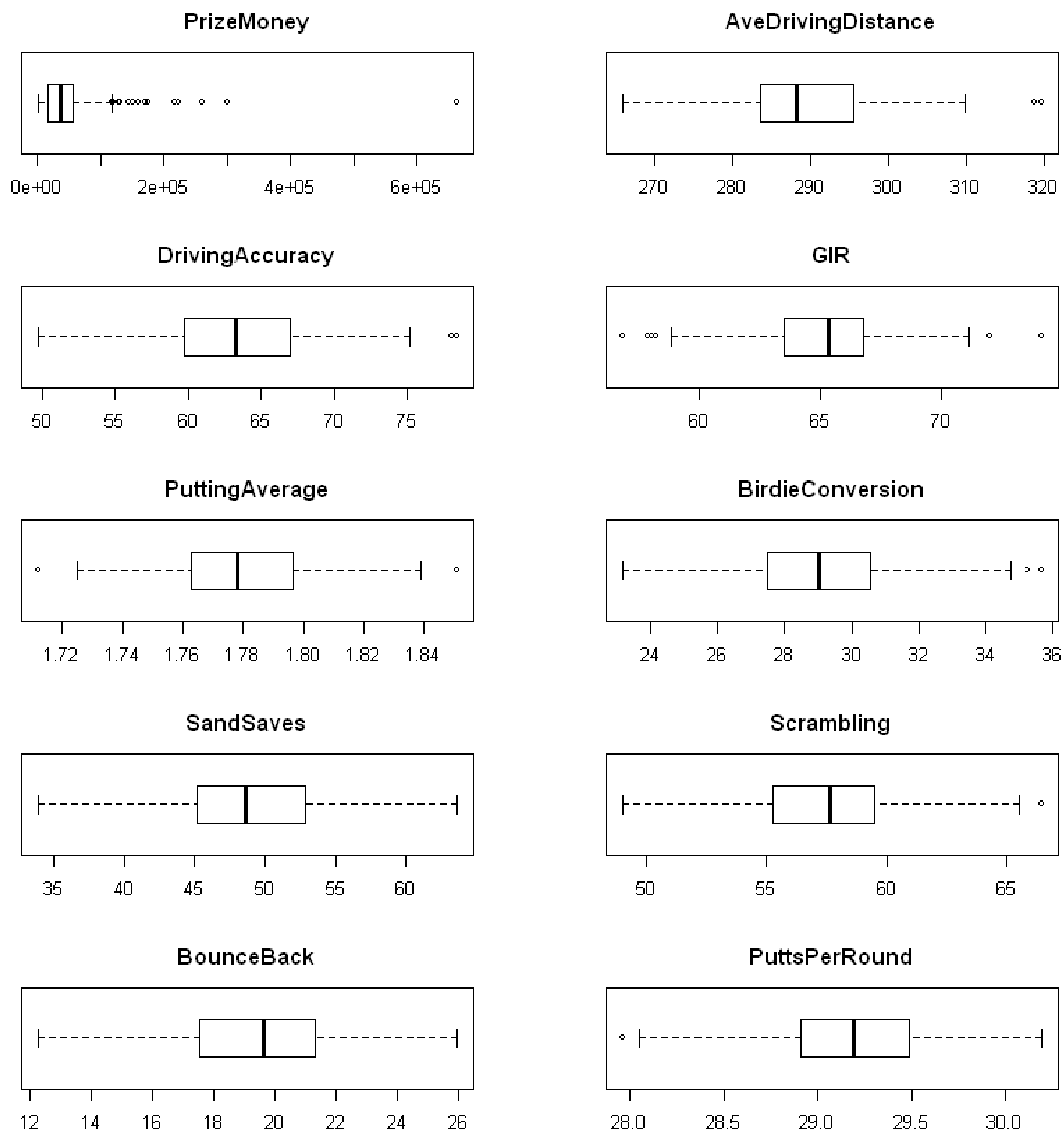
In [7]:
```
summary(pga[,2:11])
```

```
   PrizeMoney     AveDrivingDistance DrivingAccuracy      GIR       
 Min.   :  2240   Min.   :265.9      Min.   :49.75   Min.   :56.87  
 1st Qu.: 17369   1st Qu.:283.6      1st Qu.:59.76   1st Qu.:63.52  
 Median : 36645   Median :288.2      Median :63.24   Median :65.36  
 Mean   : 50891   Mean   :289.5      Mean   :63.38   Mean   :65.19  
 3rd Qu.: 57915   3rd Qu.:295.5      3rd Qu.:66.97   3rd Qu.:66.77  
 Max.   :662771   Max.   :319.6      Max.   :78.43   Max.   :74.15  
 PuttingAverage  BirdieConversion   SandSaves       Scrambling   
 Min.   :1.712   Min.   :23.17    Min.   :33.91   Min.   :49.02  
 1st Qu.:1.763   1st Qu.:27.51    1st Qu.:45.13   1st Qu.:55.26  
 Median :1.778   Median :29.01    Median :48.66   Median :57.65  
 Mean   :1.780   Mean   :28.98    Mean   :48.97   Mean   :57.49  
 3rd Qu.:1.796   3rd Qu.:30.55    3rd Qu.:52.87   3rd Qu.:59.46  
 Max.   :1.851   Max.   :35.66    Max.   :63.64   Max.   :66.45  
   BounceBack    PuttsPerRound  
 Min.   :12.29   Min.   :27.96  
 1st Qu.:17.56   1st Qu.:28.91  
 Median :19.62   Median :29.19  
 Mean   :19.60   Mean   :29.20  
 3rd Qu.:21.31   3rd Qu.:29.48  
 Max.   :25.93   Max.   :30.19  
```

(1b)

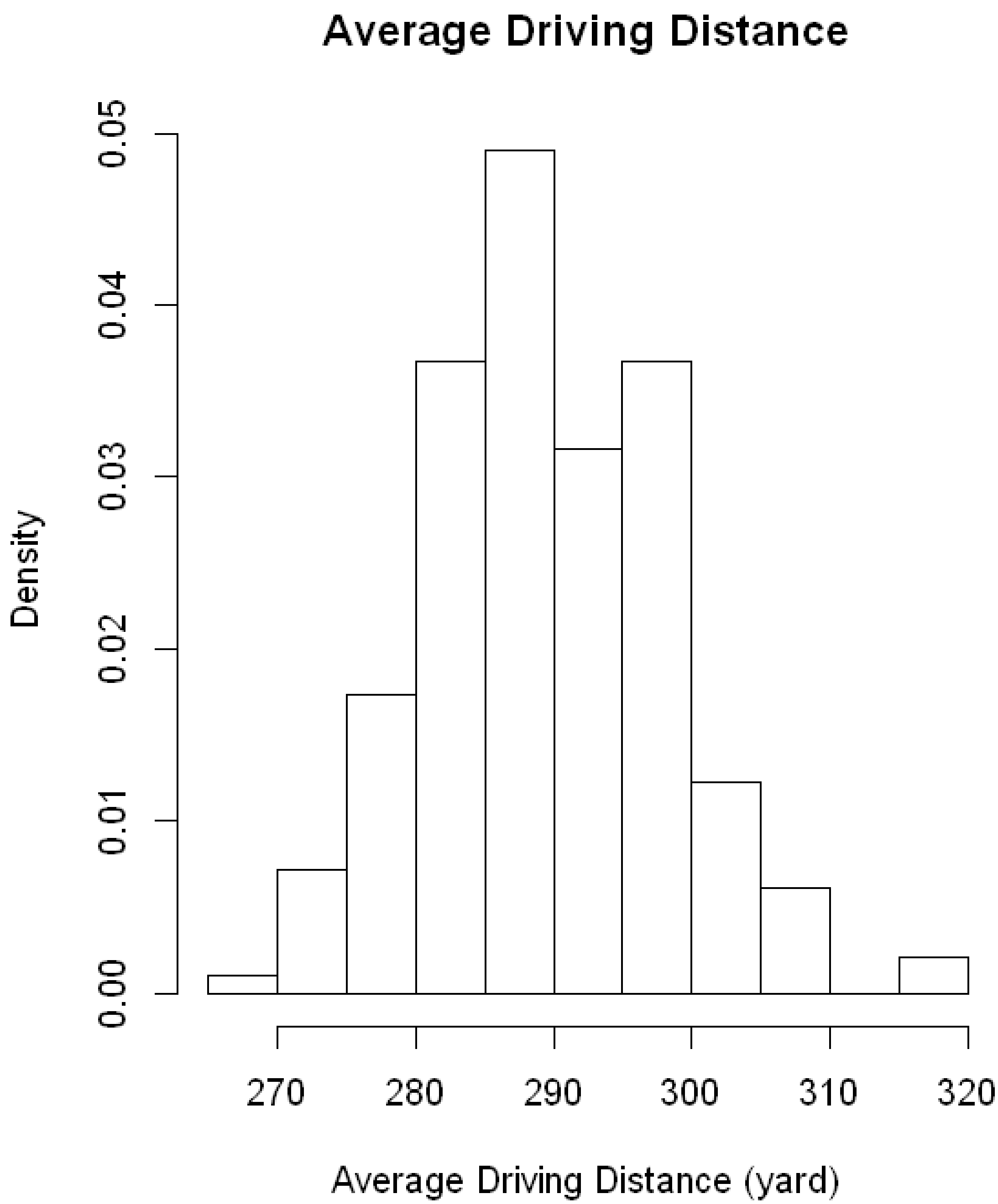
In [8]:

```
j <- 2
par(mfrow = c(5,2), mar = c(3,3,3,3))
for (i in pga[,2:11]) {
    boxplot(i, horizontal = T, main = names(pga)[j])
    j <- j+1
}
```



(1c) Density histogram of AveDrivingDistance

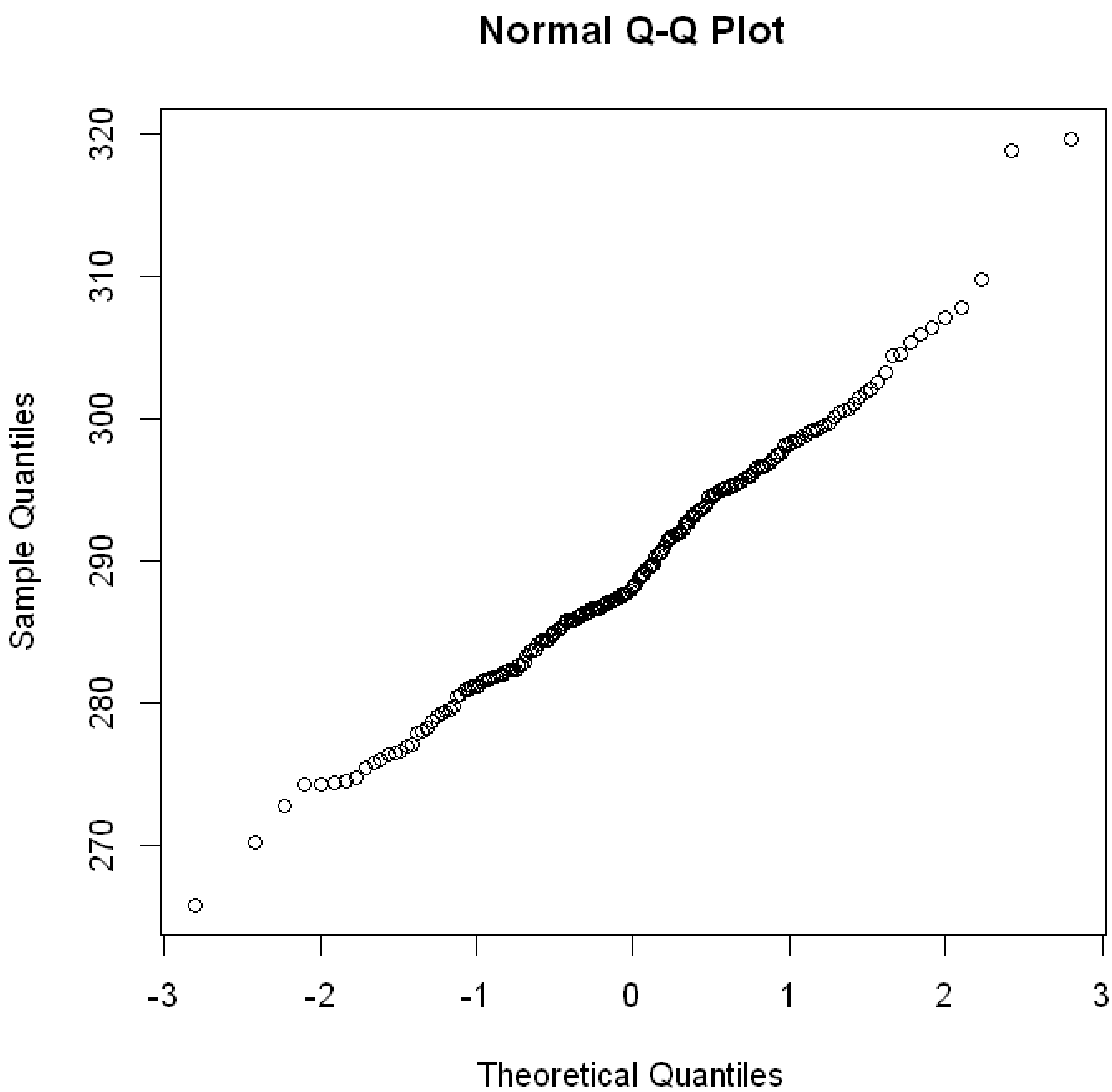
In [111]:
```
hist(pga$AveDrivingDistance, prob = T, main = 'Average Driving Distance', xlab
 = 'Average Driving Distance (yard)')
options(repr.plot.width=5, repr.plot.height=6)
```



(1d) Normal plot of AveDrivingDistance

In [110]:

```
qqnorm(pga$AveDrivingDistance)
options(repr.plot.width=5, repr.plot.height=6)
```



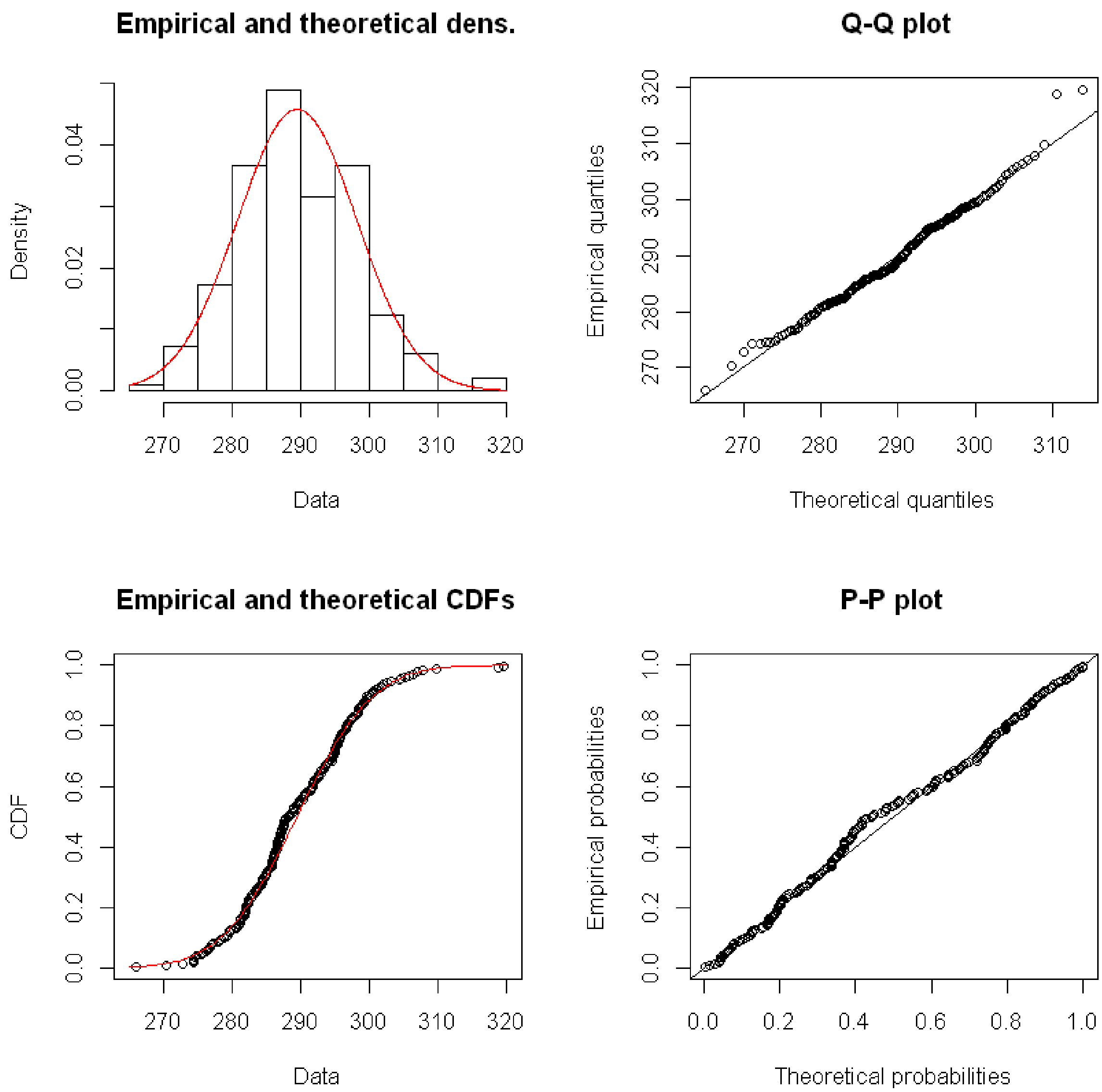
(1e) Fit a model

Normal plot

In [11]:

```
fitPgaNorm <- fitdist(pga$AveDrivingDistance, "norm")
```

In [12]:
```
plot(fitPgaNorm)
```



In [13]:
```
fitPgaNorm
```

```
Fitting of the distribution ' norm ' by maximum likelihood
Parameters:
       estimate Std. Error
mean 289.498469  0.6220149
sd     8.708209  0.4398310
```

(1f) 75th percentile of AveDrivingDistance from quantile command

In [14]:
```
quantile(pga$AveDrivingDistance, 0.75)
```

75%: 295.525

(1g) 75th percentile of AveDrivingDistance from fitted model

In [15]:
```
qnorm(0.75, 289.498, 8.708)
```

295.371456744707

(1h) Bootstrap the 75th percentile

Bootstrap

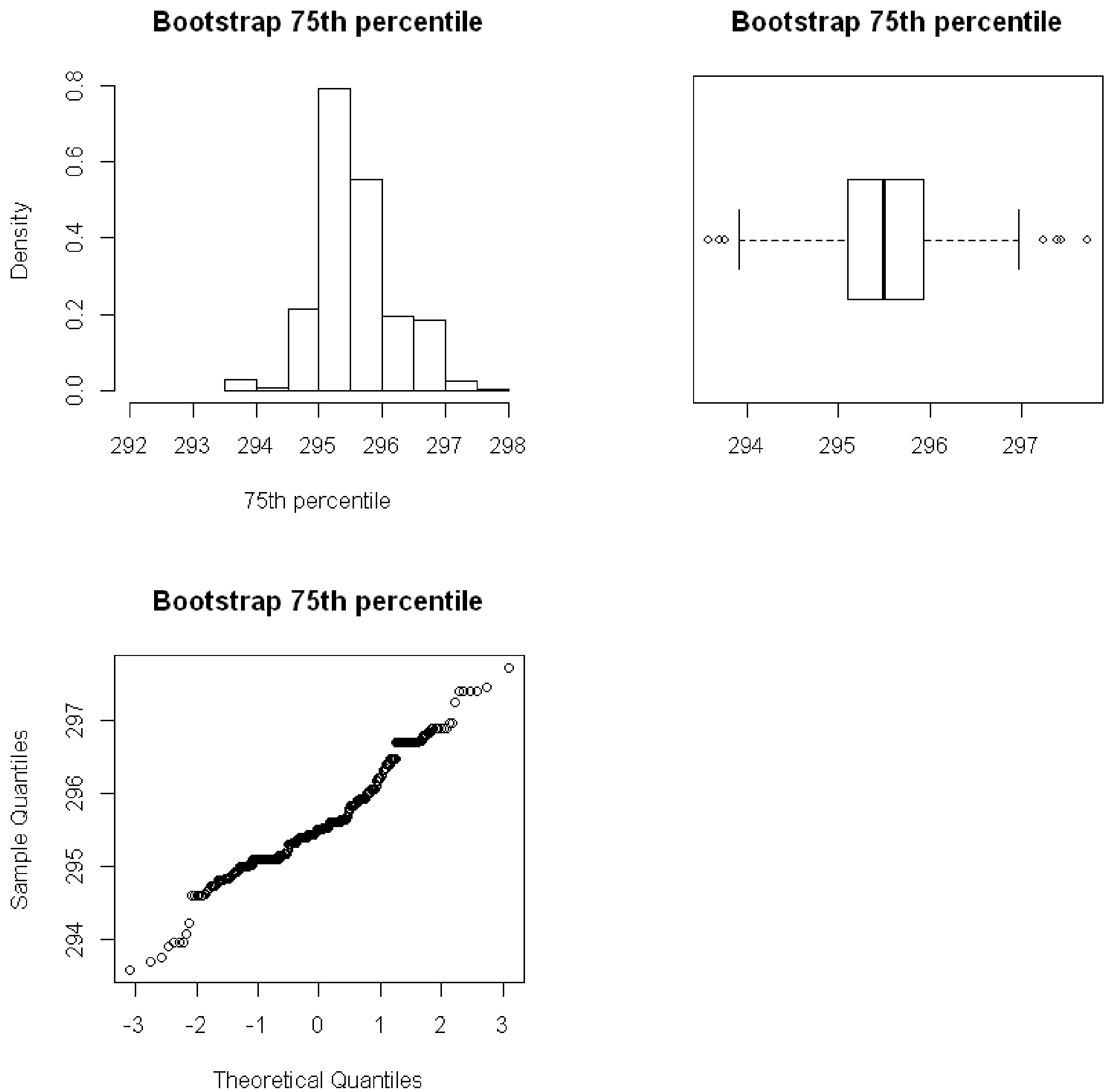
In [49]:
```
thirdQuartile <- function(d, i){
    return(quantile(d[i], 0.75))
}
thirdQuartileBoot <- boot(data = pga$AveDrivingDistance, statistic = thirdQuar
tile, R = 500)
thirdQuartileBoot
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = pga$AveDrivingDistance, statistic = thirdQuartile, 
    R = 500)


Bootstrap Statistics :
    original  bias    std. error
t1*  295.525  0.0622   0.6294172
```

Plots of sampling distribution

In [50]:

```
par(mfrow = c(2,2))
hist(thirdQuartileBoot$t, freq = F, main = "Bootstrap 75th percentile", xlab = "75th percentile", xlim = c(292,298))
boxplot(thirdQuartileBoot$t, horizontal = T, main = "Bootstrap 75th percentile")
qqnorm(thirdQuartileBoot$t, main = "Bootstrap 75th percentile")
options(repr.plot.width=7, repr.plot.height=7)
```



(1i) Shape of sampling distribution of sample 75th percentile

The sampling distribution of the sample 75th percentile resembles a normal distribution. There are a few outliers and 2 extreme outliers out of 500 data points.

(1j) 5th and 95th percentiles of sampling distribution of sample 75th percentile

In [56]:
```
quantile(thirdQuartileBoot$t, 0.05)
quantile(thirdQuartileBoot$t, 0.95)
```

5%: 294.6

95%: 296.725

2. pgatour200.xlsx analysis

(2a) Pairwise correlations

In [61]:
```
cor(pga[, 2:11])
```

	PrizeMoney	AveDrivingDistance	DrivingAccuracy	GIR	PuttingAverag
PrizeMoney	1.00000000	0.15900129	0.024677039	0.41021935	-0.3130515
AveDrivingDistance	0.15900129	1.00000000	-0.590599303	0.16460354	0.0859594
DrivingAccuracy	0.02467704	-0.59059930	1.000000000	0.41635604	-0.0255826
GIR	0.41021935	0.16460354	0.416356043	1.00000000	0.0588073
PuttingAverage	-0.31305150	0.08595947	-0.025582688	0.05880737	1.0000000
BirdieConversion	0.41342953	0.37568272	-0.252125225	0.02685014	-0.7679593
SandSaves	0.22187452	-0.23669494	0.035407734	-0.08107691	-0.2650921
Scrambling	0.28472059	-0.38033753	0.396059676	0.19435094	-0.1989427
BounceBack	0.33620030	0.23750860	0.001746659	0.29275929	-0.3185672
PuttsPerRound	-0.11249143	0.25656855	0.060313852	0.48083985	0.7916828

(2b) Create and add natural logarithm of PrizeMoney to dataframe

In [52]:
```
lnPrize <- log(pga$PrizeMoney)
pga$lnPrize <- lnPrize
```

(2c) Correlation of lnPrize with all variables except for PrizeMoney

In [65]:
```
cor(lnPrize, pga[, 3:11])
```

AveDrivingDistance	DrivingAccuracy	GIR	PuttingAverage	BirdieConversion	SandSaves	S
0.07587079	0.1816729	0.5048932	-0.4301117	0.4673991	0.2414879	

(2d) Scatterplot of lnPrize versus GIR with LOESS line. Identify outliers + summarize.

In [151]:
```
GIR <- pga$GIR
options(repr.plot.width=6, repr.plot.height=6)
scatter.smooth(GIR, lnPrize, main = "lnPrize vs GIR", xlab = "Greens In Regula
tion (%)", ylab = "Natural log of Prize Money")
```

The scatterplot demonstrates a moderate positive linear relationship between Natural log of Prize Money and Greens in Regulation (r = 0.505). There are no obvious outliers.

(2e) Bootstrap sampling distribution of correlation between lnPrize and GIR

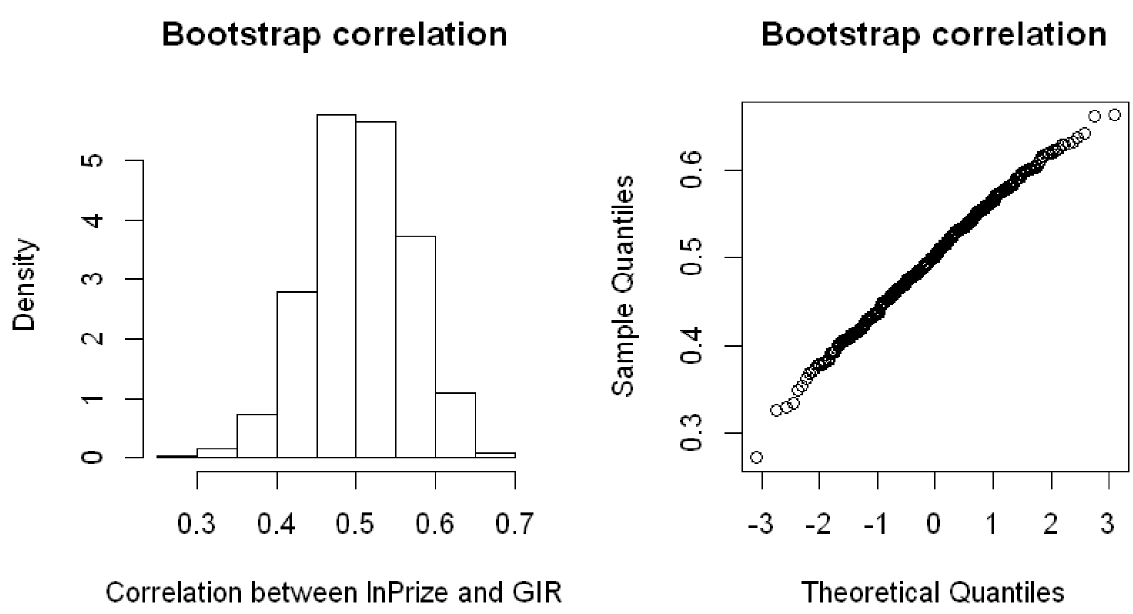
Bootstrap

In [53]:
```
corre <- function(x, i){
    return(cor(x$lnPrize[i], x$GIR[i]))
}
correBoot <- boot(data = pga, statistic = corre, R = 500)
correBoot
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = pga, statistic = corre, R = 500)


Bootstrap Statistics :
     original       bias    std. error
t1* 0.5048932 -0.001418225  0.06268673
```

Plots of sampling distribution

In [54]:
```
par(mfrow = c(1,2))
options(repr.plot.width=7, repr.plot.height=4)
hist(correBoot$t, freq = F, main = "Bootstrap correlation", xlab = "Correlatio
n between lnPrize and GIR")
qqnorm(correBoot$t, main = "Bootstrap correlation")
```



(2f) Fit the model lnPrize~GIR

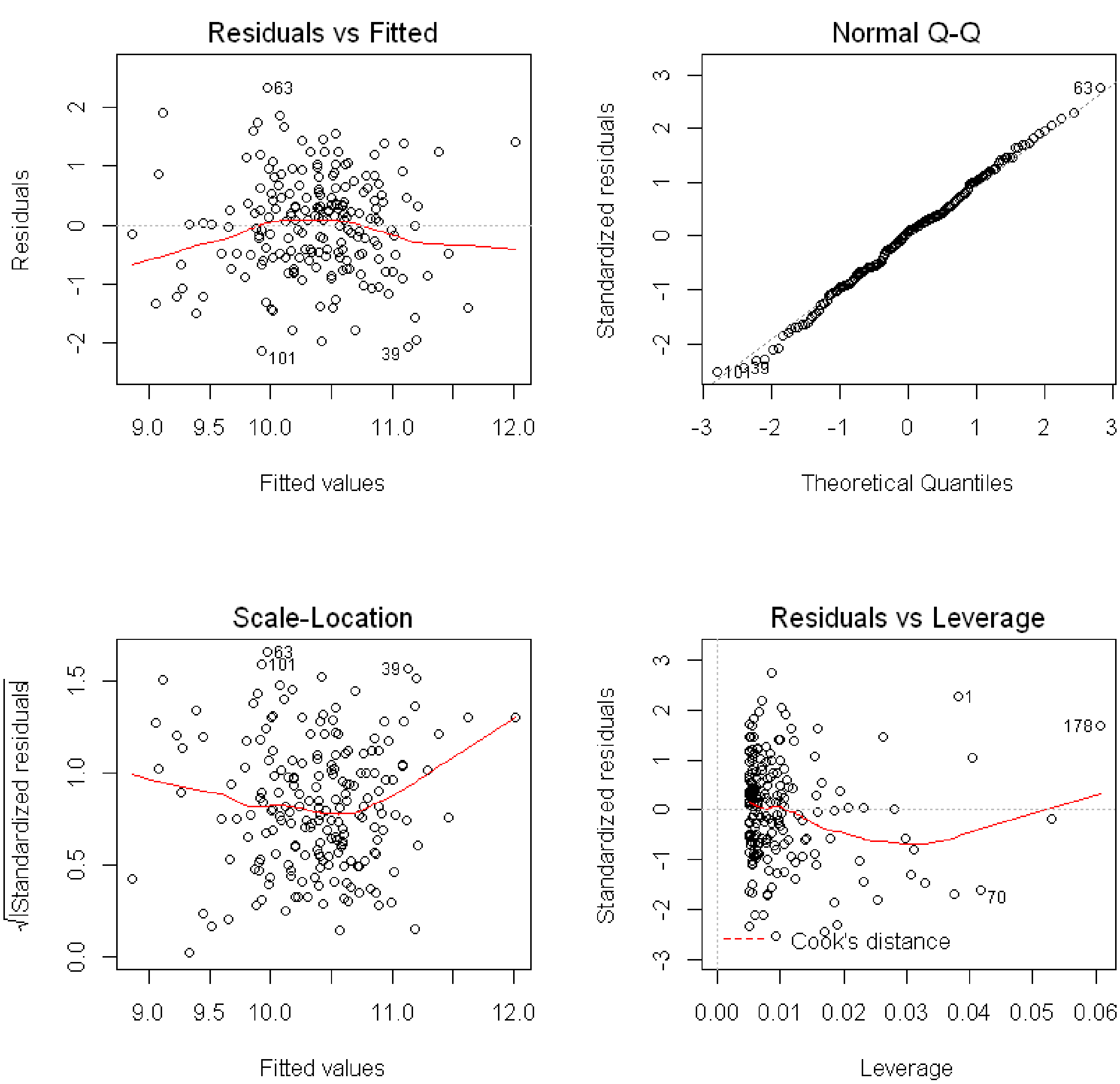
In [154]:
```
lnPrizeGIR <- lm(lnPrize~GIR)
```

In [155]:
```
summary(lnPrizeGIR)
```

```
Call:
lm(formula = lnPrize ~ GIR)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.13396 -0.55742  0.06891  0.52960  2.32133 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.47207    1.45580  -1.011    0.313    
GIR          0.18179    0.02231   8.147 4.49e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8483 on 194 degrees of freedom
Multiple R-squared:  0.2549,	Adjusted R-squared:  0.2511 
F-statistic: 66.37 on 1 and 194 DF,  p-value: 4.486e-14
```

Fitted equation: lnPrize = 0.18179 * GIR - 1.47207

In [156]:
```
par(mfrow = c(2,2))
options(repr.plot.width=7, repr.plot.height=7)
plot(lnPrizeGIR)
```



(2g) Estimate lnPrize when GIR = 65

In [161]:
```
predict.lm(lnPrizeGIR, data.frame(GIR=65), se.fit = T)
```

$fit
1: 10.3442535019428
$se.fit
0.0607328205851393
$df
194
$residual.scale
0.848270263557247

(2h) Transform the fitted value in part 2g back to $ units

In [171]:
```
'$ unit'
exp(10.34425)
```

'$ unit'

31077.8305524319

3. Titanic-Survival-Data.xlsx analysis

(3a) Summarize Age

In [3]:
```
titanic <- read_excel('Titanic-Survival-Data.xlsx')
```

In [4]:
```
str(titanic)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':	2201 obs. of  5 variables:
 $ Class       : chr  "Coach" "Coach" "Coach" "Coach" ...
 $ Gender      : chr  "Female" "Female" "Female" "Female" ...
 $ Age         : num  20 21 26 26 36 41 41 45 45 48 ...
 $ Status      : chr  "Survived" "Survived" "Survived" "Died" ...
 $ ChildorAdult: chr  "Adult" "Adult" "Adult" "Adult" ...
```

10% trimmed mean

In [5]:
```
mean(titanic$Age, trim = 0.1)
```

47.4599659284497

Standard Deviation

In [6]:
```
sd(titanic$Age)
```

19.5976186008994

In [7]:
```
summary(titanic$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   31.00   48.00   47.12   64.00   80.00 
```

(3b) Tabulate survivors vs non-survivors

In [8]:
```
table(titanic$Status)
```

```
    Died Survived 
    1490      711 
```

(3c) Tabulate males vs females

In [9]:
```
table(titanic$Gender)
```

```
Female   Male 
   470   1731 
```

(3d) Tabulate the number in each pasenger class

In [10]:
```
table(titanic$Class)
```

```
Coach First 
 1876   325 
```

(3e) Cross-tabulate survivors vs non-survivors by gender

In [11]:
```
table(titanic$Status, titanic$Gender)
```

```
          
           Female Male
  Died        126 1364
  Survived    344  367
```

(3f) Cross-tabulate survivors vs non-survivors by passenger class

In [12]:
```
table(titanic$Status, titanic$Class)
```

```
          Coach First
Died       1368   122
Survived    508   203
```

(3g) Cross-tabulate survivors vs non-survivors by ChildorAdult

In [13]:
```
table(titanic$Status, titanic$ChildorAdult)
```

```
          Adult Child
Died       1438    52
Survived    654    57
```

(3h) Estimate proportion of males that survived and its standard error

Sample proportion

In [14]:
```
pMale <- 367/(367 + 1364)
pMale
```

0.212016175621028

Estimated standard error

In [15]:
```
SEpMale <- sqrt(pMale*(1-pMale)/(367 + 1364))
SEpMale
```

0.00982414164386597

(3i) Estimate proportion of females that survived and its standard error

Sample proportion

In [16]:
```
pFemale <- 344/(344 + 126)
pFemale
```

0.731914893617021

Estimated standard error

In [17]:
```
SEpFemale <- sqrt(pFemale*(1-pFemale)/(344 + 126))
SEpFemale
```

0.0204323211894421

(3j) Estimate probability of survival for children and its standard error

Sample proportion

In [18]:
```
pCoach <- 508/(508 + 1368)
pCoach
```

0.270788912579957

Estimated standard error

In [19]:
```
SEpCoach <- sqrt(pCoach*(1-pCoach)/(344 + 126))
SEpCoach
```

0.02049713407075

(3k) Estimate probability of survival for children and its standard error

Sample proportion

In [20]:
```
pChild <- 57/(52+57)
pChild
```

0.522935779816514

Estimated standard error

In [21]:
```
SEpChild <- sqrt(pChild*(1-pChild)/(52 + 57))
SEpChild
```

0.0478409012760106

4. Titanic-Survival-Data.xlsx analysis

(4a) Estimate median age of males that survive

Subset male survivors

In [32]:
```
maleSurvivors = subset(titanic, Gender == 'Male', Status = 'Survived')
```

Median age of male survivors

In [33]:
```
median(maleSurvivors$Age)
```

48

(4b) Estimate median age of females that survive

Subset female survivors

In [34]:
```
femaleSurvivors = subset(titanic, Gender == 'Female', Status = 'Survived')
```

Median age of female survivors

In [35]:
```
median(femaleSurvivors$Age)
```

47

(4c) Bootstrap sampling distribution of median age of males that survived

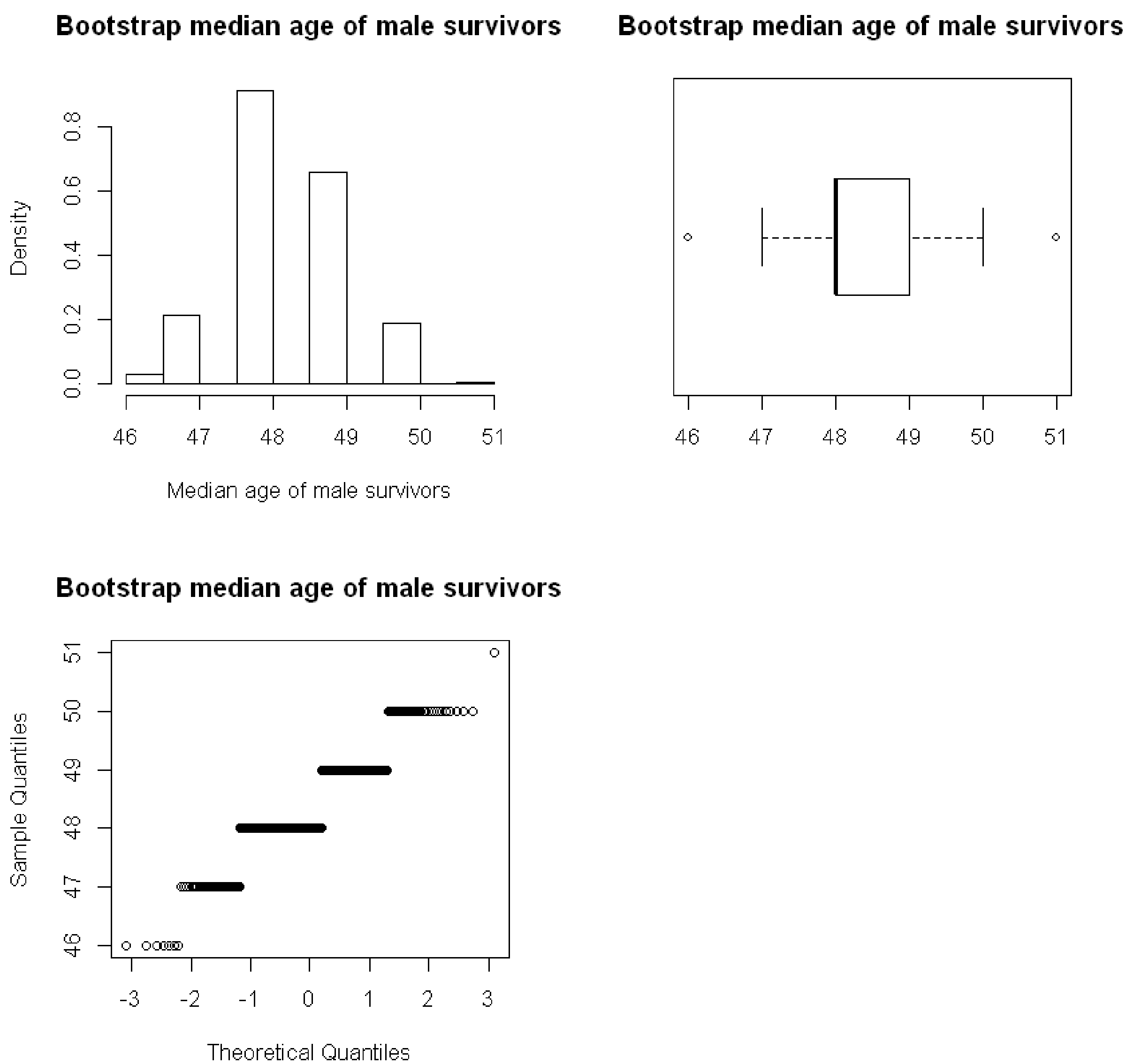
Bootstrap

In [55]:
```
medianFc <- function(x, i){
    return(median(x[i]))
}
medianMaleBoot <- boot(data = maleSurvivors$Age, statistic = medianFc, 500)
medianMaleBoot
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = maleSurvivors$Age, statistic = medianFc, R = 500)


Bootstrap Statistics :
    original  bias    std. error
t1*       48   0.388   0.8572782
```

Plots of sampling distribution

In [56]:
```
par(mfrow = c(2,2))
options(repr.plot.width=7, repr.plot.height=7)
hist(medianMaleBoot$t, freq = F, main = "Bootstrap median age of male survivor
s", xlab = "Median age of male survivors")
boxplot(medianMaleBoot$t, horizontal = T, main = "Bootstrap median age of male
 survivors")
qqnorm(medianMaleBoot$t, main = "Bootstrap median age of male survivors")
```



(4d) Bootstrap sampling distribution of median age of females that survived

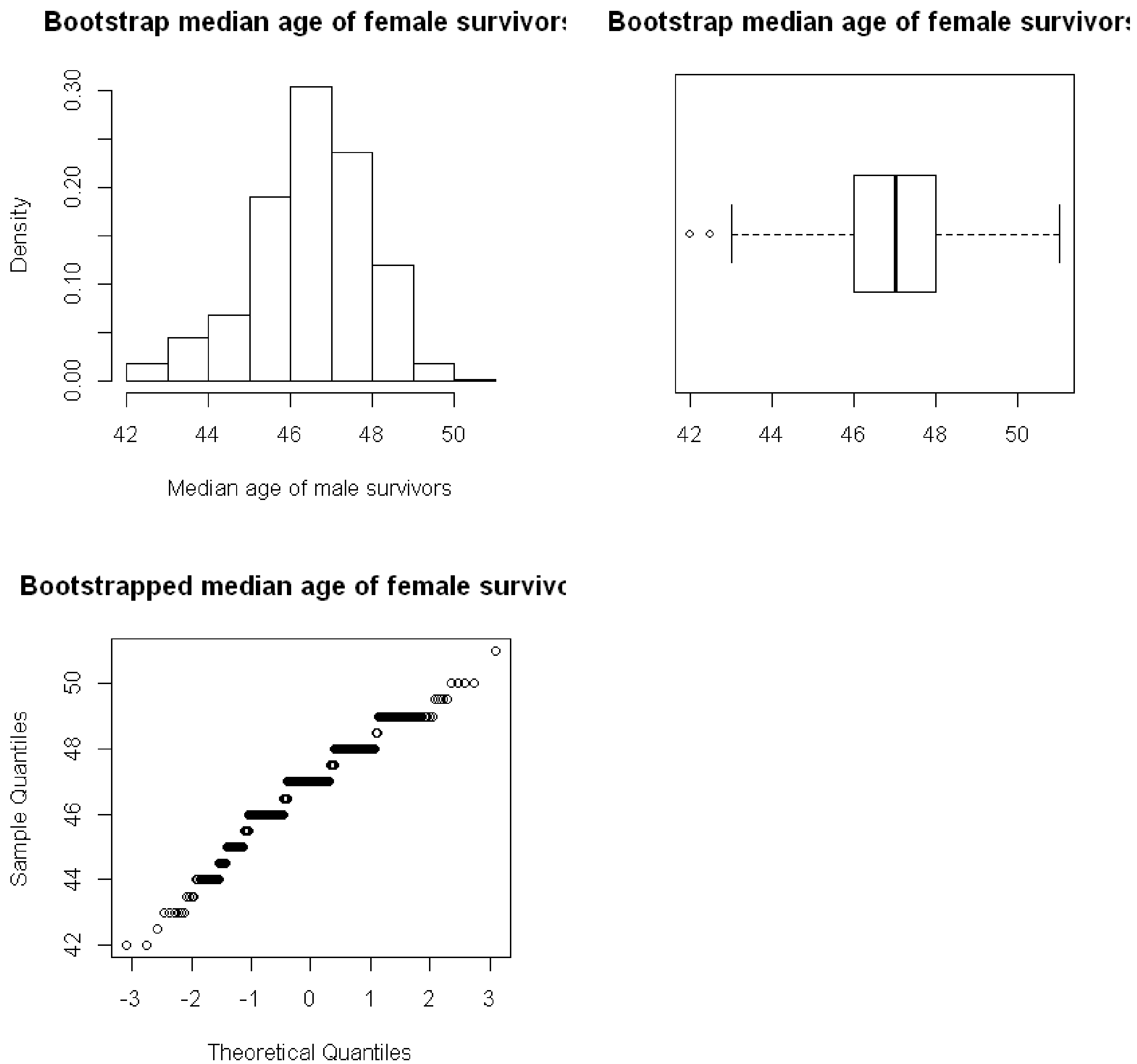
Bootstrap

In [57]:
```
medianFemaleBoot <- boot(data = femaleSurvivors$Age, statistic = medianFc, 500
)
medianFemaleBoot
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = femaleSurvivors$Age, statistic = medianFc, R = 500)


Bootstrap Statistics :
    original  bias    std. error
t1*       47  -0.058    1.457355
```

Plots of sampling distribution

In [58]:
```
par(mfrow = c(2,2))
options(repr.plot.width=7, repr.plot.height=7)
hist(medianFemaleBoot$t, freq = F, main = "Bootstrap median age of female surv
ivors", xlab = "Median age of male survivors")
boxplot(medianFemaleBoot$t, horizontal = T, main = "Bootstrap median age of fe
male survivors")
qqnorm(medianFemaleBoot$t, main = "Bootstrapped median age of female survivor
s")
```



(4e) Estimate median age of first class passengers that survived

Subset first class survivors

In [59]:
```
firstClassSurvivors = subset(titanic, Class == 'First', Status = 'Survived')
```

Median age of first class survivors

In [60]:
```
median(firstClassSurvivors$Age)
```

50

(4f) Estimate median age of coach class passengers that survived

Subset coach class survivors

In [61]:
```
coachClassSurvivors = subset(titanic, Class == 'Coach', Status = 'Survived')
```

Median age of coach class survivors

In [62]:
```
median(coachClassSurvivors$Age)
```

48

(4g) Bootstrap sampling distribution of median age of first class passengers that survived

Bootstrap

In [63]:
```
medianFirstClassBoot <- boot(data = firstClassSurvivors$Age, statistic = media
nFc, 500)
medianFirstClassBoot
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = firstClassSurvivors$Age, statistic = medianFc, R = 500)


Bootstrap Statistics :
    original  bias    std. error
t1*       50  -0.494    1.614676
```

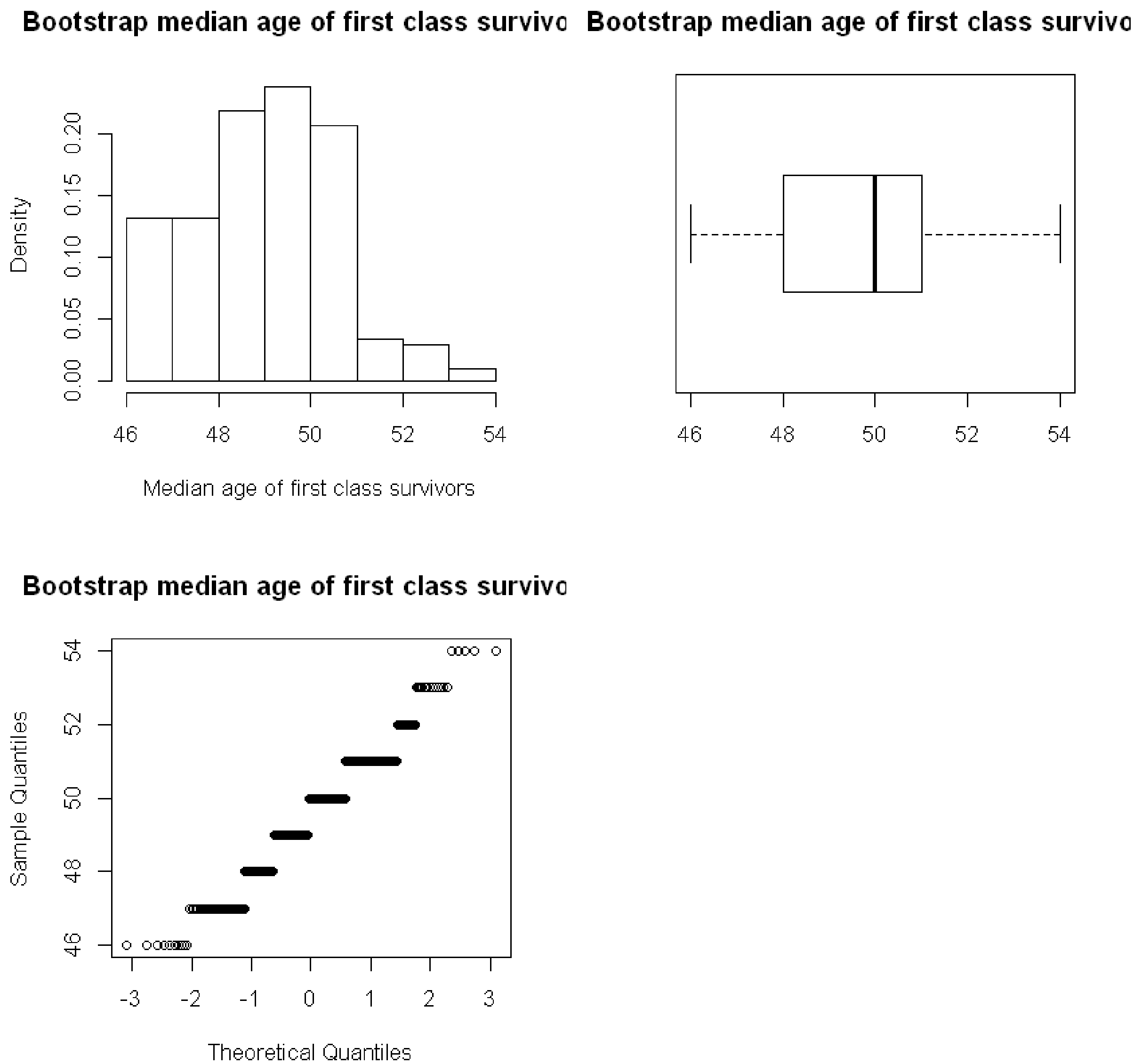
Deciles of sampling distribution

In [72]:
```
quantile(medianFirstClassBoot$t, probs = seq(0.1,0.9,0.1))
```

10%	47
20%	48
30%	49
40%	49
50%	50
60%	50
70%	50
80%	51
90%	51

Plots of sampling distribution

In [64]:
```r
par(mfrow = c(2,2))
options(repr.plot.width=7, repr.plot.height=7)
hist(medianFirstClassBoot$t, freq = F, main = "Bootstrap median age of first class survivors", xlab = "Median age of first class survivors")
boxplot(medianFirstClassBoot$t, horizontal = T, main = "Bootstrap median age of first class survivors")
qqnorm(medianFirstClassBoot$t, main = "Bootstrap median age of first class survivors")
```



(4h) Bootstrap sampling distribution of median age of coach class passengers that survived

In [65]:
```
medianCoachClassBoot <- boot(data = coachClassSurvivors$Age, statistic = media
nFc, 500)
medianCoachClassBoot
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = coachClassSurvivors$Age, statistic = medianFc, R = 500)


Bootstrap Statistics :
    original  bias    std. error
t1*       48  -0.188    0.829477
```

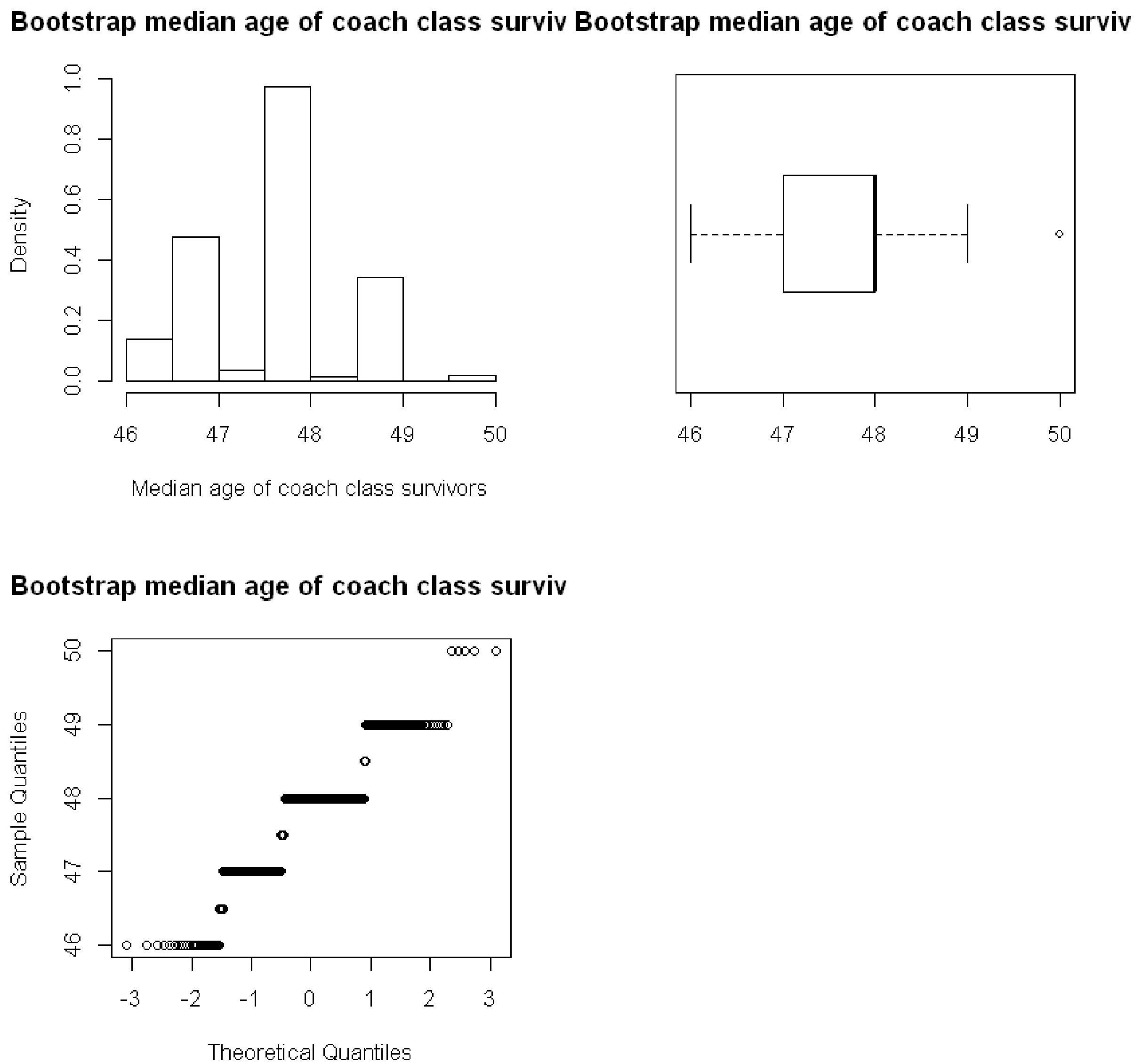
Quartiles of sampling distribution

In [78]:
```
quantile(medianFirstClassBoot$t, c(.25,.5,.75))
```

25%	48
50%	50
75%	51

Plots of sampling distribution

In [66]:
```
par(mfrow = c(2,2))
options(repr.plot.width=7, repr.plot.height=7)
hist(medianCoachClassBoot$t, freq = F, main = "Bootstrap median age of coach c
lass survivors", xlab = "Median age of coach class survivors")
boxplot(medianCoachClassBoot$t, horizontal = T, main = "Bootstrap median age o
f coach class survivors")
qqnorm(medianCoachClassBoot$t, main = "Bootstrap median age of coach class sur
vivors")
```



End