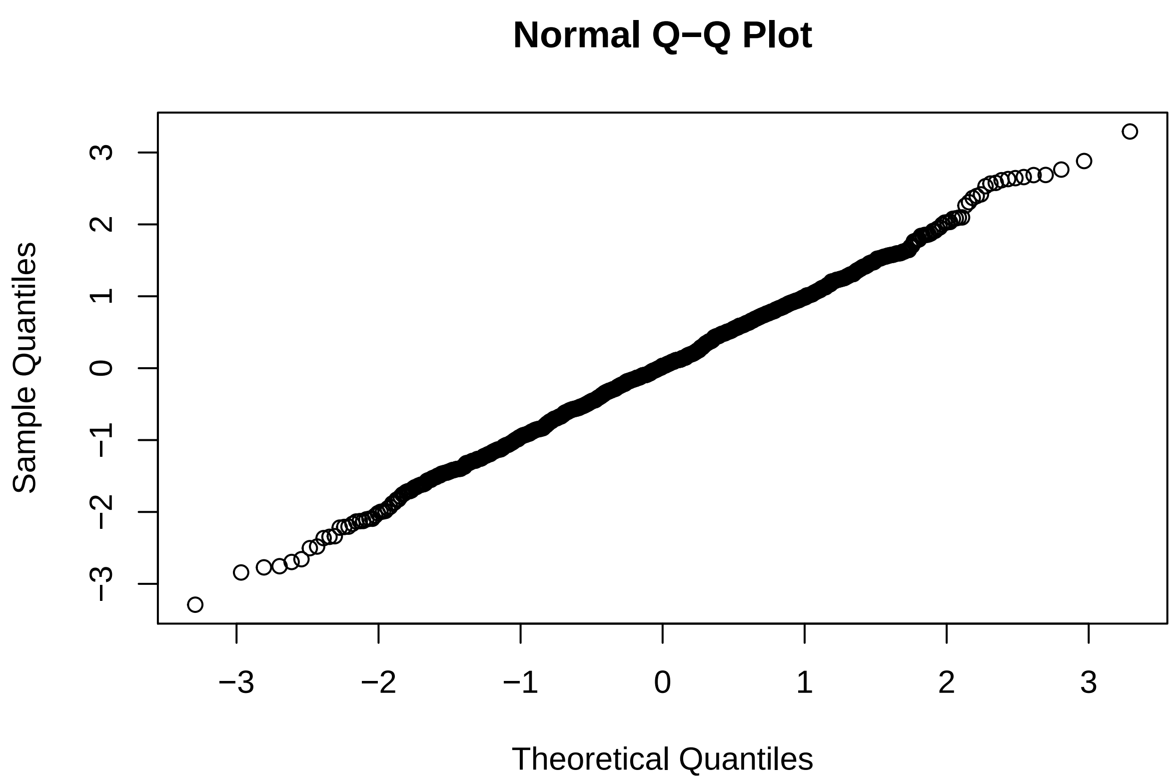
HW6

Yigao Li

October 29, 2017

Problem 1

```r
n <- 1000
z <- replicate(n, sum(runif(12))-6)
qqnorm(z)
```



$X_i \sim Uniform(0,1)$, $E[Z] = E[X_1 + X_2 + \cdots + X_{12} - 6] = E[X_1] + E[X_2] + \cdots + E[X_{12}] - 6 = \frac{1}{2} \times 12 - 6 = 0$, $Var(Z) = Var(X_1 + X_2 + \cdots + X_{12} - 6) = Var(X_1) + Var(X_2) + \cdots + Var(X_{12}) = \frac{1}{12} \times 12 = 1$

Problem 2

(a) Expected value of a sample mean is $\frac{1}{\lambda} = 10$

(b)

```r
n <- 1000
samples <- replicate(n, mean(rexp(30, 1/10)))
print(sum(samples >= 12)/n)
```

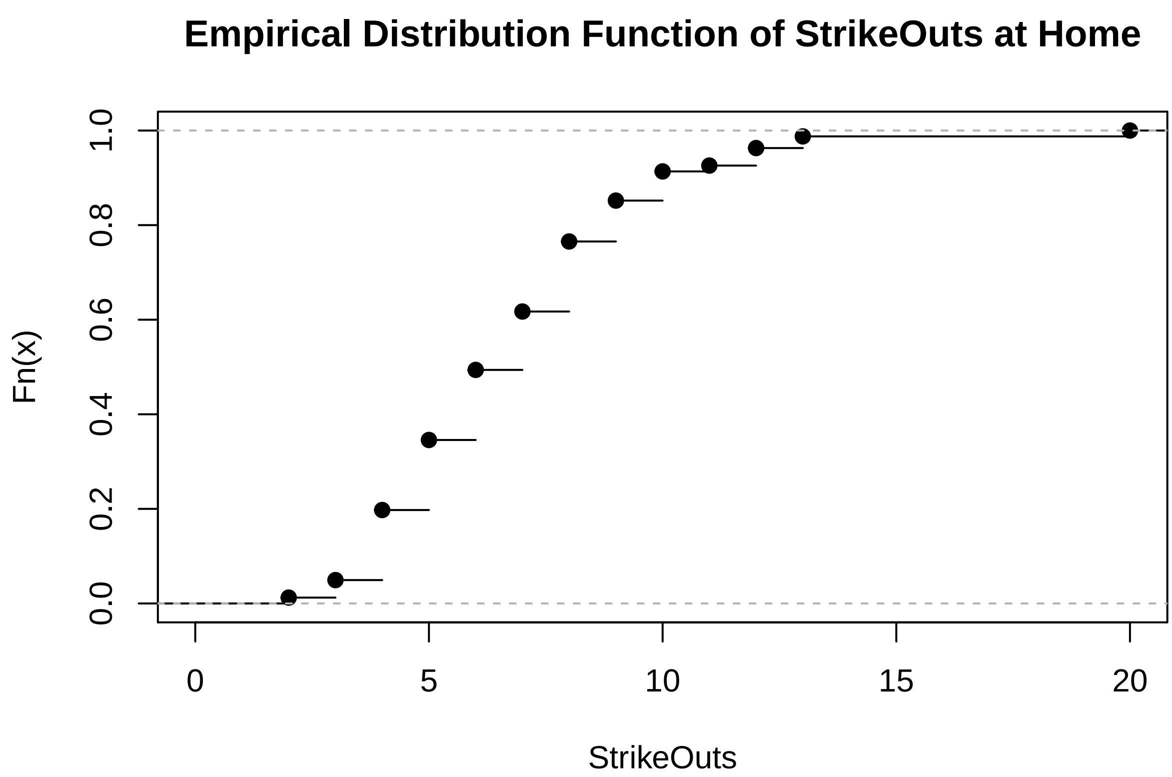
```
## [1] 0.125
```

(c) The probability of the sample mean no less than 12 is greater than 0.05. There is not enough evidence that a mean of 12 is unusual.

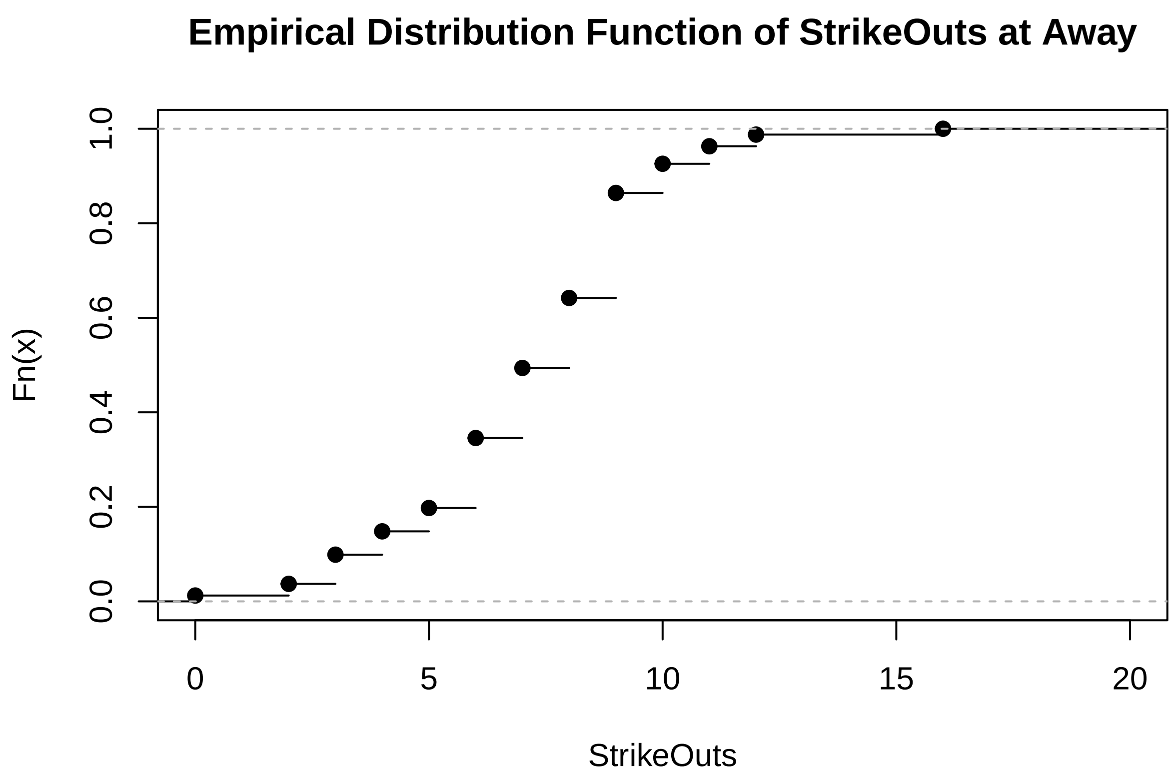
Problem 3

(a)

```r
phillies <- read.csv("D:/Courses/ANLY 511/Phillies2009.csv")
strikeoutsHome <- subset(phillies$StrikeOuts, phillies$Location == "Home")
strikeoutsAway <- subset(phillies$StrikeOuts, phillies$Location == "Away")
plot(ecdf(strikeoutsHome), main = "Empirical Distribution Function of StrikeOuts at Home",
     xlab = "StrikeOuts", xlim = c(0,20))
```




```r
plot(ecdf(strikeoutsAway), main = "Empirical Distribution Function of StrikeOuts at Away",
     xlab = "StrikeOuts", xlim = c(0,20))
```



(b)

```r
aggregate(StrikeOuts ~ Location, data = phillies, FUN = mean)
```

```
##   Location StrikeOuts
## 1     Away   7.308642
## 2     Home   6.950617
```

(c)

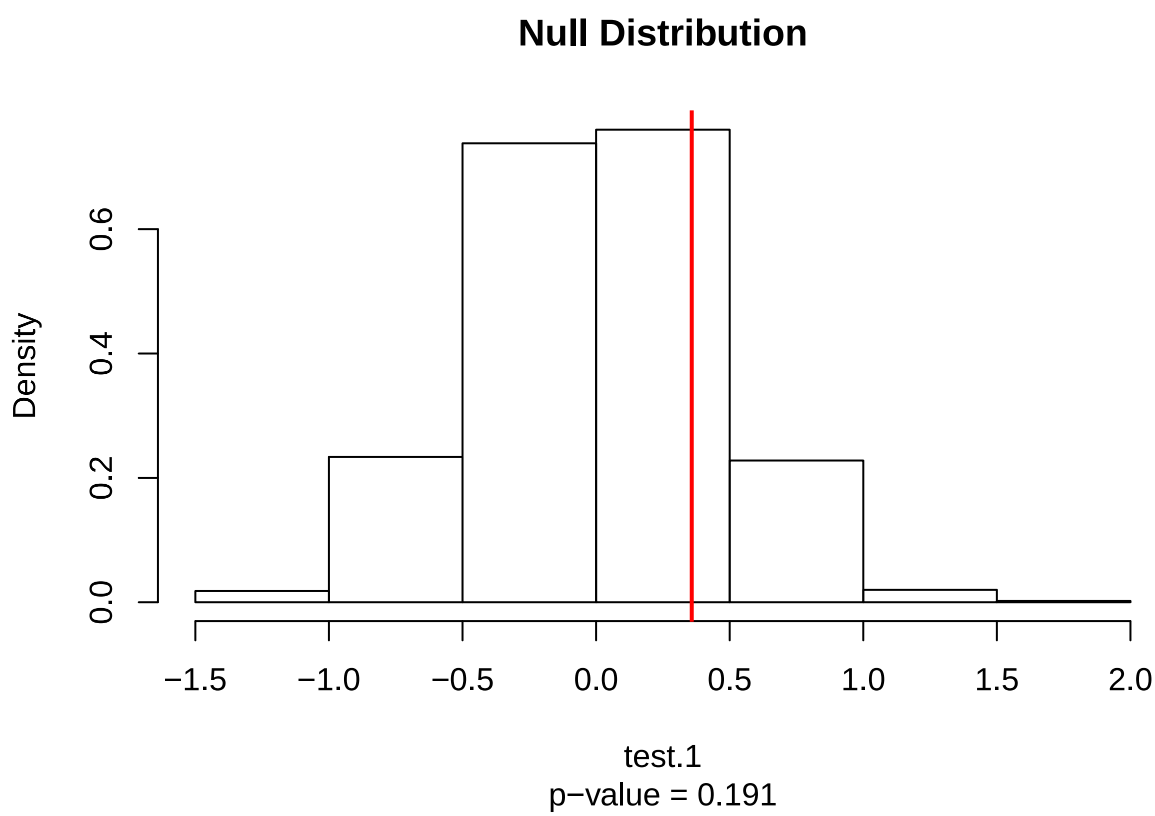
$H_0 : \mu_{StrikeOutsHome} = \mu_{StrikeOutsAway}$
$H_1 : \mu_{StrikeOutsHome} < \mu_{StrikeOutsAway}$

```r
mytest.1 <- function(mydf){
  agg <- aggregate(StrikeOuts ~ Location, data = mydf, FUN = mean)
  return(agg$StrikeOuts[1] - agg$StrikeOuts[2])
}

permute.sample.1 <- function(mydf){
  n <- dim(mydf)[1]
  mydf$Location <- mydf$Location[sample(n, n, replace = F)]
  return(mytest.1(mydf))
}

phillies.permute <- phillies
N <- 1000
test.1 <- replicate(N, permute.sample.1(phillies.permute))
hist(test.1, main = "Null Distribution", prob = T,
```

```r
     sub = paste("p-value =", mean(test.1 > mytest.1(phillies))))
abline(v = mytest.1(phillies), col = 2, lwd = 2)
```



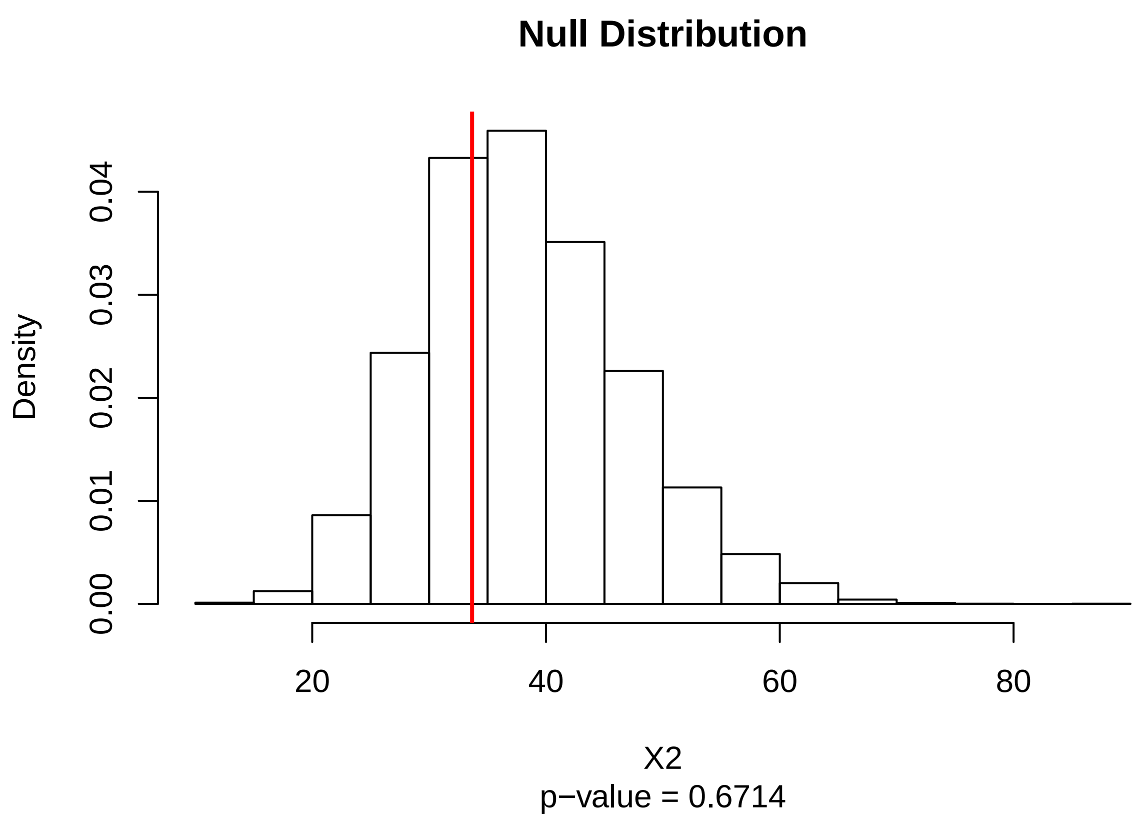
Because p-value is greater than 0.05, we fail to reject H_0 at the significance level 0.05.

Problem 4

H_0: winning numbers are randomly drawn.
H_1: winning numbers are not randomly drawn.

```r
lottery <- read.csv("D:/Courses/ANLY 511/Lottery.csv")
myX2 <- function(x, n, pvec){
  k <- length(pvec)
  pvec <- pvec/sum(pvec)
  expected <- n*pvec
  return(sum((x - expected)^2/expected))
}
lotterytable <- table(lottery)
freq <- as.vector(lotterytable)
mySim.1 <- function(n, pvec){
  x <- rmultinom(1, n, pvec)
  x <- as.vector(x)
  return(myX2(x, n, pvec))
}
myX2test <- function(x.obs, n, pvec, N){
  myX2dist <- replicate(N, mySim.1(n, pvec))
```

```
  X2.obs <- myX2(x.obs, n, pvec)
  p.value <- mean(myX2dist > X2.obs)
  hist(myX2dist, main = "Null Distribution", sub = paste("p-value =", p.value), xlab = "X2", prob = T)
  abline(v = X2.obs, col = 2, lwd = 2)
  return(p.value = p.value)
}
myX2test(freq, 500, rep(1,39), 10000)
```




```
## [1] 0.6714
```

Since p-value is greater than 0.05. We fail to reject H_0 at significance level 0.05.

Problem 5

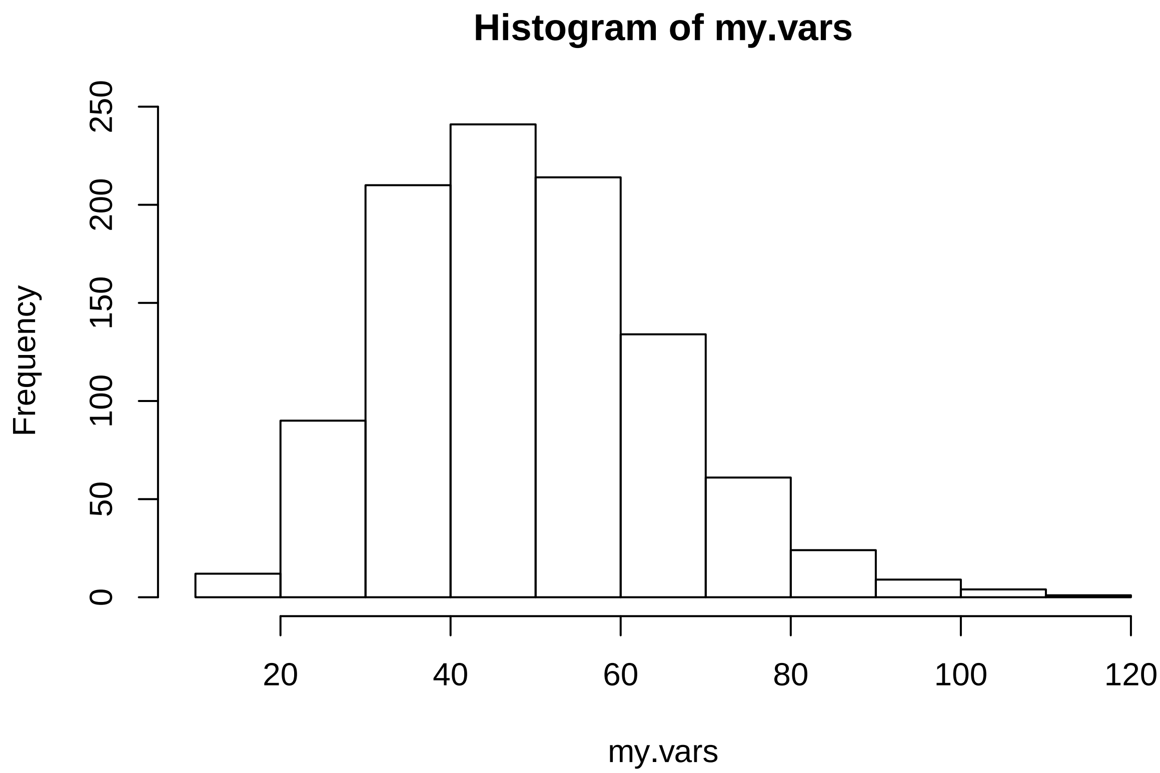
```
samplef <- function(x){var(rnorm(20, 25, 7))}
my.vars <- numeric(1000)
my.vars <- sapply(my.vars, samplef)
mean(my.vars)
```

```
## [1] 49.21409
```

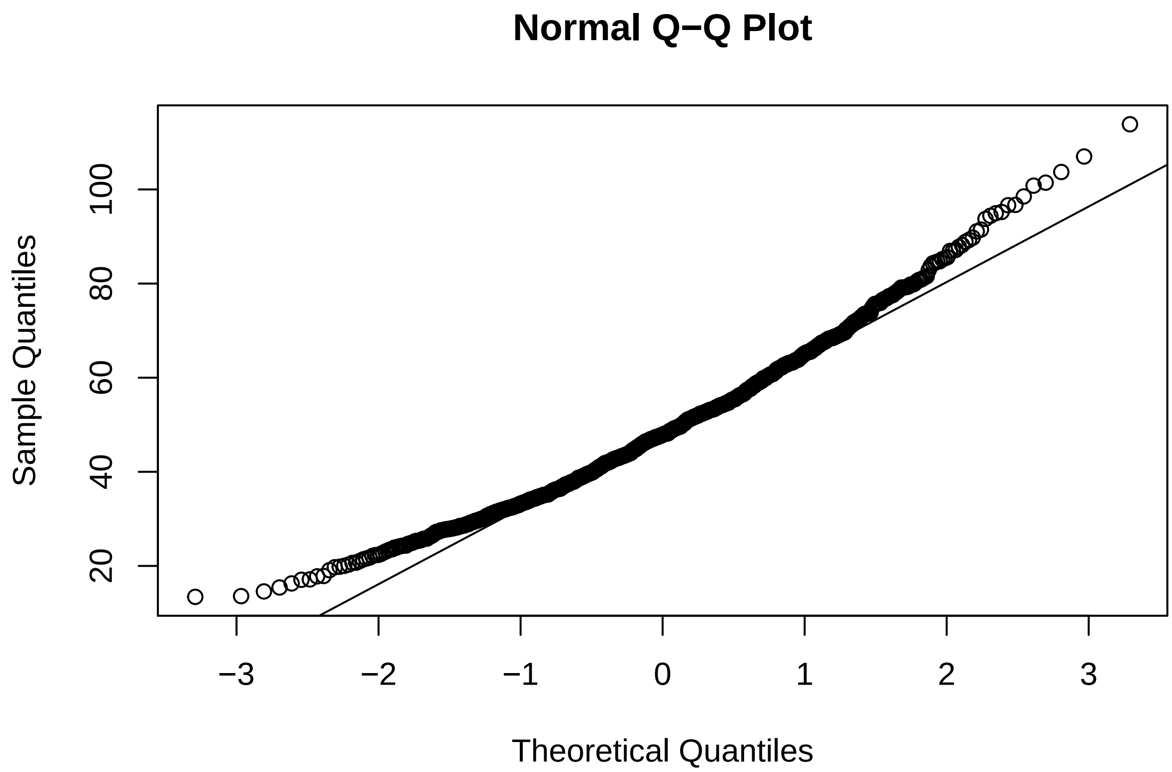
```
var(my.vars)
```

```
## [1] 255.2655
```

```
hist(my.vars)
```




```
qqnorm(my.vars)
qqline(my.vars)
```



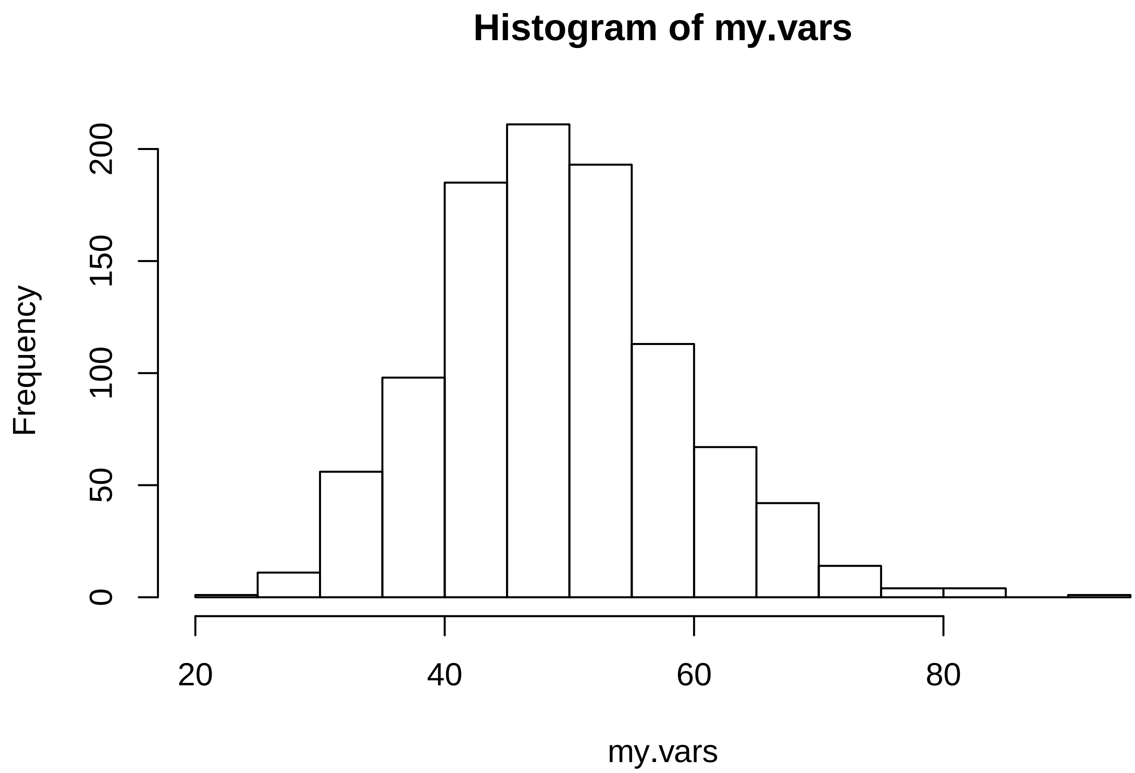

```
samplef <- function(x){var(rnorm(50, 25, 7))}
my.vars <- numeric(1000)
my.vars <- sapply(my.vars, samplef)
mean(my.vars)
```

```
## [1] 49.03478
```

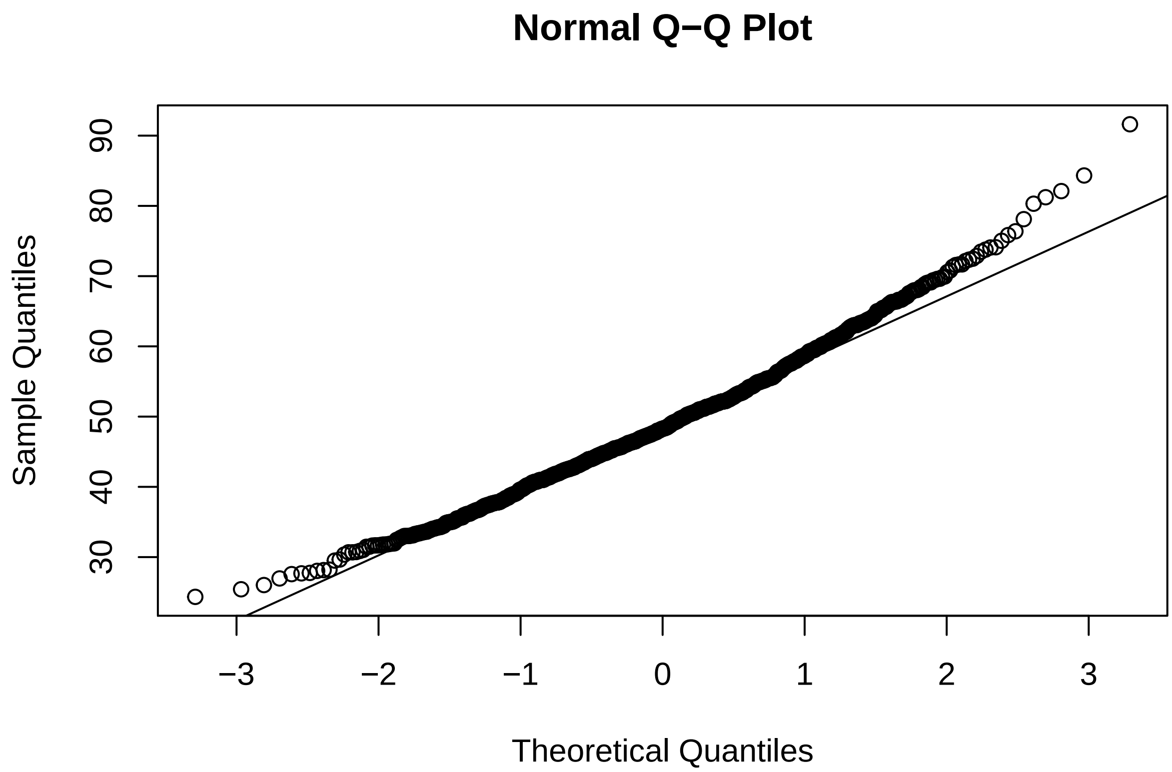
```
var(my.vars)
```

```
## [1] 95.15494
```

```
hist(my.vars)
```




```r
qqnorm(my.vars)
qqline(my.vars)
```



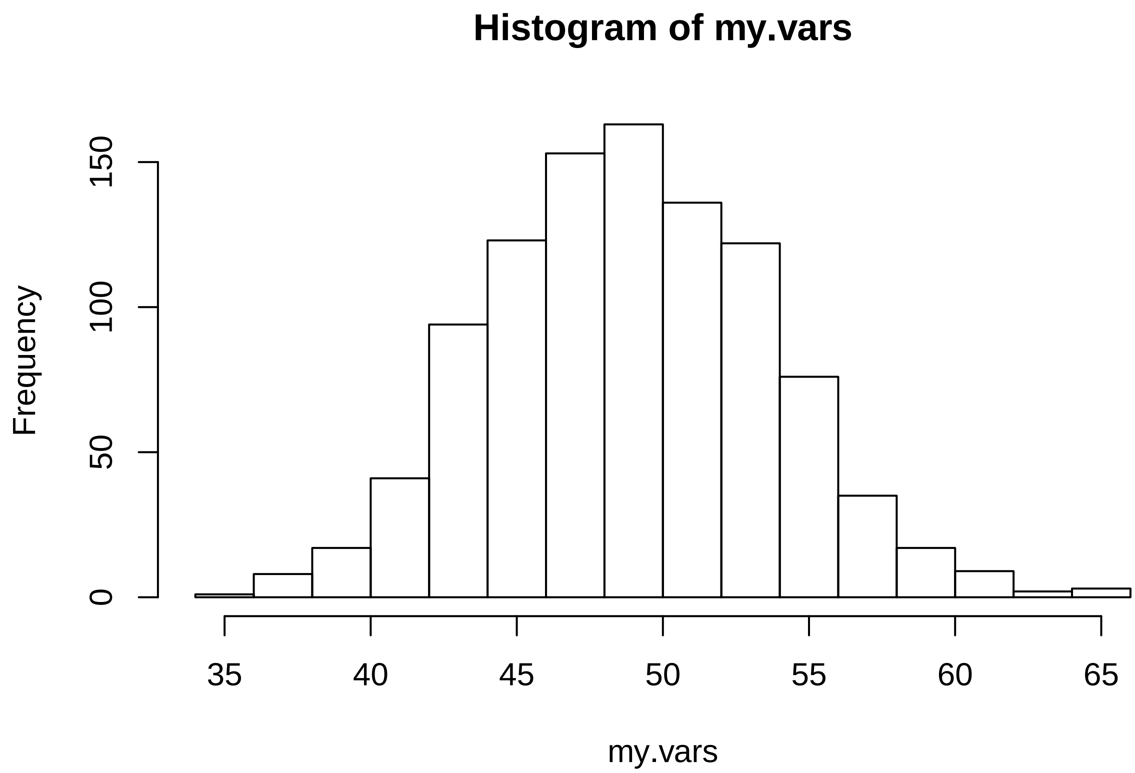

```r
samplef <- function(x){var(rnorm(200, 25, 7))}
my.vars <- numeric(1000)
my.vars <- sapply(my.vars, samplef)
mean(my.vars)
```

```
## [1] 48.85905
```

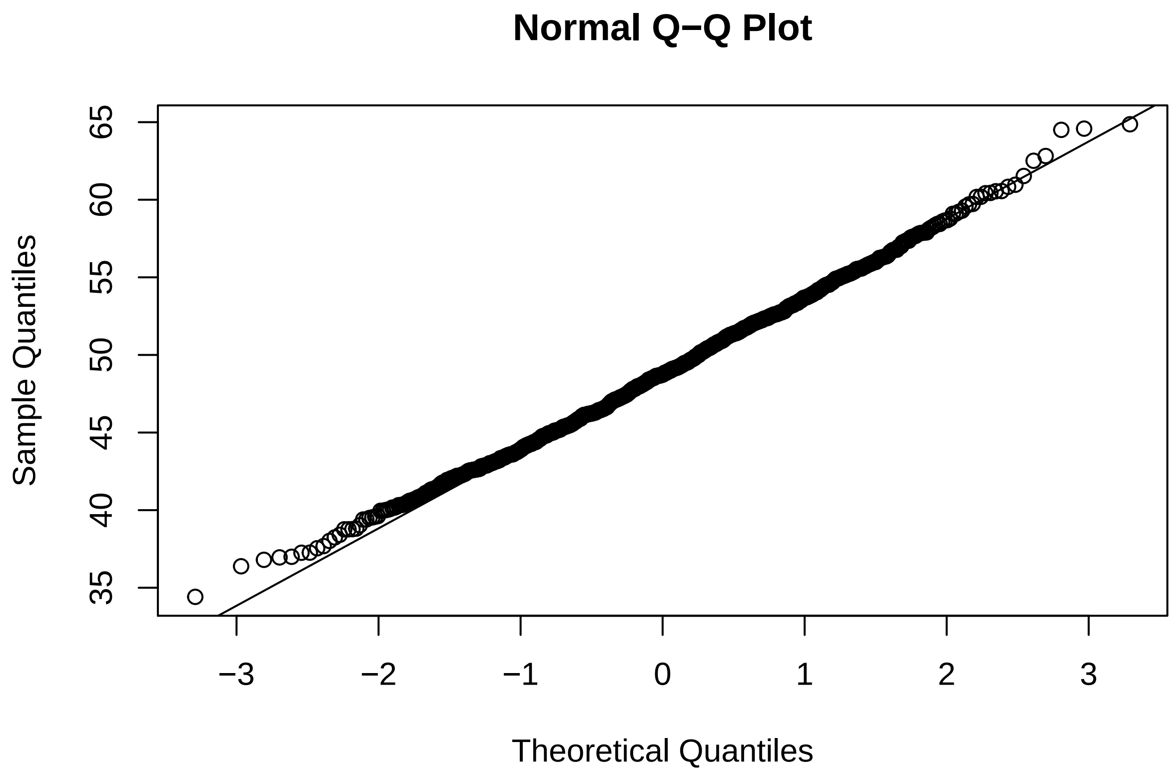
```r
var(my.vars)
```

```
## [1] 23.21553
```

```r
hist(my.vars)
```




```
qqnorm(my.vars)
qqline(my.vars)
```



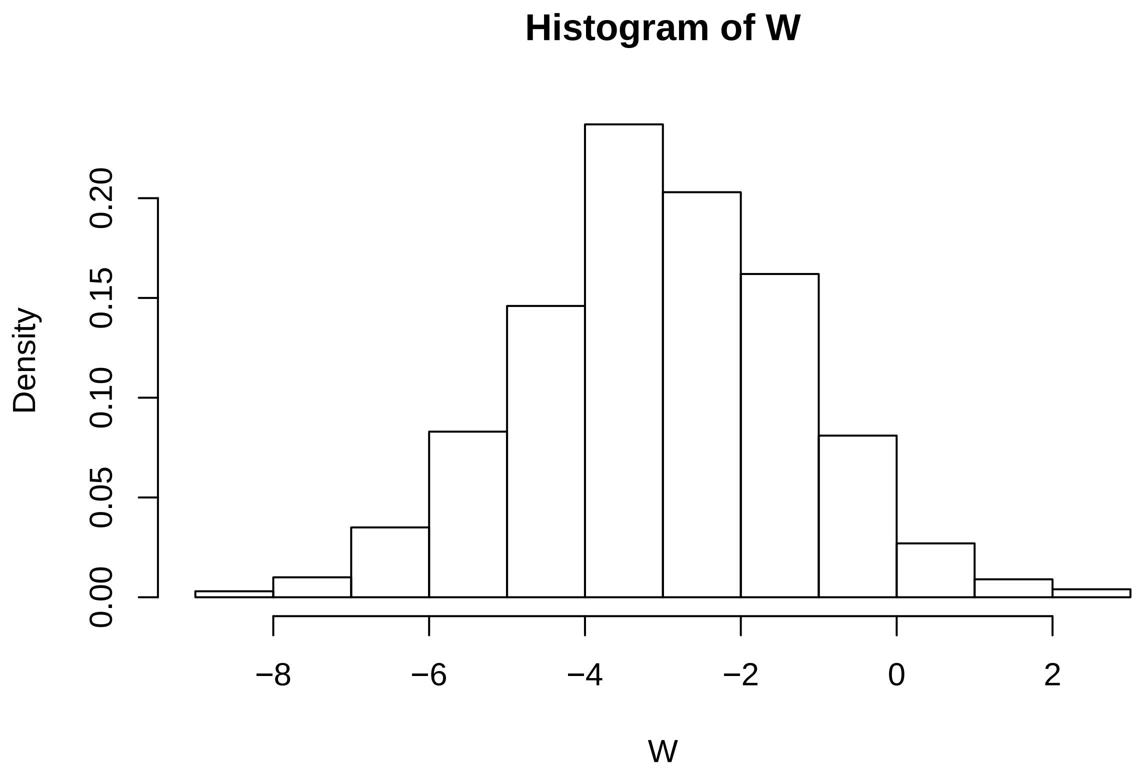
Sampling distribution appears to be normally distributed when $n = 200$.

Problem 6

(a) $W = \bar{X} - \bar{Y} = \frac{1}{9}X_1 + \frac{1}{9}X_2 + \cdots + \frac{1}{9}X_9 - \frac{1}{12}Y_1 - \frac{1}{12}Y_2 - \cdots - \frac{1}{12}Y_{12}$.
Because W is a linear combination of normal distributions X_i, Y_i, W is also a normal distribution with mean $E[W] = \frac{1}{9}E[X_1] + \frac{1}{9}E[X_2] + \cdots + \frac{1}{9}E[X_9] - \frac{1}{12}E[Y_1] - \frac{1}{12}E[Y_2] - \cdots - \frac{1}{12}E[Y_{12}] = 7 - 10 = -3$, and variance $Var(W) = \frac{1}{81}Var(X_1) + \frac{1}{81}Var(X_2) + \cdots + \frac{1}{81}Var(X_9) + \frac{1}{144}Var(Y_1) + \frac{1}{144}Var(Y_2) + \cdots + \frac{1}{144}Var(Y_{12}) = \frac{1}{9} \times 3^2 + \frac{1}{12} \times 5^2 = \frac{37}{12}$

(b)

```
W <- numeric(1000)
for (i in 1:1000){
  x <- rnorm(9, 7, 3)
  y <- rnorm(12, 10, 5)
  W[i] <- mean(x) - mean(y)
}
hist(W, prob = T)
```



(c)

```r
mean(W< (-1.5))
```

```
## [1] 0.798
```

```r
pnorm(-1.5, -3, sqrt(37/12))
```

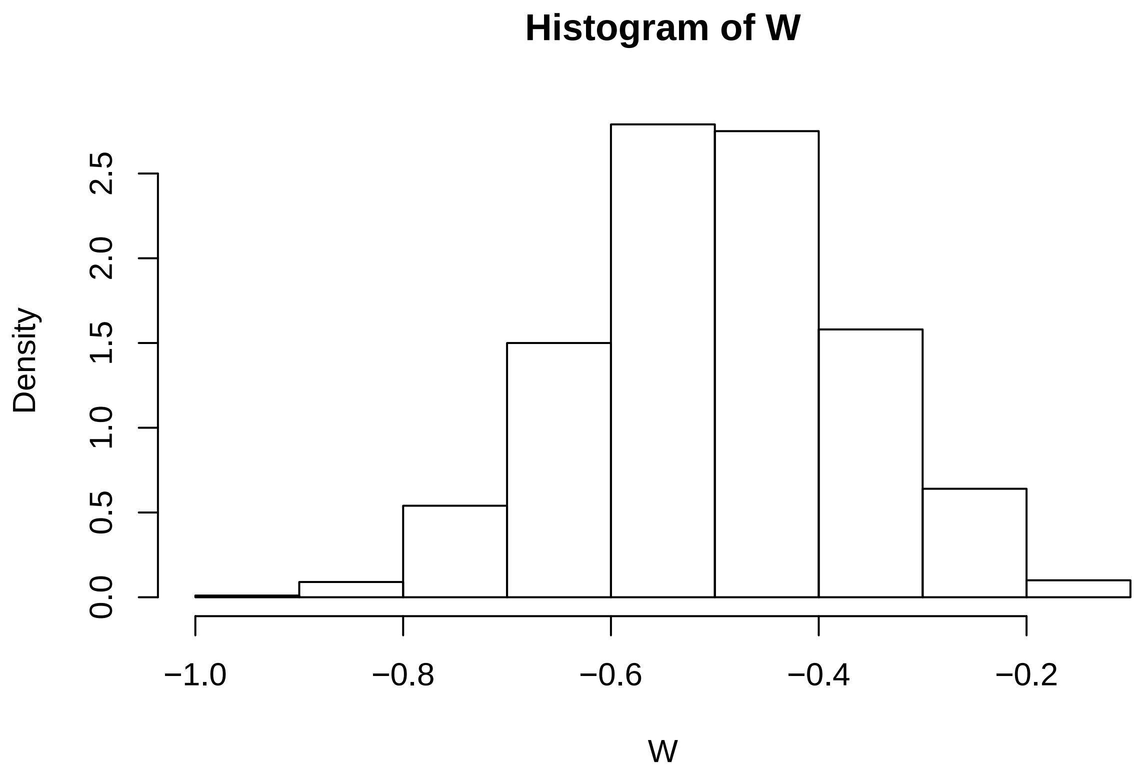
```
## [1] 0.8035146
```

Problem 7

(a) W is approximately normally distributed with mean $E[W] = \frac{1}{9}E[X_1] + \frac{1}{9}E[X_2] + \cdots + \frac{1}{9}E[X_9] - \frac{1}{12}E[Y_1] - \frac{1}{12}E[Y_2] - \cdots - \frac{1}{12}E[Y_{12}] = 0.5 - 1 = -0.5$, and variance $Var(W) = \frac{1}{81}Var(X_1) + \frac{1}{81}Var(X_2) + \cdots + \frac{1}{81}Var(X_9) + \frac{1}{144}Var(Y_1) + \frac{1}{144}Var(Y_2) + \cdots + \frac{1}{144}Var(Y_{12}) = \frac{1}{9} \times \frac{1}{12} + \frac{1}{12} \times \frac{1}{12} = \frac{7}{432}$

(b)

```r
W <- numeric(1000)
for (i in 1:1000){
  x <- runif(9)
  y <- runif(12, 0.5, 1.5)
  W[i] <- mean(x) - mean(y)
}
hist(W, prob = T)
```



(c) (Note: this part maybe is wrong because the maximum possible value for W is 0.5, which is always less than 0.6)

```
mean(W< 0.6)
```

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

```
pnorm(0.6, -0.5, sqrt(7/432))
```

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

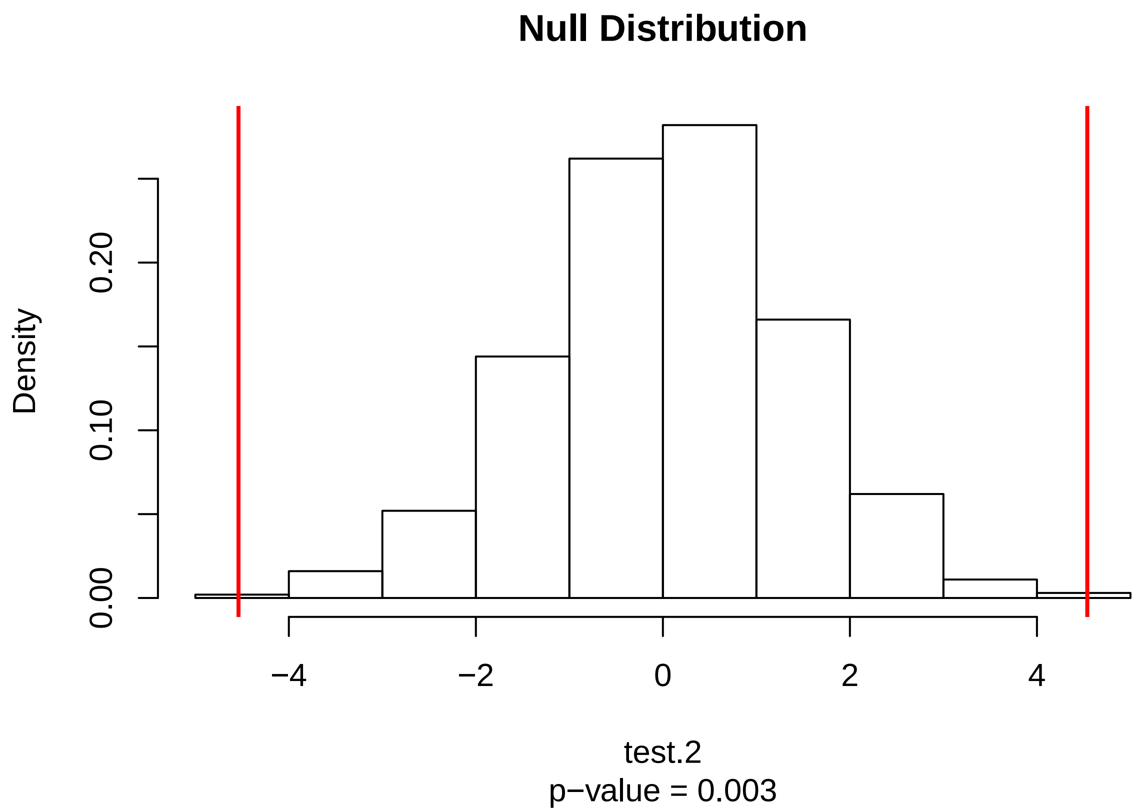
Problem 8

```
titani <- read.csv("D:/Courses/ANLY 511/Titanic.csv")
mytest.2 <- function(mydf){
  agg <- aggregate(Age ~ Survived, data = mydf, FUN = mean)
  return(agg$Age[1] - agg$Age[2])
}

permute.sample.2 <- function(mydf){
  n <- dim(mydf)[1]
  mydf$Survived <- mydf$Survived[sample(n, n, replace = F)]
  return(mytest.2(mydf))
}

titani.permute <- titani
```

```r
N <- 1000
test.2 <- replicate(N, permute.sample.2(titani.permute))
hist(test.2, main = "Null Distribution", prob = T,
     sub = paste("p-value =", mean(abs(test.2) > mytest.2(titani))), xlim = c(-5,5))
abline(v = mytest.2(titani), col = 2, lwd = 2)
abline(v = -mytest.2(titani), col = 2, lwd = 2)
```



Because p-value is smaller than 0.05, we reject H_0 at the significance level 0.05.