ANLY 520: Assignment 6

Brody Vogel

10/26/2017

Preparation

```
set.seed(1234)
```

Problem 1

The variance of one sample from the U(0,1) distribution is $\sigma^2 = 1/12$. So the variance of Z, which is 6 subtracted from the sum of 12 independent samples from the U(0,1) distribution, is $Var(Z) = Var(\sum_{i=1}^{12}(X_i) - 6) = Var(\sum_{i=1}^{12} X_i) = 12 \times Var(X_i) = 12 \times 1/12 = 1$, by the variance rules $Var(\sum_{i=1}^{j}(X_i)) = j \times Var(X_i)$ and $Var(\alpha X) = Var(X)$. Therefore Var(Z) = StD(Z) = 1.

The mean of Z = E[Z] = $\frac{12\times(1+0)}{2} - 6 = 0$. So the mean of Z = mean(X~N(0,1)), and we know from above that StD(Z) = StD(X~N(0,1)).

Thus, Z ~ N(0,1).

Problem 2

a) $E[\overline{x}] = \lambda^{-1} = (1/10)^{-1} = 10$.

b) It looks like ~13% of sample means are $\geq$ 12.

```
expys <- replicate(1000, mean(rexp(30, 1/10)))

p1 <- length(expys[expys >= 12])/length(expys)

p1
```

```
## [1] 0.119
```

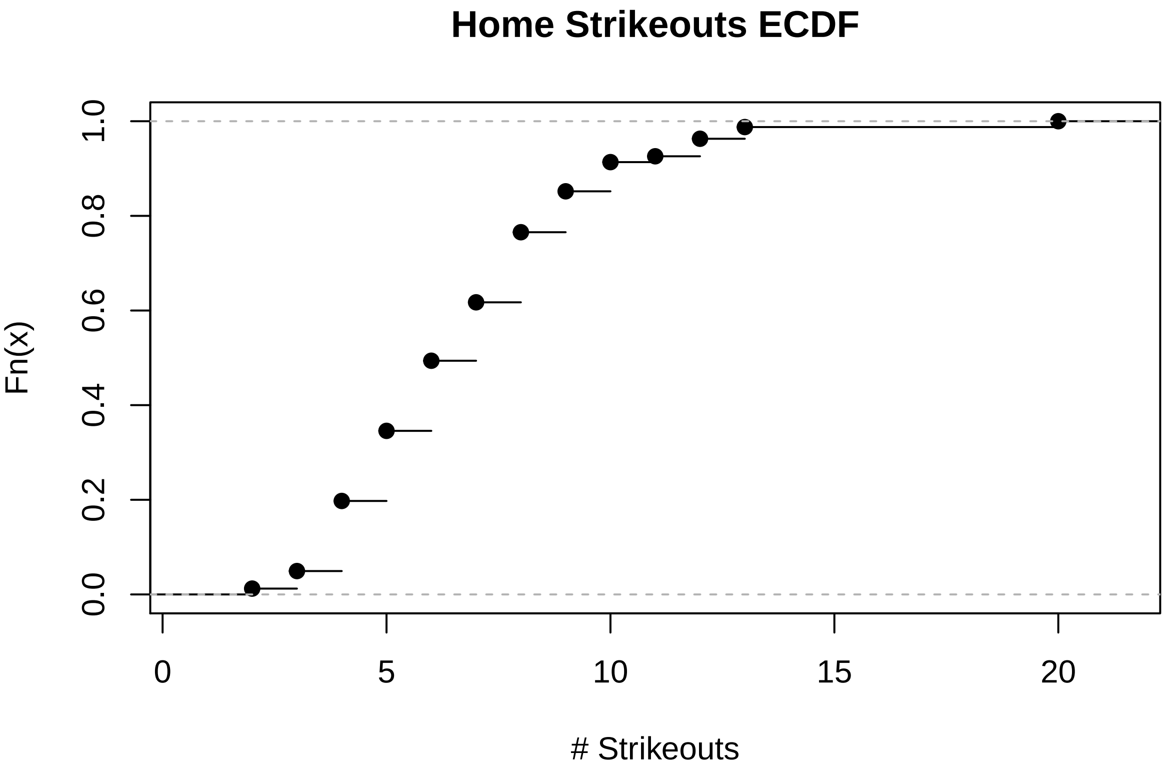
c) A mean of 12 is somewhat unusual for a sample size of 30 from an Exp(1/10) distribution, according to my simulation. The p-value is much too large to be statistically significant, though, and so I have no good reason to doubt the validity of my friend's claim.

Problem 3

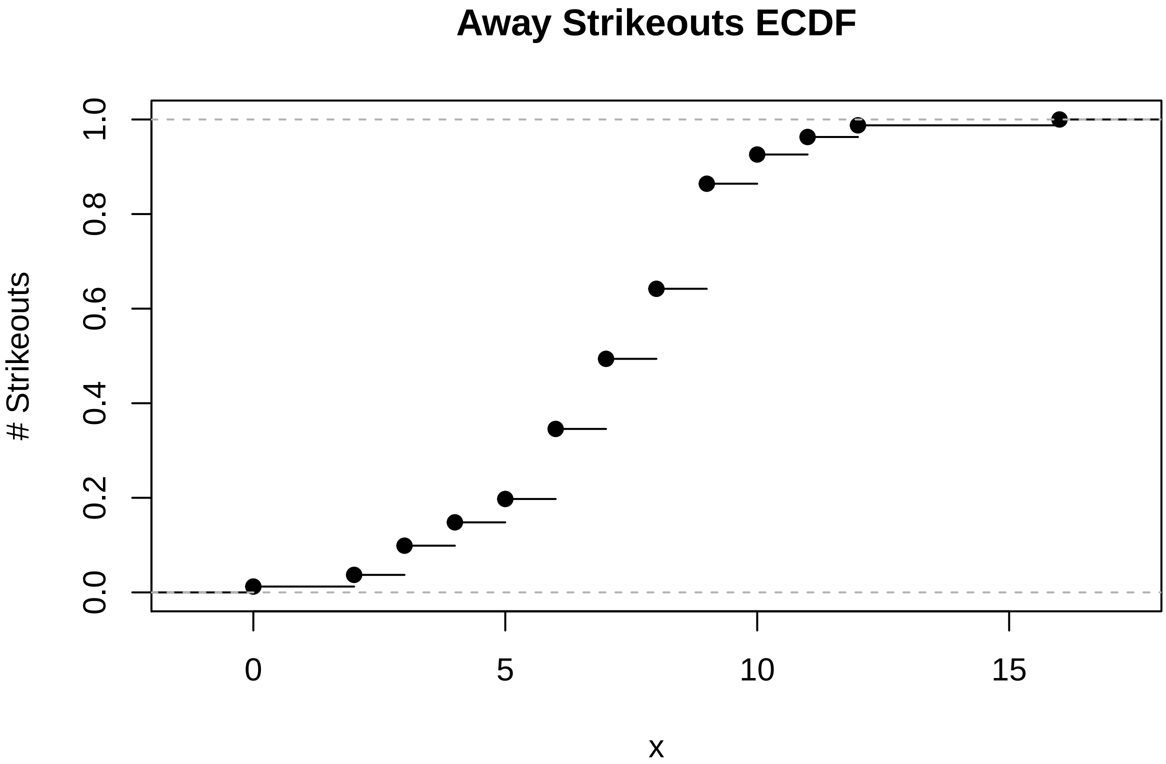
a) It looks like the Phillies struck out, on average, more on the road than at home, but the ECDF of their home strikeouts is heavily-affected by one 20-strikeout game, which makes it hard to estimate.

```
phils <- read.csv('/Users/brodyvogel/Desktop/Data/Phillies2009.csv', sep = ',')

homes <- phils[phils$Location == 'Home',]
aways <- phils[phils$Location == 'Away',]
```

```r
plot(ecdf(homes$StrikeOuts), main = 'Home Strikeouts ECDF', xlab = '# Strikeouts')
```



```r
plot(ecdf(aways$StrikeOuts), main = 'Away Strikeouts ECDF', ylab = '# Strikeouts')
```



b) $\mu_{home} \approx 6.951$; $\mu_{away} \approx 7.309$.

```
muHome <- mean(homes$StrikeOuts)
muAway <- mean(aways$StrikeOuts)

muHome
```

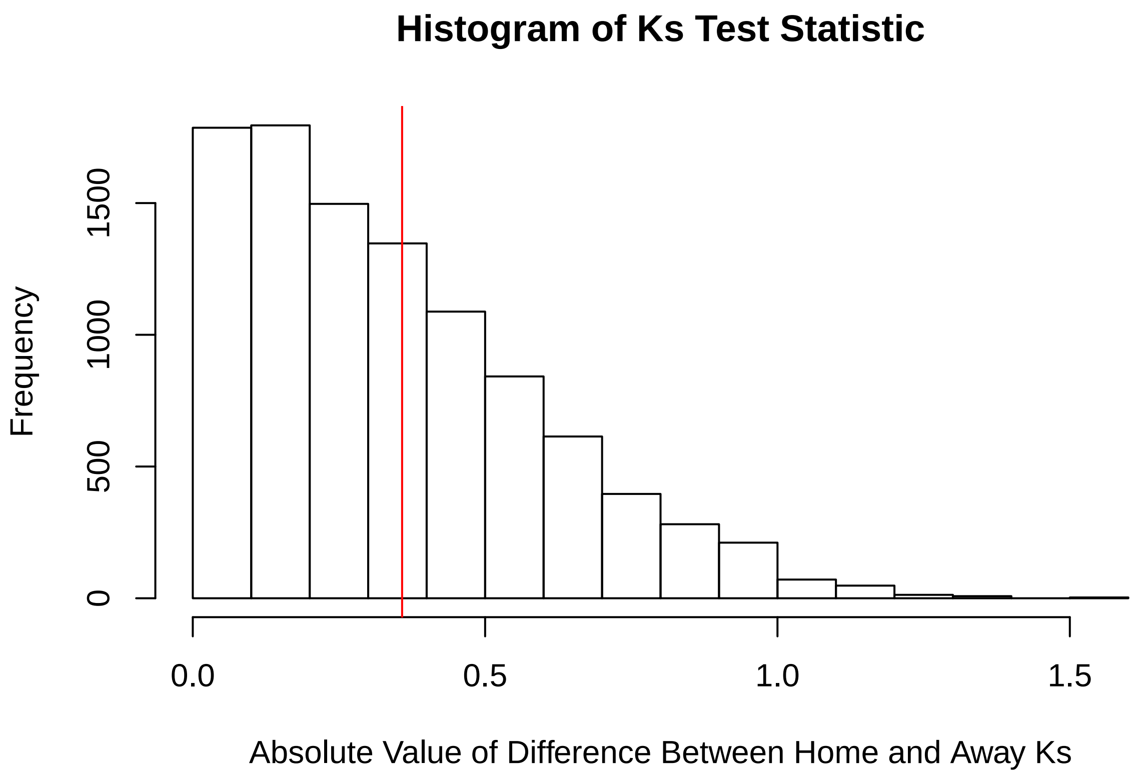
```
## [1] 6.950617
```

```
muAway
```

```
## [1] 7.308642
```

c) The difference between means is not statistically significant, with a simulated p-value of ≈ .427. I used the absolute value of the difference because the hypothesis we're testing doesn't presuppose that the Phillies strikeout more/less at home/on the road, just that the means are different.

```
a <- length(homes$StrikeOuts)

testStats <- c()
for (x in 1:10000) {
  samp <- sample(nrow(phils), a, replace = FALSE)
  helper1 <- phils[samp, ]
  helper2 <- phils[-samp, ]
  testStat <- abs(mean(helper1$StrikeOuts) - mean(helper2$StrikeOuts))
  testStats <- c(testStats, testStat)
}


hist(testStats, main = 'Histogram of Ks Test Statistic', xlab = 'Absolute Value of Difference Between Ho

testSTAT <- abs(muHome - muAway)

abline(v = testSTAT, col = 'red')
```



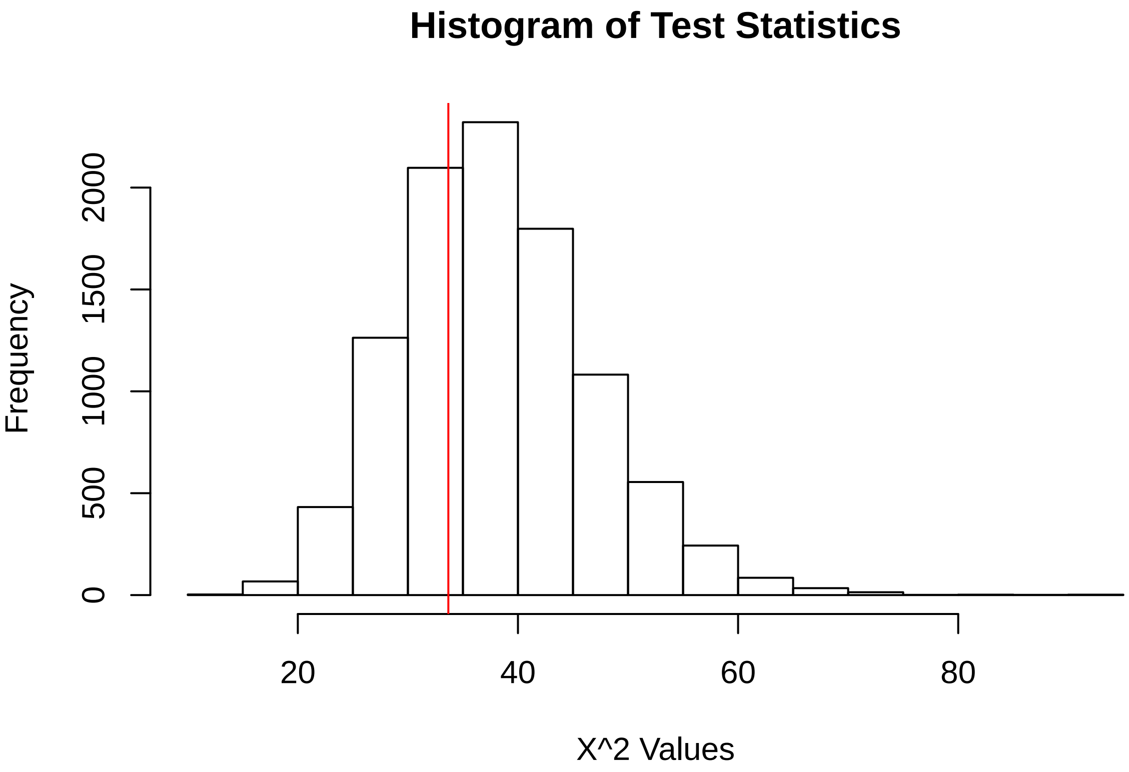
```
p2 <- length(testStats[testStats >= testSTAT])/length(testStats)
p2
```

```
## [1] 0.421
```

Problem 4

From my simulation, I got a p-value of $\approx .67$, which is roughly equivalent to the X^2 approximation of .669. This means that there is about a 67 percent chance of getting a value at least as extreme as that gleaned from the lottery data. From this, I conclude that we cannot reject the null hypothesis; that is, there is no evidence that the lottery doesn't follow a multinomial distribution, and so the lottery draws are fair.

```
lot <- read.csv('/Users/brodyvogel/Desktop/Data/Lottery.csv', sep = ',')

expected <- length(lot$Win)/39

wins <- table(lot$Win)

X2 <- sum((wins-expected)^2/expected)

tester2 <- function(n) {
  testStats2 <- c()
  for (x in 1:n) {
    randos <- rmultinom(1, 500, rep(1,39))
    testStat2 <- sum((randos-expected)^2/expected)
    testStats2 <- c(testStats2, testStat2)
  }
  return(testStats2)
}

testSTATS2 <- tester2(10000)
hist(testSTATS2, main = 'Histogram of Test Statistics', xlab = 'X^2 Values')
abline(v = X2, col = 'red')
```



```
p3 <- length(testSTATS2[testSTATS2 >= X2])/length(testSTATS2)
p3
```

```
## [1] 0.6695
```

```
### by the Pearson-Fisher approximation
1 - pchisq(X2, df = 38)
```

```
## [1] 0.669616
```

Problem 5

No, the sampling distribution doesn't look to be normally-distributed; the histogram is right-skewed and the qqplot varies significantly from the imposed normal line. This, I think, is because the sample size is 20, which is less than the rule of thumb for the Central Limit Theorem of 30.

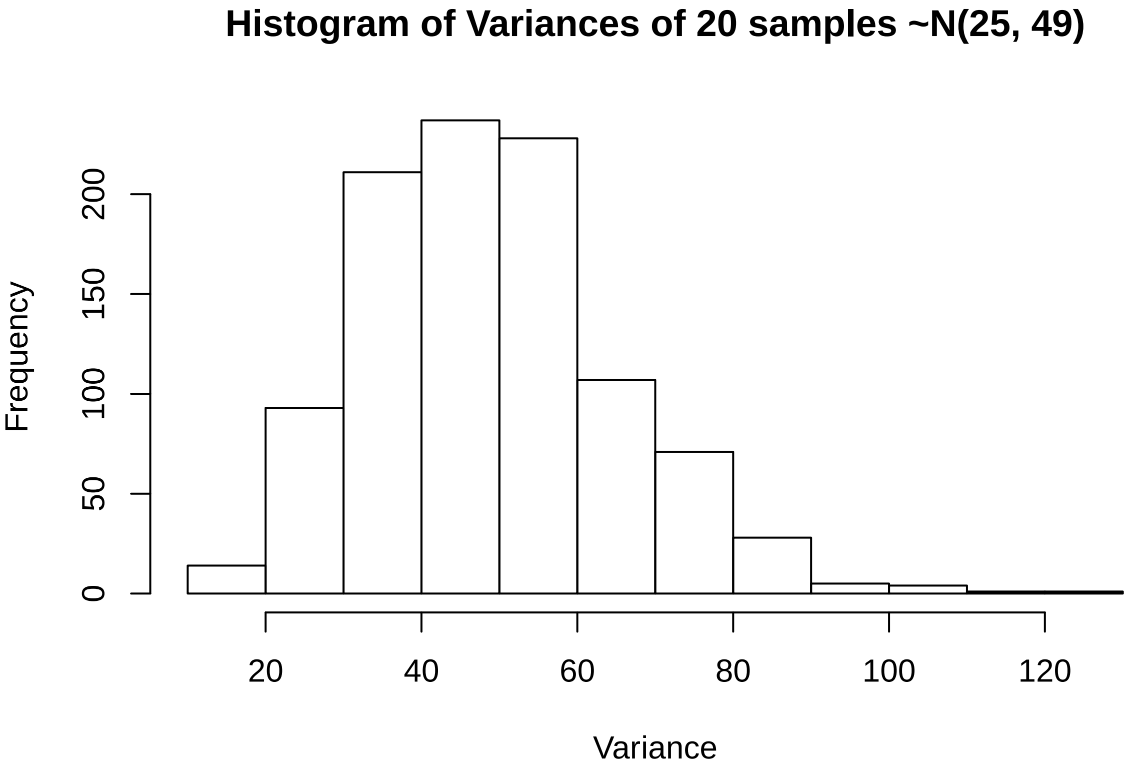
```
my.vars <- sapply(replicate(1000, list(rnorm(20, 25, 7))), function(x)var(x))

mean(my.vars)
```

```
## [1] 48.93441
```

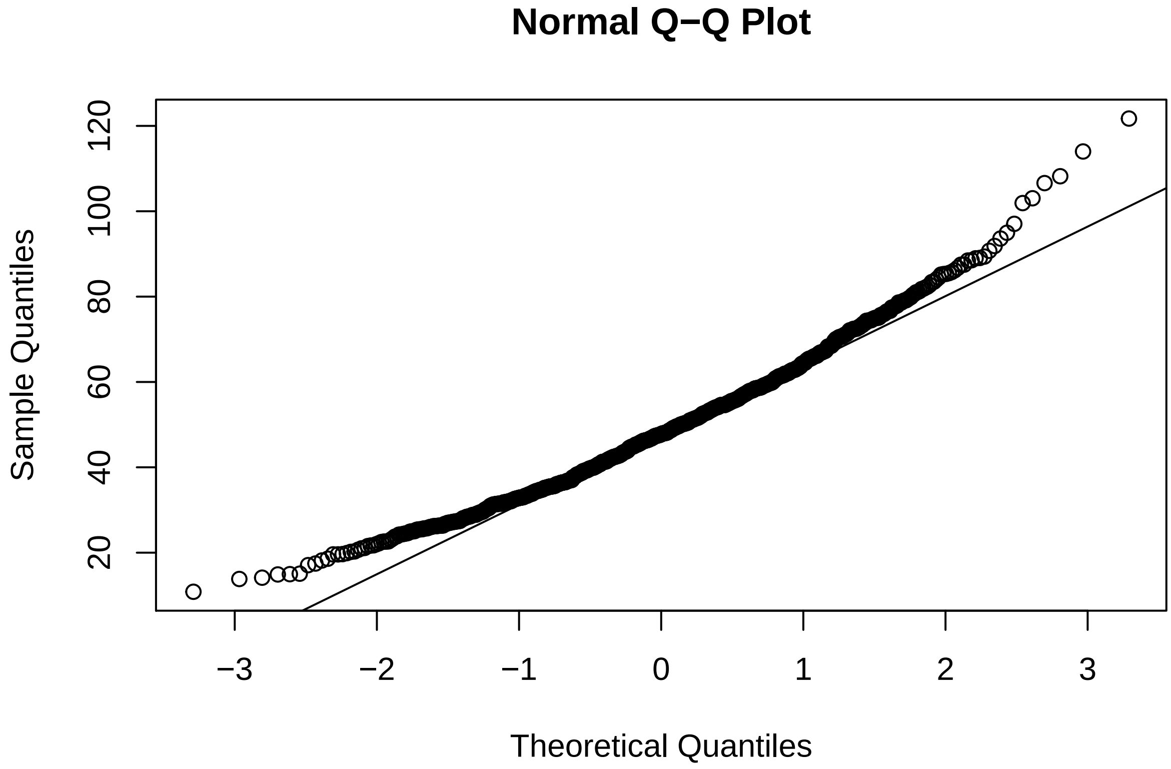
```
var(my.vars)
```

```
## [1] 259.7597
```

```
hist(my.vars, main = 'Histogram of Variances of 20 samples ~N(25, 49)', xlab = 'Variance')
```



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



For n = 50, the histogram and qqplot look to be normally-distributed, although it's not perfect, as the histogram shows a little right-skewness and the qqplot strays a bit from the normal line. The plots do, however, look to be closer to the normal distribution than those for n = 20.

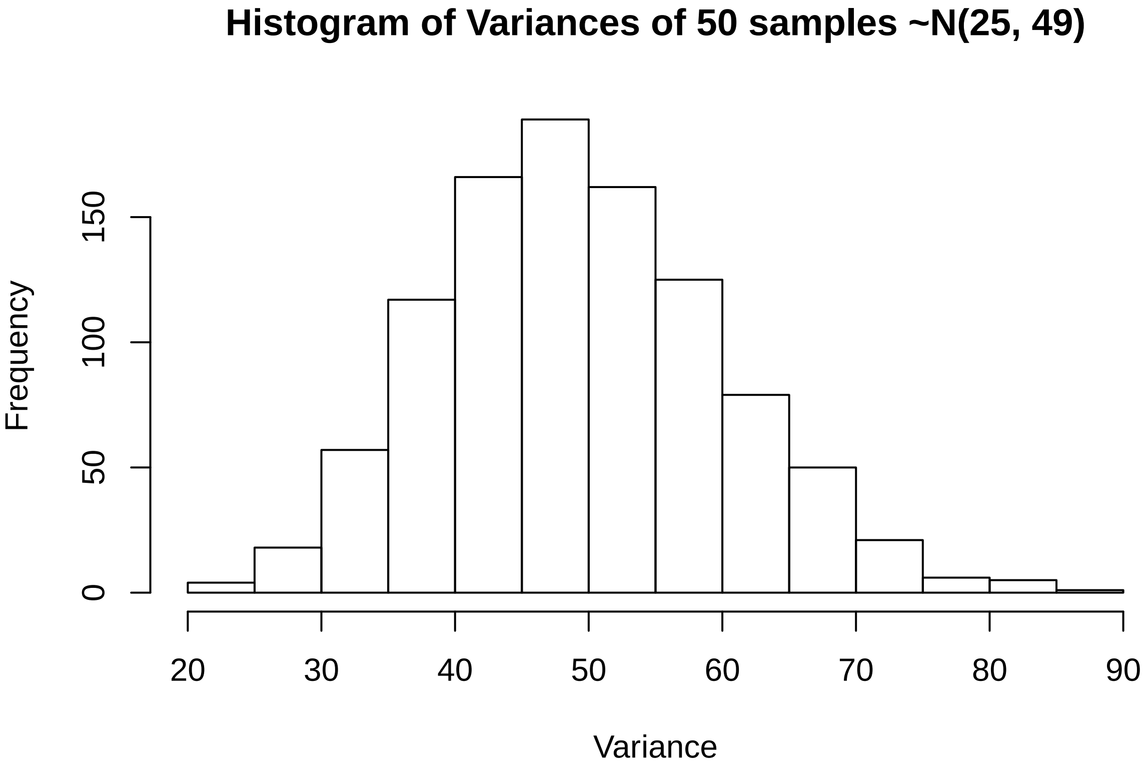
```
my.vars1 <- sapply(replicate(1000, list(rnorm(50, 25, 7))), function(x)var(x))

mean(my.vars1)
```

```
## [1] 49.38867
```

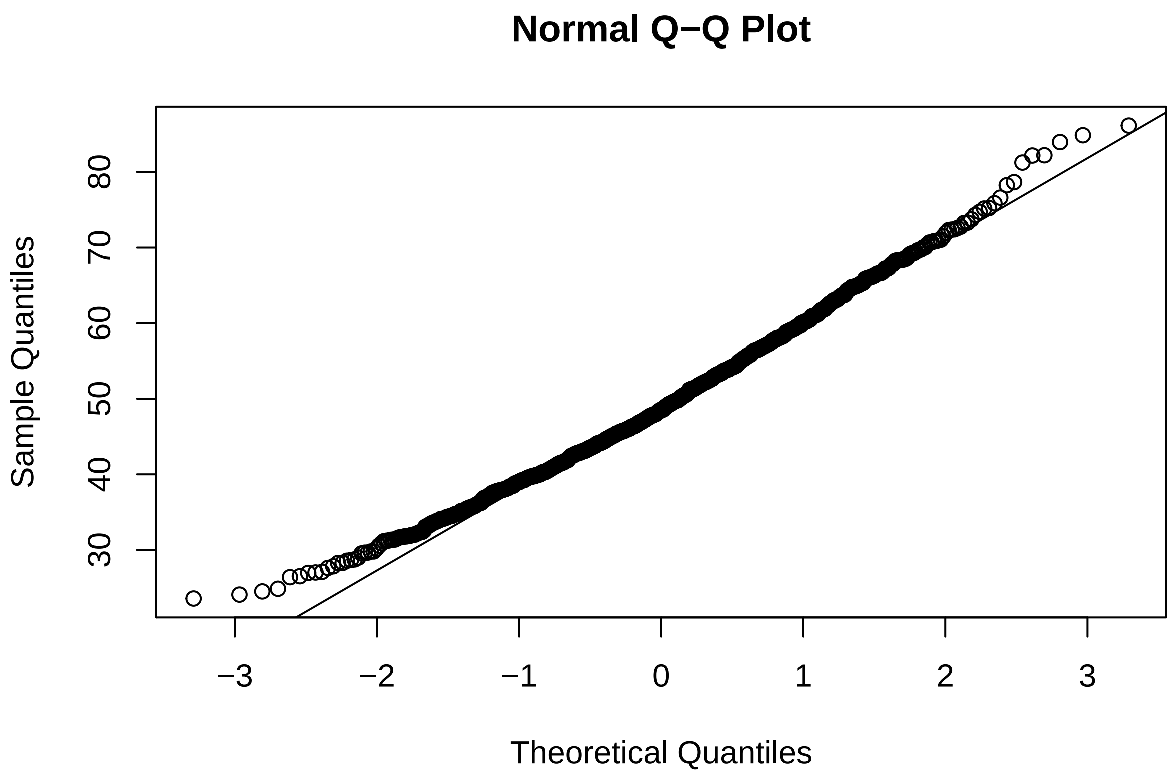
```
var(my.vars1)
```

```
## [1] 112.228
```

```
hist(my.vars1, main = 'Histogram of Variances of 50 samples ~N(25, 49)', xlab = 'Variance')
```



```
qqnorm(my.vars1)
qqline(my.vars1)
```



For n = 200, the histogram and qqplot certainly look normally-distributed. The histogram has a bell shape, and the qqplot barely varies from the imposed normal line. The plots don't look THAT much more normal than those for n = 50, though, which is good evidence that the distribution follows the aforementioned rule of thumb that says we only need 30 samples, in most cases, to approximate a normal distribution.

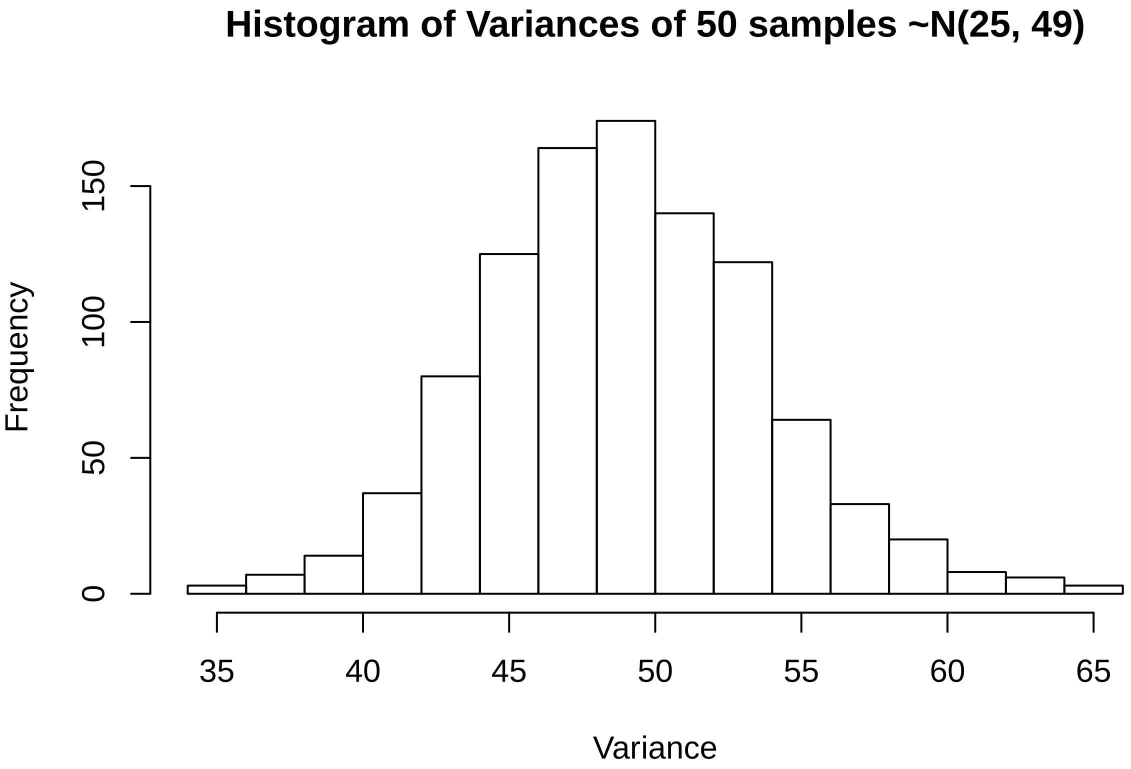
```
my.vars2 <- sapply(replicate(1000, list(rnorm(200, 25, 7))), function(x)var(x))

mean(my.vars2)
```

```
## [1] 48.94714
```

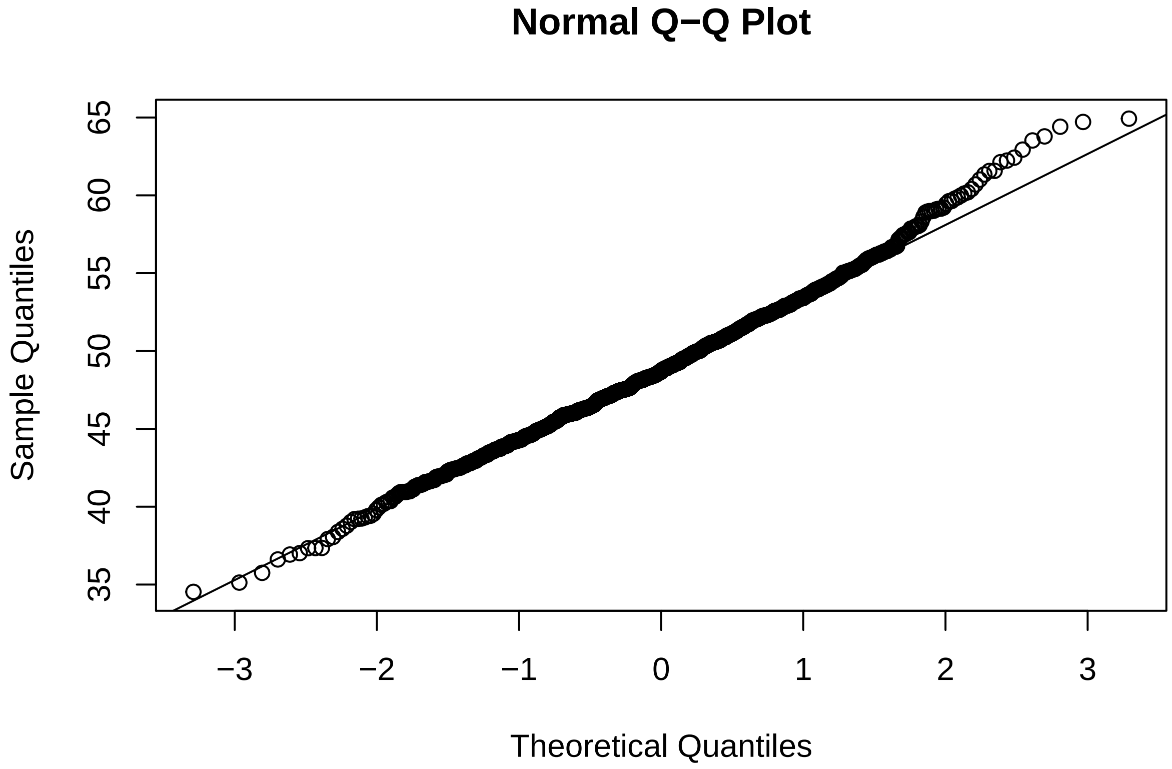
```
var(my.vars2)
```

```
## [1] 22.5001
```

```
hist(my.vars2, main = 'Histogram of Variances of 50 samples ~N(25, 49)', breaks = 20, xlab = 'Variance'
```



```
qqnorm(my.vars2)
qqline(my.vars2)
```



Problem 6

a) By Theorem A.4, $Var(\overline{x}) = \frac{\sigma^2}{n}$. So Var(X) = 9/9 = 1 and Var(Y) = 25/12 = 2.083. Then, by Theorem A.10, Var(X - Y) = Var(X) + Var(Y), and mean(X - Y) = mean(X) - mean(Y). So $\sigma^2(W) = 3.083$, $\sigma(W) \approx 1.756$, and $mean(W) = -3$. Putting this together, the sampling distribution of W is approximately W ~ N(-3, 1.756), by the Central Limit Theorem.

b) My simulation produced results very close to the theoretical expectations. I got $\sigma^2(W) \approx 3.1$, $\sigma(W) \approx 1.76$, and $mean(W) \approx -3.01$.

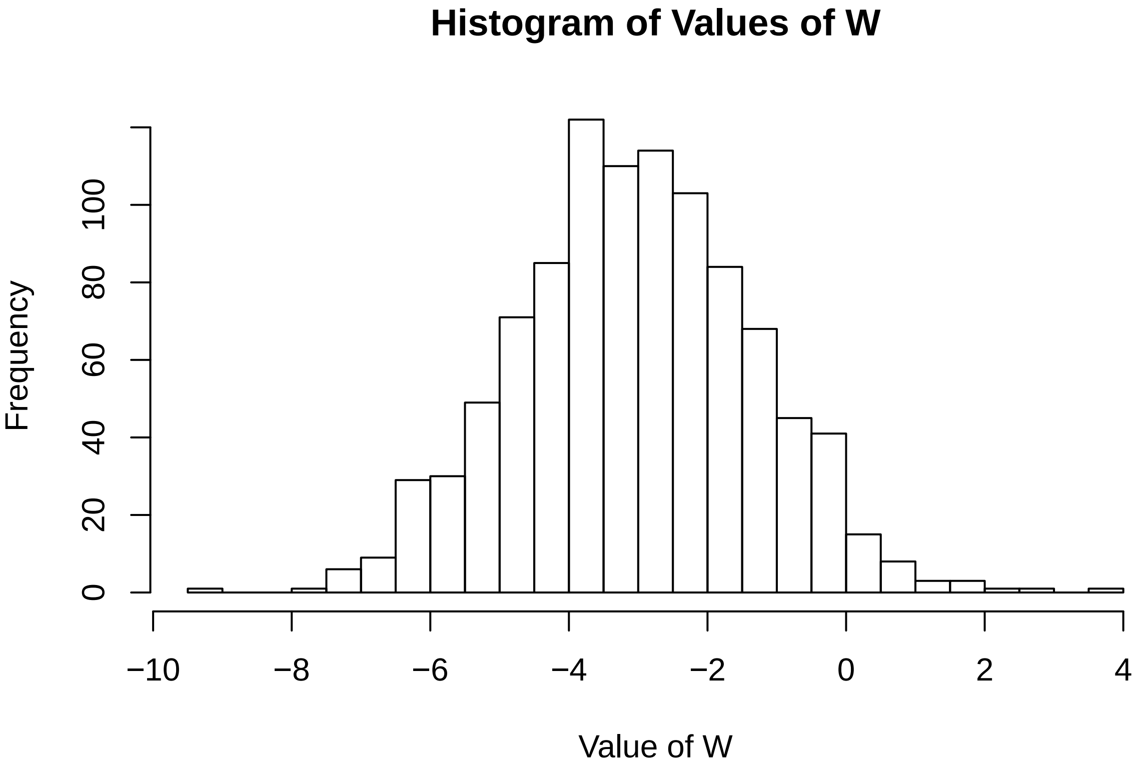
```
my.vars3 <- replicate(1000, mean(rnorm(9, 7, 3)) - mean(rnorm(12, 10, 5)))

mean(my.vars3)
```

```
## [1] -3.035411
```

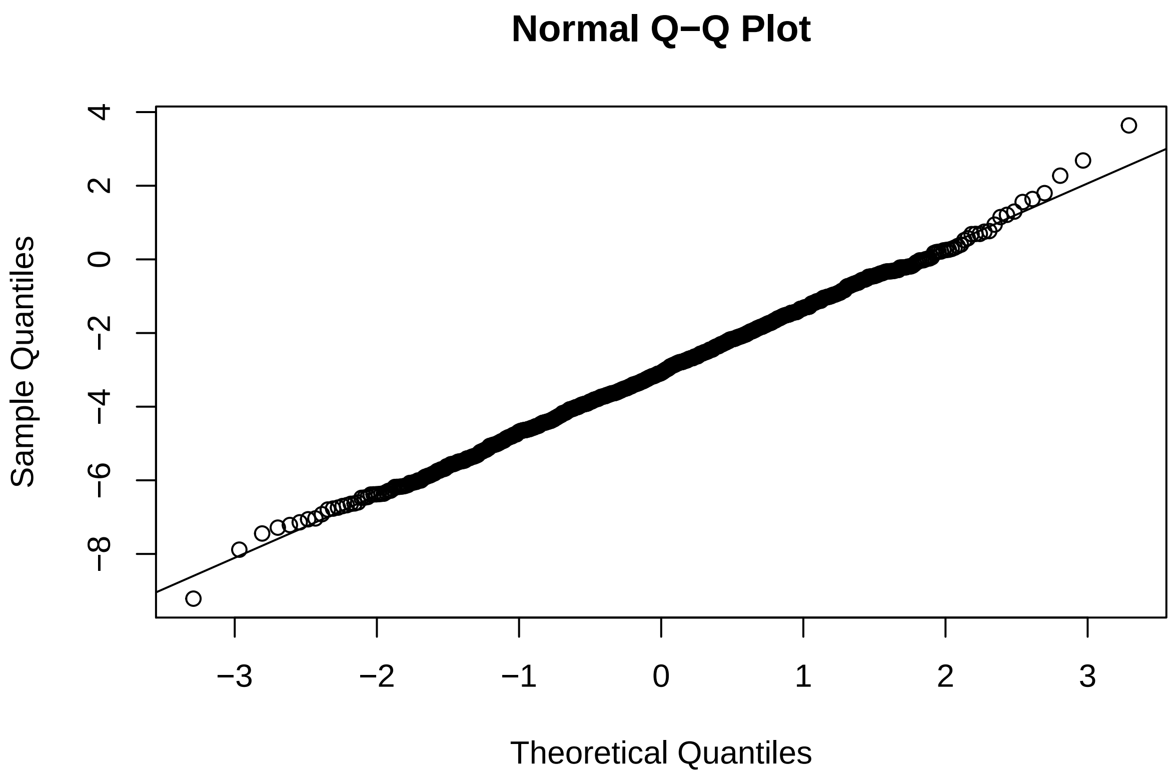
```
sd(my.vars3)
```

```
## [1] 1.705901
```

```
hist(my.vars3, main = 'Histogram of Values of W', breaks = 20, xlab = 'Value of W')
```



```
qqnorm(my.vars3)
qqline(my.vars3)
```



```
### C
length(my.vars3[my.vars3 < -1.5])/length(my.vars3)
```

```
## [1] 0.814
```

```
pnorm(-1.5, -3, 1.756)
```

```
## [1] 0.8035068
```

c) My simulation produced $P(W < -1.5) \approx .791$, which is very close to the theoretical value of $P(drawfromN(-3, 1.756^2) < -1.5) \approx .803$.

Problem 7

a) By the Central Limit Theorem, the sampling distribution of the test statistic $W = \overline{X} - \overline{Y}$ will be approximately normal for a large enough sample. The estimated parameters of the distribution will be, then, $\mu = E[\overline{X} - \overline{Y}] = E[\overline{X}] - E[\overline{Y}] = 1/2 \times (1 + 0) - 1/2 \times (1.5 + .5) = 1/2 - 1 = -.5$ and $\sigma^2 = Var(\overline{X}) + Var(\overline{Y}) = \frac{Var(X_i)}{n_X} + \frac{Var(Y_i)}{n_Y} = \frac{1/12}{9} + \frac{1/12}{12} = .0092 + .00694 \approx .01614$, both of which can be derived using the rules referenced in Problem 6.

So we'd expect W ~ N(-.5, .01614), i.e., with standard error $\sqrt(.01614) = .1271$.

b) The simulations and plots seem to agree with the theoretical approximations that W ~ N(-.5, .01614), with standard error $\approx .1271$.

```
my.vars4 <- replicate(10000, mean(runif(9)) - mean(runif(12, .5, 1.5)))

mean(my.vars4)
```

```
## [1] -0.4999648
```

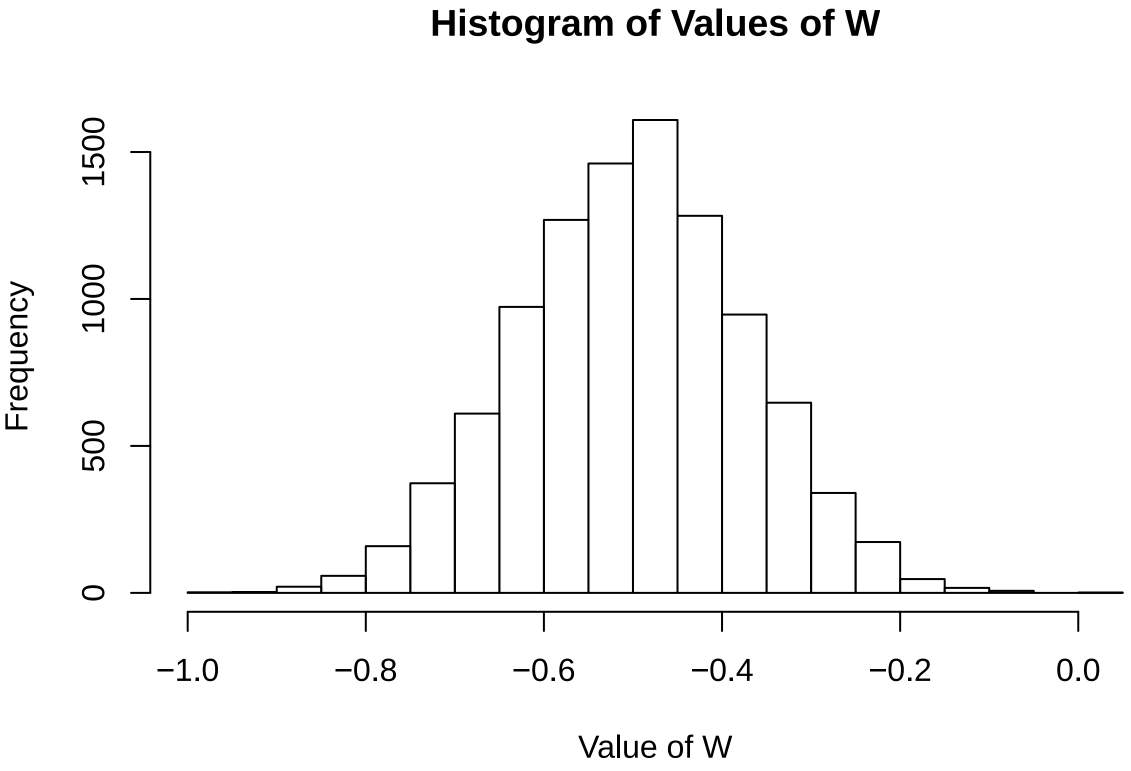
```
var(my.vars4)
```

```
## [1] 0.01625955
```

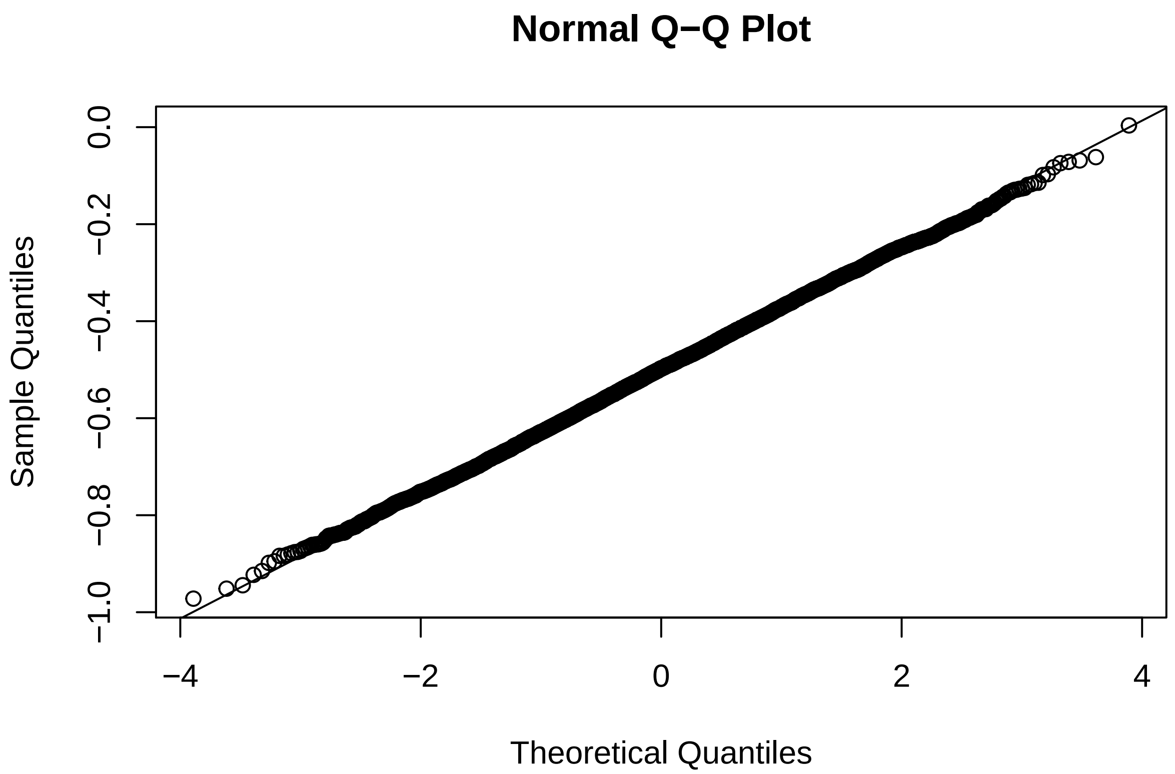
```
sd(my.vars4)
```

```
## [1] 0.1275129
```

```
hist(my.vars4, main = 'Histogram of Values of W', breaks = 20, xlab = 'Value of W')
```



```
qqnorm(my.vars4)
qqline(my.vars4)
```



```
### C - p-value = 1
length(my.vars4[my.vars4 < .6])/length(my.vars4)
```

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

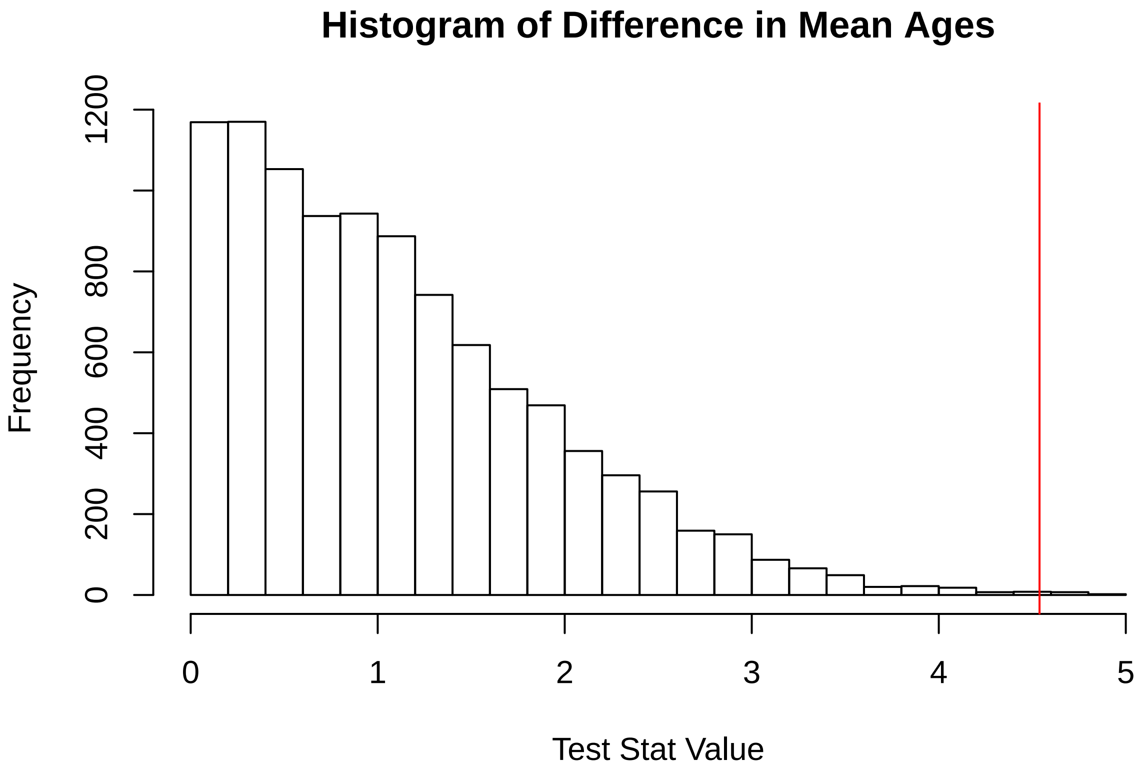
c) The range of values for W is [-1.5, .5], so every possible value is < .6. This is mirrored in the simulation above, which produces a p-value of 1.

Problem 8

There is strong reason to reject the null hypothesis that the mean ages of survivors and victims is the same. The simulation below produced a very small p-value for the null hypothesis of .0001, which is corroborated by the histogram.

For a two-sided test, we use the absolute value of the test statistic, since the hypotheses say nothing about *how* the means differ, just that they do.

```
ti <- read.csv('/Users/brodyvogel/Desktop/Data/Titanic.csv', sep = ',')

survivors <- ti[ti$Survived == 1, ]
casualties <- ti[ti$Survived == 0, ]

survAge <- mean(survivors$Age)
casAge <- mean(casualties$Age)

testStat3 <- abs(survAge - casAge)

testSTATS3 <- c()

for (x in 1:10000) {
  samp1 <- sample(nrow(ti), length(survivors$Age), replace = FALSE)
  helper3 <- ti[samp1, ]
  helper4 <- ti[-samp1, ]

  testSTATS3 <- c(testSTATS3, abs(mean(helper3$Age) - mean(helper4$Age)))
}

hist(testSTATS3, breaks = 20, main = 'Histogram of Difference in Mean Ages', xlab = 'Test Stat Value')
abline(v = testStat3, col = 'red')
```



```
p4 <- length(testSTATS3[testSTATS3 >= testStat3])/length(testSTATS3)

p4
```

```
## [1] 0.001
```