hw02

Victor Huang

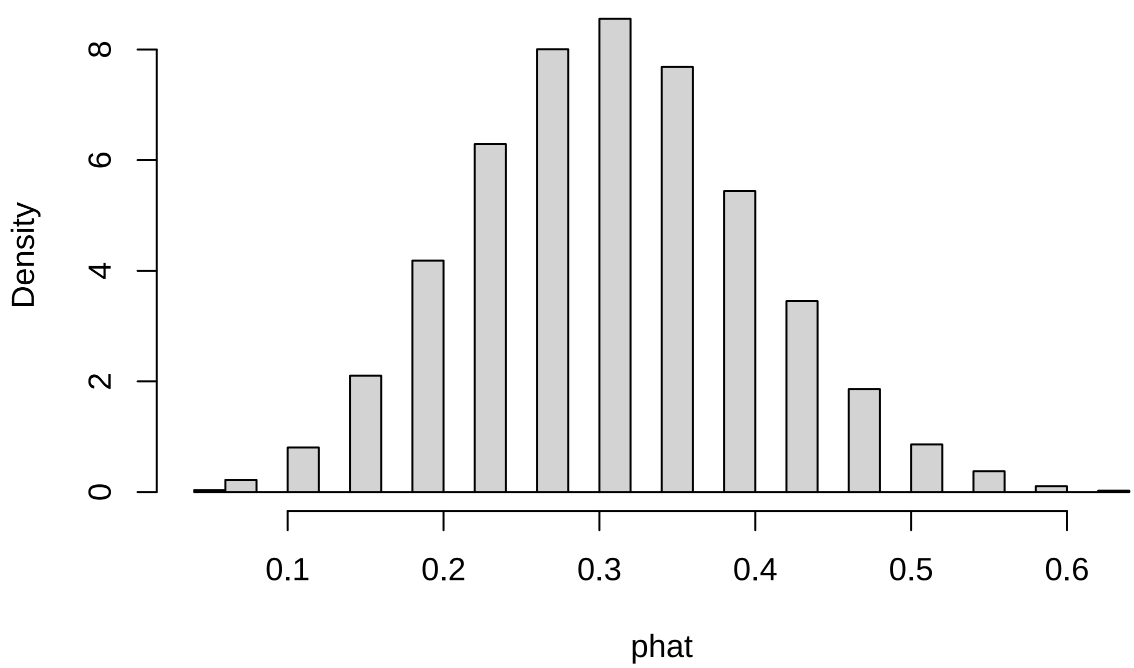
1/20/2022

Q1. (Exercise 4.6)

a) We see that the sampling distribution is normally distributed, with the estimate mean at 0.315 with estimate standard error being 0.0922

b) Using the following equation $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.316(1-0.316)}{25}} = 0.09298258$ we get that theoretical standard error to be 0.09298258. This is relatively close to our initial 0.092 estimated standard error

```r
set.seed(69420)
Recidivism <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Recidivism.csv")
N <- 10^4
phat <- numeric(N)
for (i in 1:N)
{
  samp <- sample(Recidivism$Recid, 25)
  phat[i] <- mean(samp == "Yes")
}
hist(phat, freq = FALSE, main = NULL, breaks = 30)
```



```r
mean(phat) #0.315
```

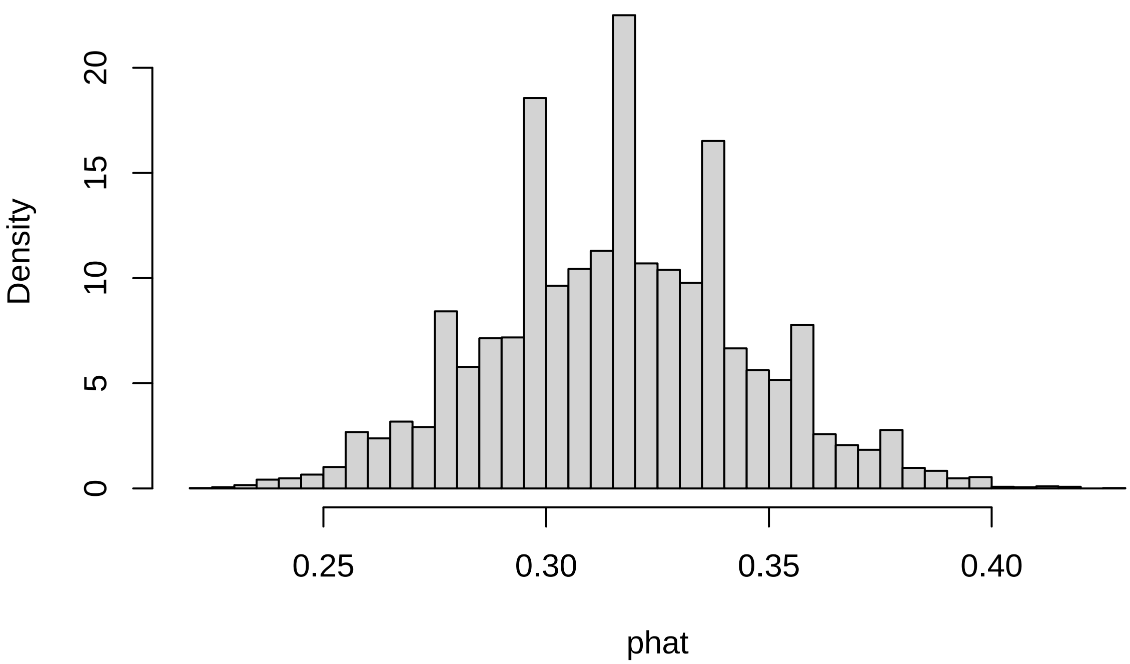
```
## [1] 0.317352
```

```r
sd(phat) #0.092
```

```
## [1] 0.09227394
```

c) Using the following equation $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.316(1-0.316)}{250}} = 0.02940367$ we get that theoretical standard error to be 0.02940367. This pair of theoretical standard error and estimated standard error is closer than the previous one with n = 25 cases.

```r
set.seed(69420)
Recidivism <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Recidivism.csv")
N <- 10^4
phat <- numeric(N)
for (i in 1:N)
{
  samp <- sample(Recidivism$Recid, 250)
  phat[i] <- mean(samp == "Yes")
}
hist(phat, freq = FALSE, main = NULL, breaks = 30)
```




```r
mean(phat)
```

```
## [1] 0.3169336
```

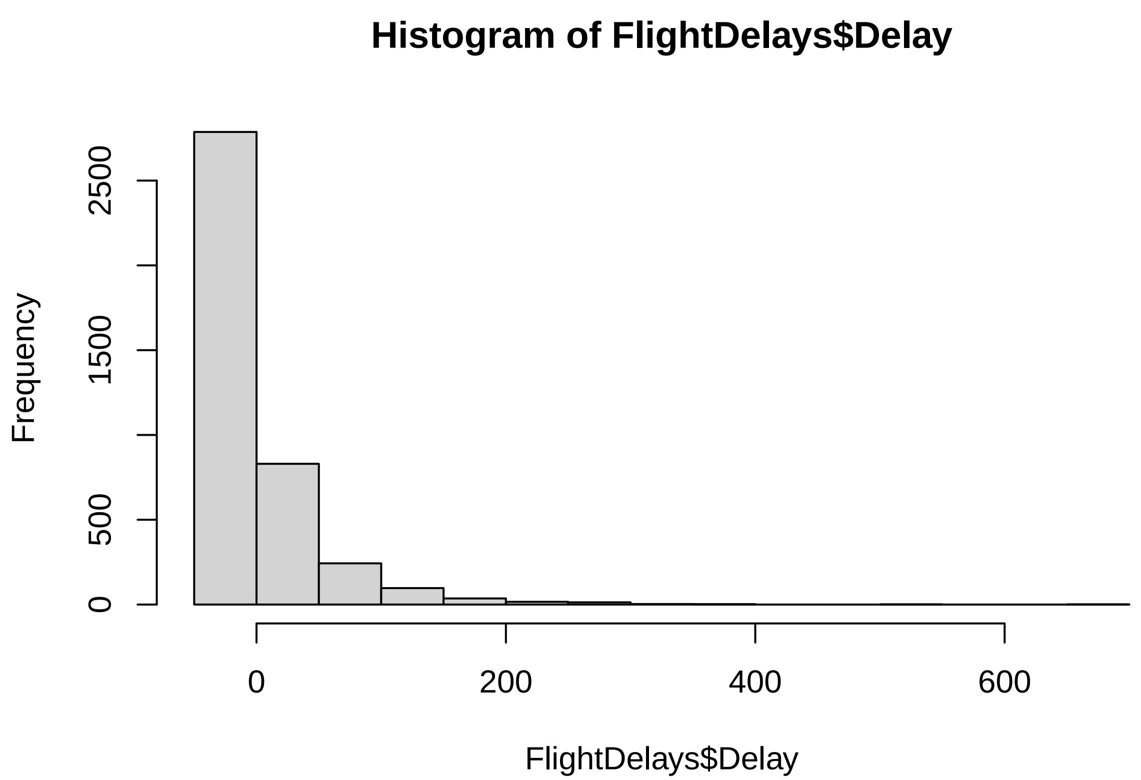
```r
sd(phat)
```

```
## [1] 0.02905113
```

Q2. (Exercise 4.7)

a) Looking at the histogram below, we see that the data is skewed to the right, We get a mean of 11.7379 and a standard deviation 41.6305

```r
FlightDelays <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/FlightDelays.csv")
hist(FlightDelays$Delay)
```




```r
mean(FlightDelays$Delay)
```

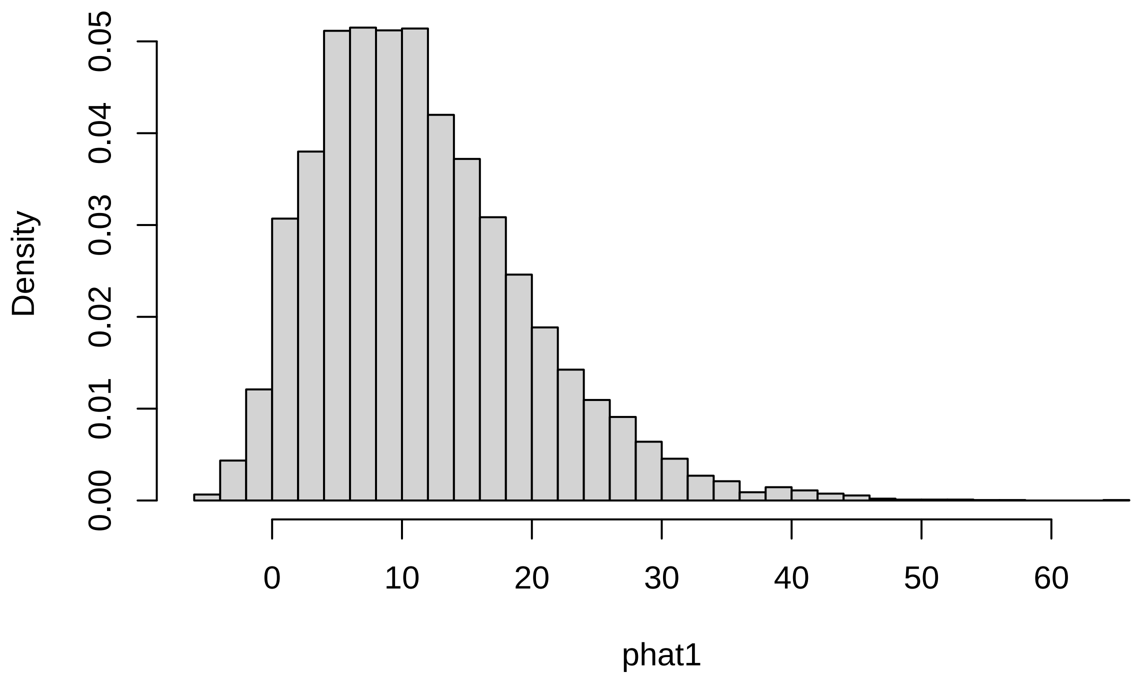
```
## [1] 11.7379
```

```r
sd(FlightDelays$Delay)
```

```
## [1] 41.6305
```

b) Looking at the histogram below, we see that the simulated sampling distribution is skewed to the right, We get an estimated mean of 11.64506 and a standard error 8.281396.

```r
set.seed(69420)
FlightDelays <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/FlightDelays.csv")
N <- 10^4
phat1 <- numeric(N)
for (i in 1:N)
{
samp <- sample(FlightDelays$Delay, 25)
phat1[i] <- mean(samp)
}
hist(phat1, freq = FALSE, main = NULL, breaks = 30)
```




```r
mean(phat1)
```

```
## [1] 11.64506
```

```r
sd(phat1)
```

```
## [1] 8.281396
```

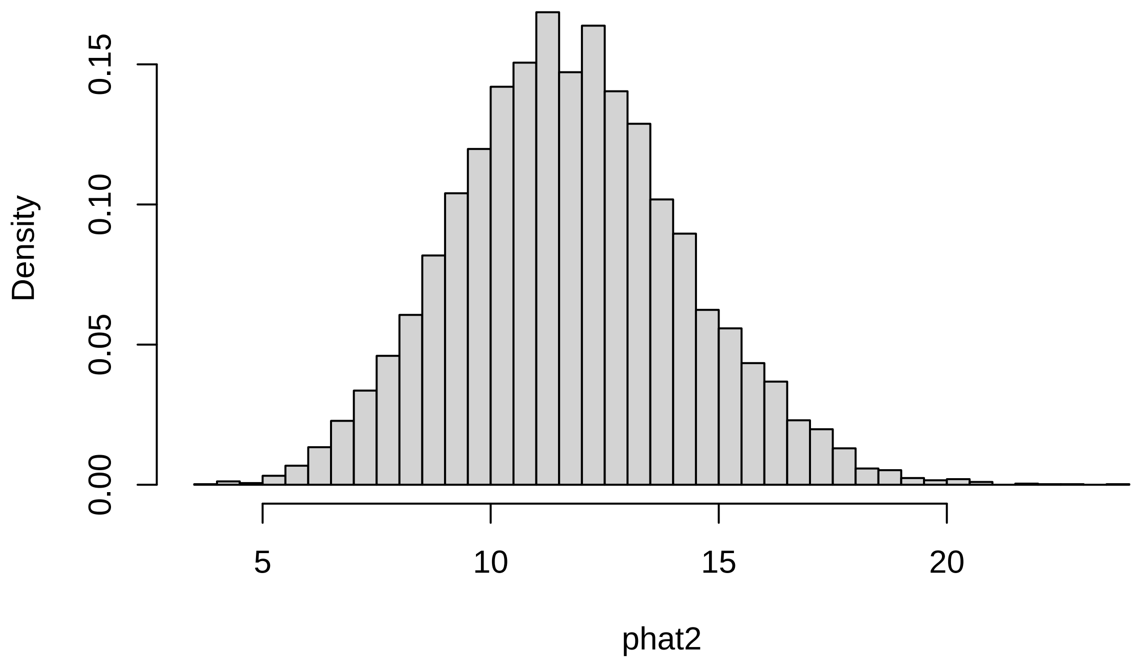
c) Our estimated standard error of 8.281396 is close to our theoretical standard error of 8.326099

```r
TSE <- sqrt(var(FlightDelays$Delay)/25)
TSE
```

```
## [1] 8.326099
```

d) Our estimated standard error is now 2.53198, which is slightly closer to the theoretical standard error of 2.632944 compared to when n = 25

```
set.seed(69420)
FlightDelays <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/FlightDelays.csv")
N <- 10^4
phat2 <- numeric(N)
for (i in 1:N)
{
samp <- sample(FlightDelays$Delay, 250)
phat2[i] <- mean(samp)
}
hist(phat2, freq = FALSE, main = NULL, breaks = 30)
```




```
mean(phat2)
```

```
## [1] 11.75512
```

```
sd(phat2)
```

```
## [1] 2.53198
```

```
TSE1 <- sqrt(var(FlightDelays$Delay)/250)
TSE1
```

```
## [1] 2.632944
```

Q3. (Exercise 4.12)

Looking at the calculations below, we see that $P(\hat{X} \leq 4.6) = 0.02385744$

```r
n <- 20
mean <- 6
variance <- 10
val <- (4.6 - 6)/(sqrt(10)/sqrt(20))
pnorm(val)
```

```
## [1] 0.02385744
```

Q4. (Exercise 4.20)

a)

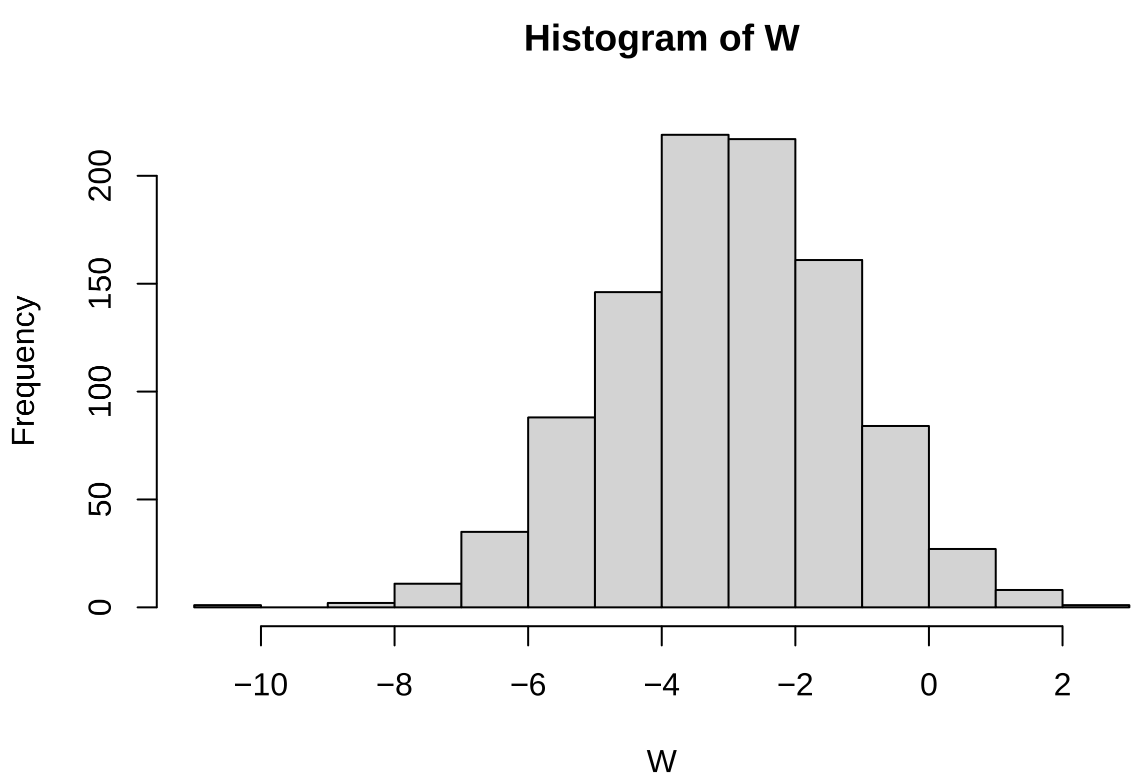
$$E(\bar{Z}) = E(\bar{X}) - E(\bar{Y}) = 7 - 10 = -3$$

$$Var(\bar{Z}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{3^2}{9} + \frac{5^2}{12} = 1 + \frac{25}{12} = \frac{37}{12} \approx 3.083$$

Because of the Central Limit Theorem, we can get that $W \sim N(-3, 3.083)$

b) We get simulated mean and standard error as -3.046 and 3.051, these are close to the theoretical mean and stdard error of -3 and 3.083

```r
set.seed(69420)
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)
```



```r
mean(W)
```

```
## [1] -3.046227
```

```r
sum((W - mean(W))^2/ length(W))
```

```
## [1] 3.050577
```

c) We get that $P(W < -1.5) = 0.82$, our exact answer is 0.804, which is pretty close to our estimated answer

```r
prob<- (-1.5 + 3)/(sqrt(3.083))
bleh <- pnorm(prob)
bleh
```

```
## [1] 0.8035274
```

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

```
## [1] 0.82
```

Q5. (Exercise 4.14)

We get the probability that between 220 and 230 people, 29.6% of people will have a high school diploma

```r
n <- 800
p <- 0.286
pbinom(220, 800, 0.286)
```

```
## [1] 0.259103
```

```r
pbinom(230, 800, 0.286)
```

```
## [1] 0.5550674
```

```r
pbinom(230, 800, 0.286) - pbinom(220, 800, 0.286)
```

```
## [1] 0.2959644
```

Q6. (Exercise 4.28)

$$f(x) = 3x^2$$

$$P(X \leq x) = \int_0^x 3x^2 dx$$

$$P(X \leq x) = x^3$$

a)

$$f(x_{min}) = n[1 - x_1^n]^{n-1} f(x_n)$$
$$f(x_{min}) = n[1 - x_1^3]^{n-1} 3x_n^2$$
$$f(x_{min}) = 3n[1 - x_1^3]^{n-1} x_n^2 (0 \leq x \leq 1)$$

b)

$$f(x_{max}) = n[f(x_n)]^{n-1} f(x_n)$$
$$f(x_{max}) = 3n \cdot x_n^{n-1} x_n^2$$
$$f(x_{max}) = 3n \cdot x_n^{3n-1} (0 \leq x_n \leq 1)$$

c)

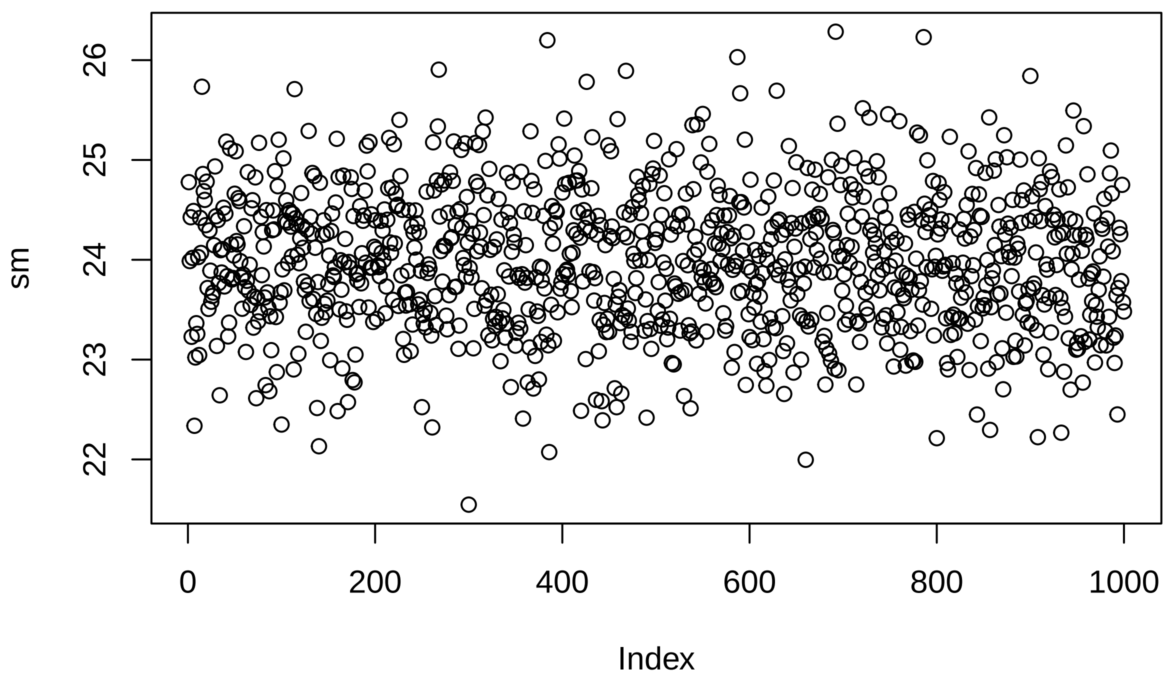
$$P[x_{max} > 0.92] = \int_{0.92}^{1} 3 \cdot 10 x_n^{3 \cdot 10 - 1} dx_n$$
$$\ldots$$

$$P[x_{max} > 0.92] = 1 - (0.92)^{30}$$

Q7. (Exercise 5.9)

a) The plot seems normally distributed.

```
set.seed(69420)
sm <- numeric(1000)
for (i in 1:1000)
{
sm[i] = mean(rgamma(200,6,0.25))
}
plot(sm)
```



```r
mean(sm)
```

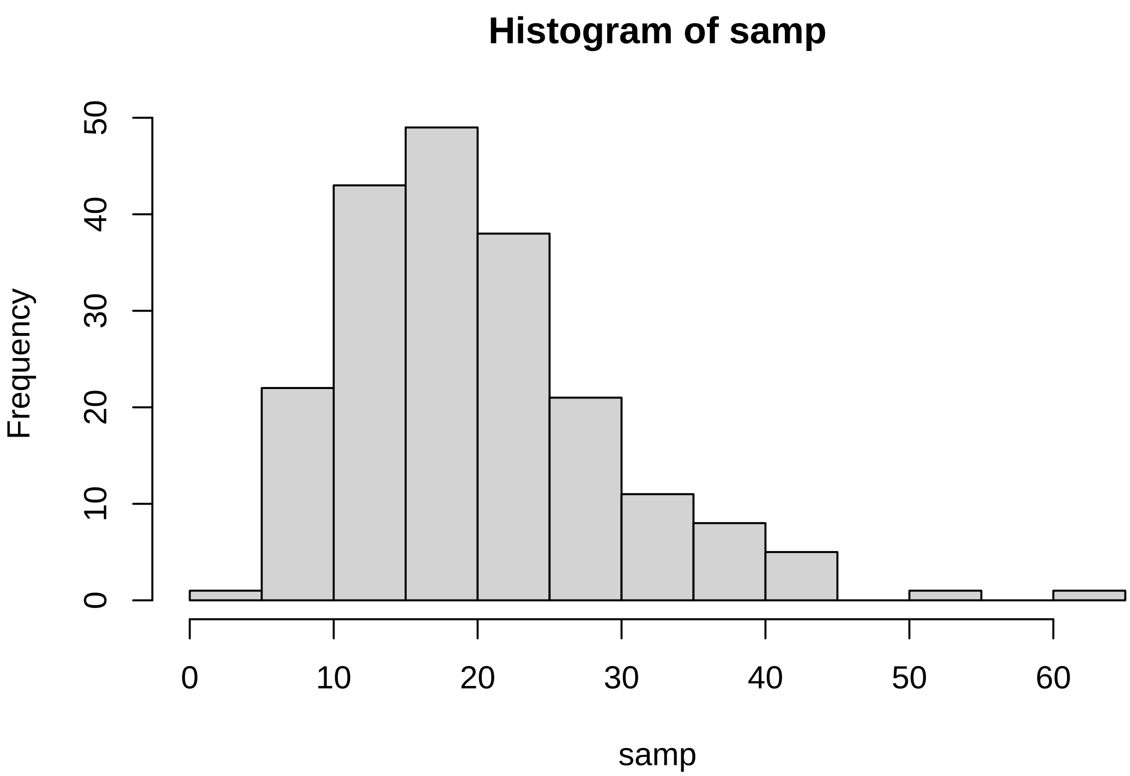
```
## [1] 23.98305
```

```r
sd(sm)
```

```
## [1] 0.698026
```

b) We get a slightly right skewed histogram with estimated mean at 20 and estimated standard error at 9.36

```r
set.seed(69420)
samp <- rgamma(200,5,0.25)
hist(samp)
```




```r
mean(samp)
```

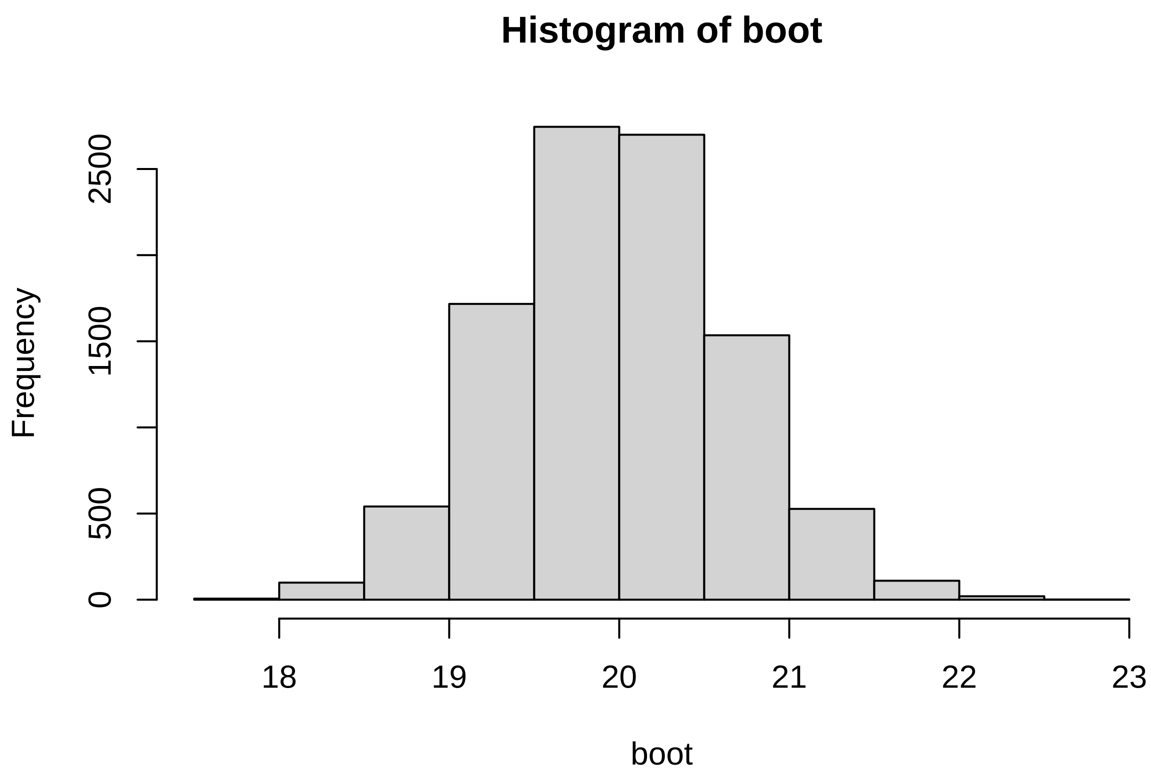
```
## [1] 20.00015
```

```r
sd(samp)
```

```
## [1] 9.359762
```

c) Bootstrap mean is 19.98796 and standard error is 0.665884

```
N <- 10^4
boot<- numeric(N)
for(i in 1:N){
  boot[i] <- mean(sample(samp, 200, replace = TRUE))
}
hist(boot)
```



d) The mean for sampling distribution and bootstrap distribution are very close (approximately 24 for both). The standard error is a bit different with smapling distribution having a 0.6881 SE and bootstrap distribution having an 0.221 SE.

```
data.frame(
    Distribution = c("Population", "Sampling distribution", "Sample",
                     "Bootstrap distribution"),
    Mean = c(20, 20,mean(sm), mean(boot)),
    SE = c(9, 0.6,sd(sm), sd(boot))
)
```

```
##             Distribution     Mean        SE
## 1             Population 20.00000 9.0000000
## 2  Sampling distribution 20.00000 0.6000000
## 3                 Sample 23.98305 0.6980260
## 4 Bootstrap distribution 19.98770 0.6649944
```

e) As the sample size gets smaller, shilw the means stay about the same, the SE gets significantly worse and more different as sample size decreases

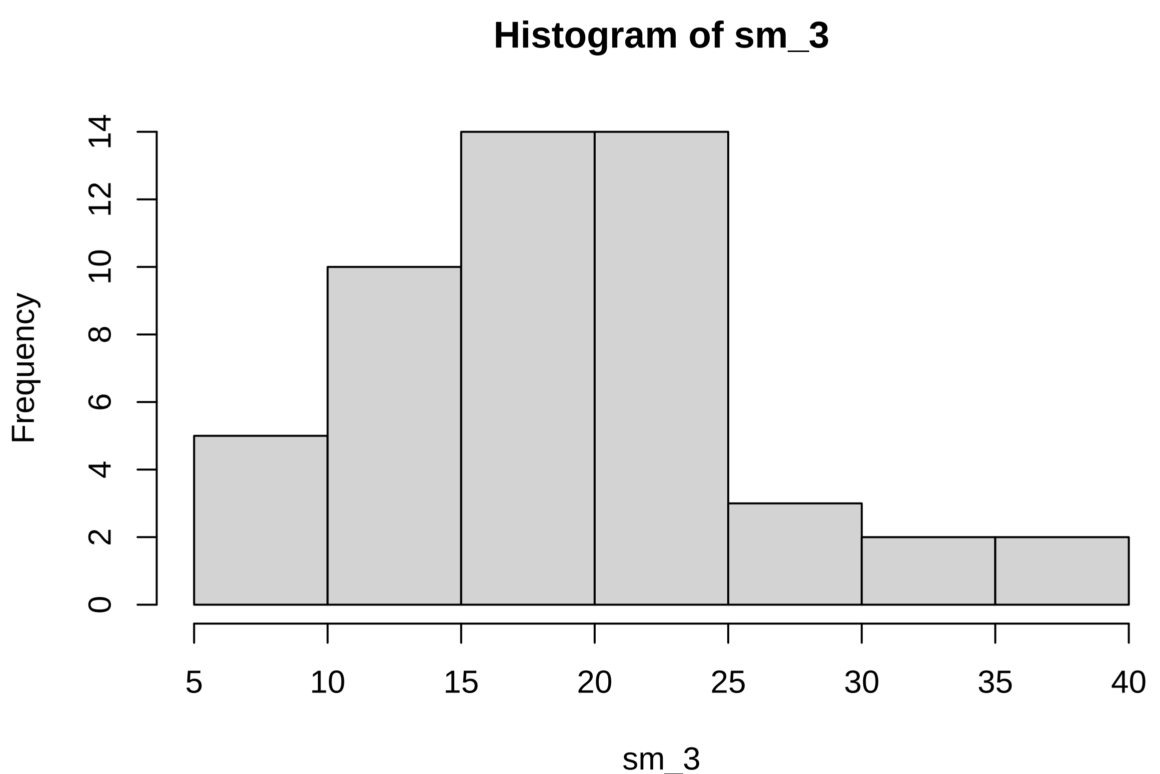
```r
set.seed(69420)
sm_1 <- numeric(1000)
for (i in 1:1000)
{
sm_1[i] = mean(rgamma(50,5,0.25))
}
mean(sm_1)
```

```
## [1] 20.03551
```

```r
sd(sm_1)
```

```
## [1] 1.245621
```

```r
sm_3 <- rgamma(50,5,0.25)
hist(sm_3)
```



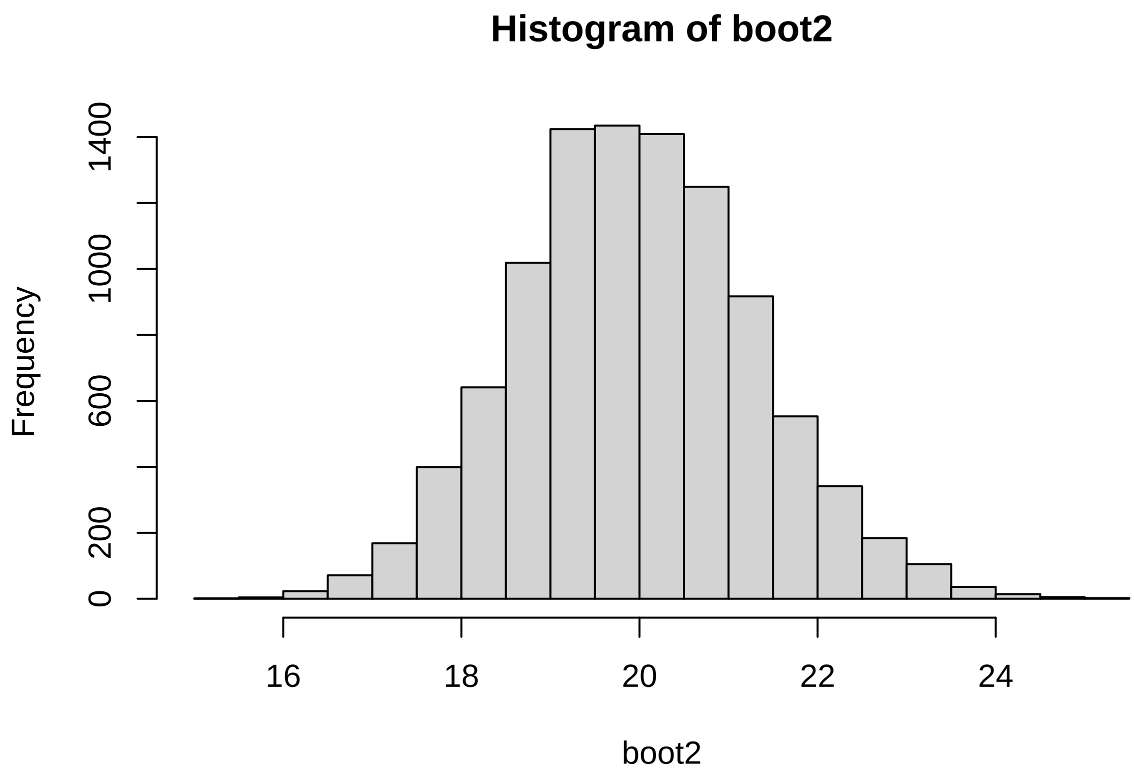

```r
mean(sm_3)
```

```
## [1] 18.77944
```

```r
sd(sm_3)
```

```
## [1] 7.337662
```

```r
N <- 10^4
boot2<- numeric(N)
for(i in 1:N){
  boot2[i] <- mean(sample(samp, 50, replace = TRUE))
}
hist(boot2)
```




```r
mean(boot2)
```

```
## [1] 19.97334
```

```r
sd(boot2)
```

```
## [1] 1.331857
```

```r
data.frame(
    Distribution = c("Population", "Sampling distribution", "Sample",
                     "Bootstrap distribution"),
    Mean = c(20, 20,mean(sm_3), mean(boot2)),
    SE = c(9, 1.262298,sd(sm_3), sd(boot2))
)
```

```
##             Distribution     Mean       SE
## 1             Population 20.00000 9.000000
## 2  Sampling distribution 20.00000 1.262298
## 3                 Sample 18.77944 7.337662
## 4 Bootstrap distribution 19.97334 1.331857
```

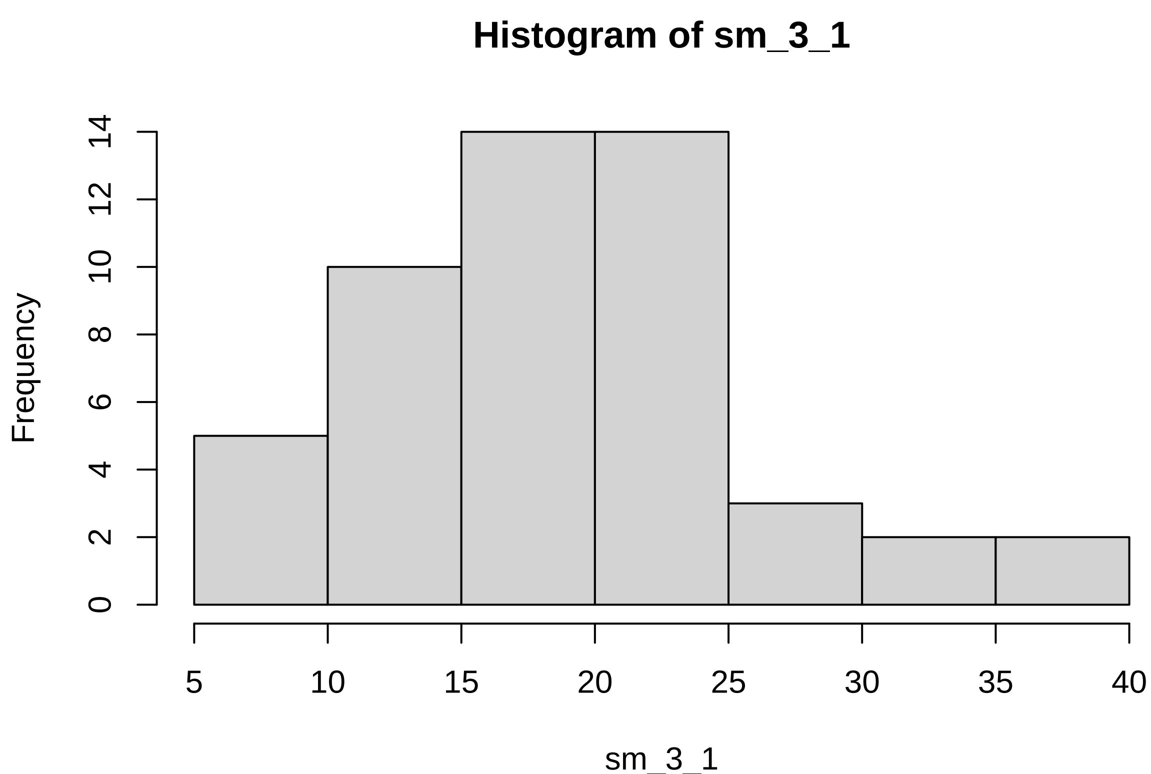
```r
set.seed(69420)
sm_1_1 <- numeric(1000)
for (i in 1:1000)
{
sm_1_1[i] = mean(rgamma(50,5,0.25))
}
mean(sm_1_1)
```

```
## [1] 20.03551
```

```r
sd(sm_1_1)
```

```
## [1] 1.245621
```

```r
sm_3_1 <- rgamma(50,5,0.25)
hist(sm_3_1)
```



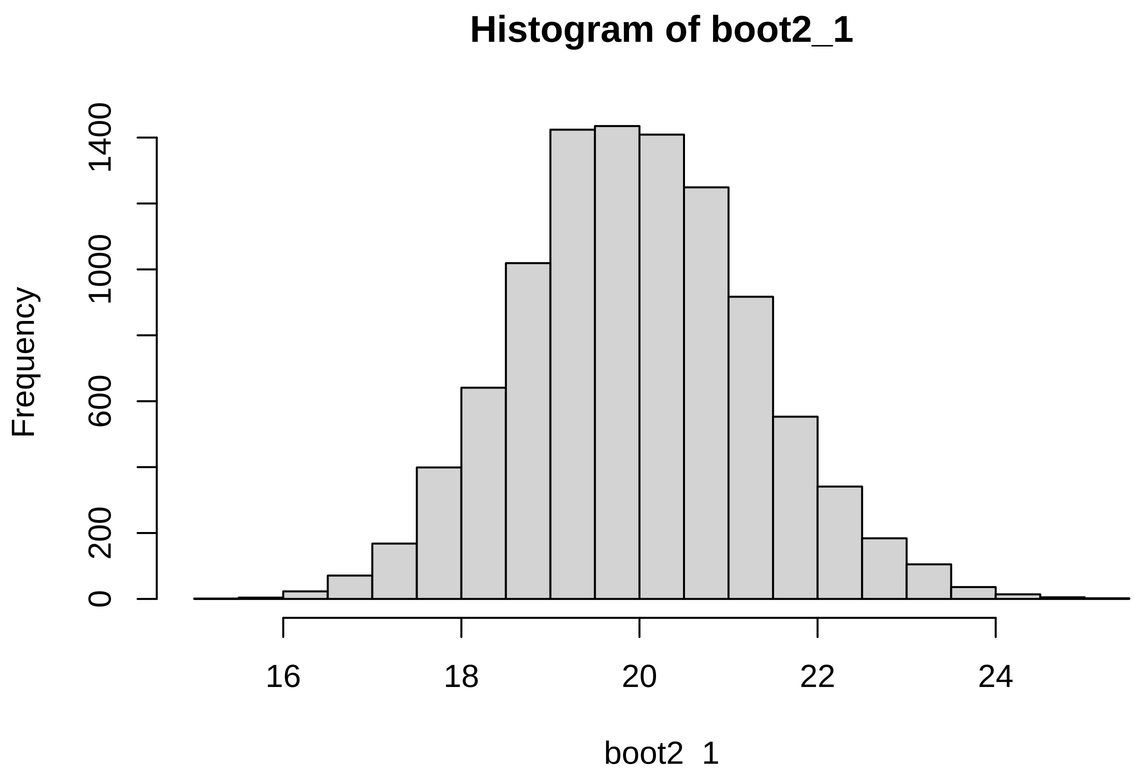

```r
mean(sm_3_1)
```

```
## [1] 18.77944
```

```r
sd(sm_3_1)
```

```
## [1] 7.337662
```

```r
N <- 10^4
boot2_1<- numeric(N)
for(i in 1:N){
  boot2_1[i] <- mean(sample(samp, 50, replace = TRUE))
}
hist(boot2_1)
```




```r
mean(boot2_1)
```

```
## [1] 19.97334
```

```r
sd(boot2_1)
```

```
## [1] 1.331857
```

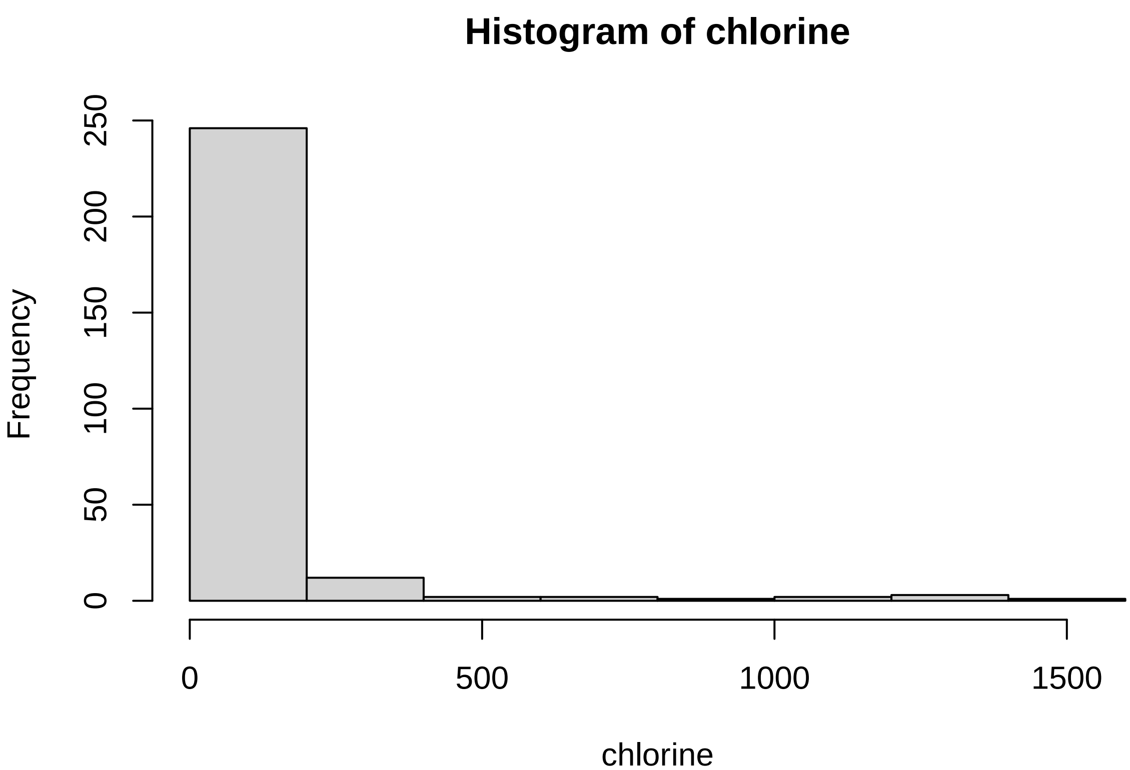
```r
data.frame(
    Distribution = c("Population", "Sampling distribution", "Sample",
                     "Bootstrap distribution"),
    Mean = c(20, 20,mean(sm_3_1), mean(boot2_1)),
    SE = c(9, 1.262298,sd(sm_3_1), sd(boot2_1))
)
```

```
##             Distribution     Mean       SE
## 1             Population 20.00000 9.000000
## 2  Sampling distribution 20.00000 1.262298
## 3                 Sample 18.77944 7.337662
## 4 Bootstrap distribution 19.97334 1.331857
```

Q8. (Exercise 5.11)

a) The histogram is extremely right skewed with a boxplot that fits that as well. The summary for the chlorine consist of min = 1.00, 1st quantile = 5.00, median = 14.05, Mean = 78.31, 3rd quantile = 55.70, and max = 1550.00

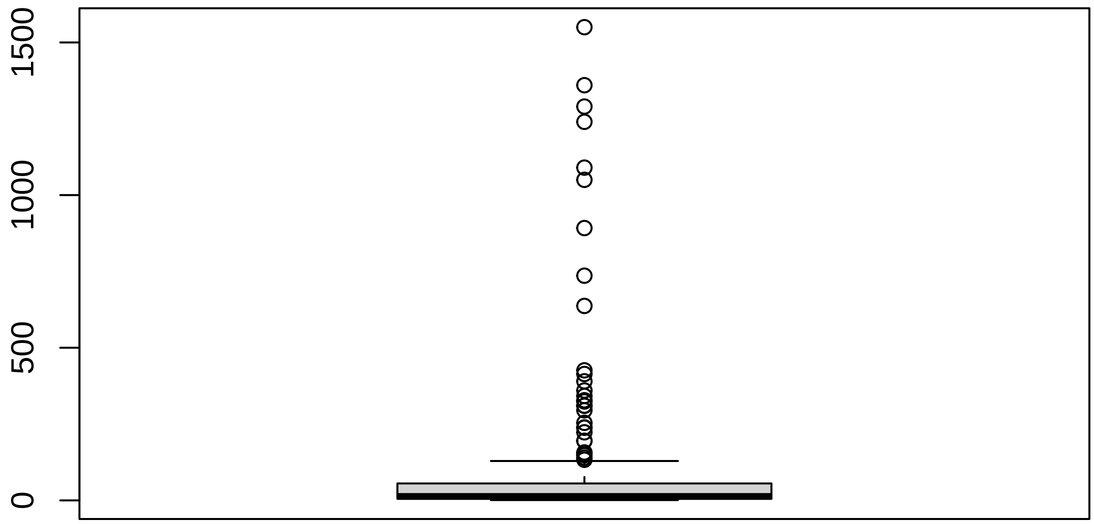
```
set.seed(69420)
Bangladesh <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Bangladesh.csv")
chlorine <-subset(Bangladesh, select = Chlorine, subset = !is.na(Chlorine), drop = T)
hist(chlorine)
```



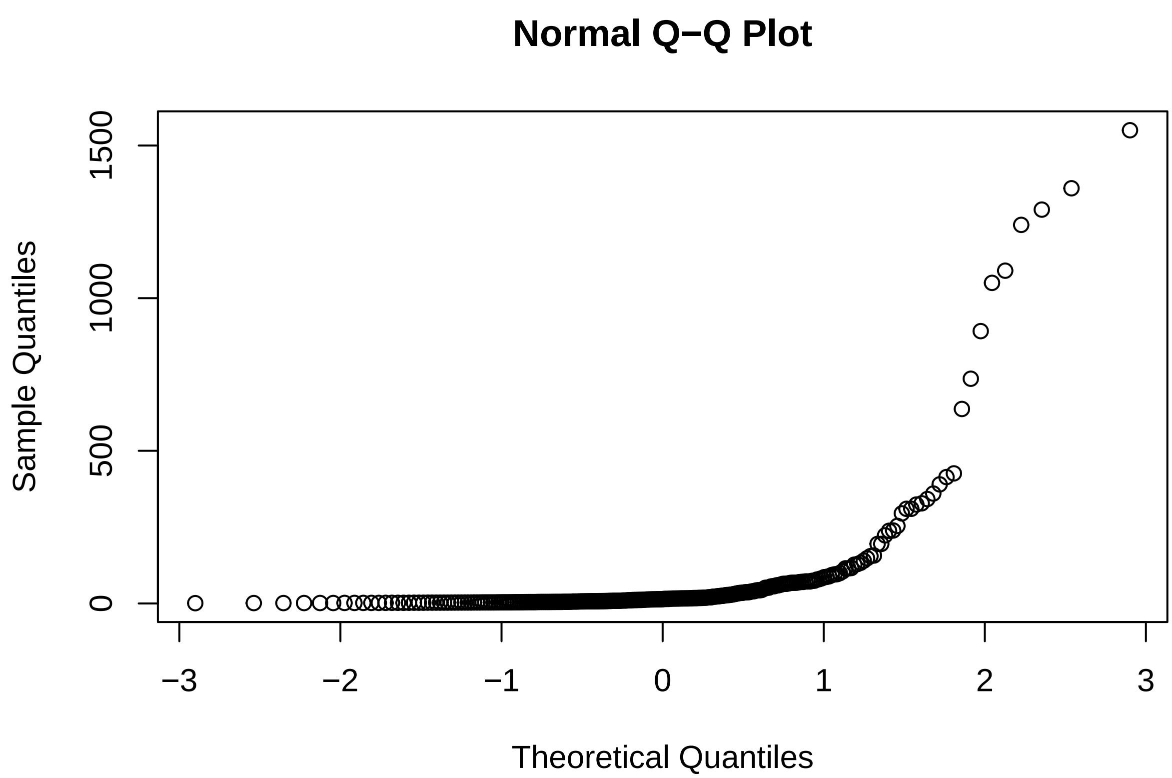

```
summary(chlorine)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    5.00   14.20   78.08   55.50 1550.00
```

```
boxplot(chlorine)
```



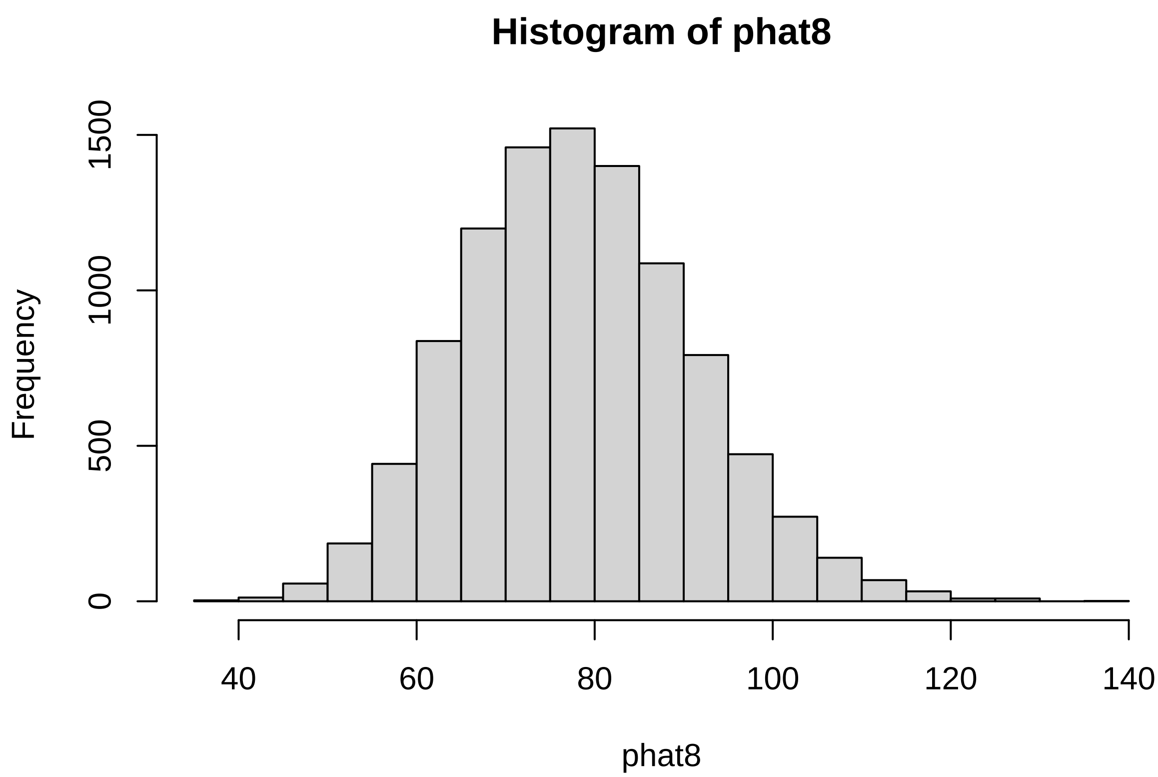

```
qqnorm(chlorine)
```



b) Bootstrap is shown below

```
set.seed(69420)
N <- 10^4
phat8 <- numeric(N)
n <- nrow(Bangladesh)
for (i in 1:N)
{
  samp8 <- sample(Bangladesh$Chlorine, size = n, replace = TRUE)
```

```r
  phat8[i] <- mean(samp8, na.rm = TRUE)
}
hist(phat8)
```




```r
mean(phat8)
```

```
## [1] 78.22395
```

```r
sd(phat8)
```

```
## [1] 12.9174
```

c) We get a 95% confidence interval from 54.92034 to 105.20690 What this means is that we are 95% confident that the mean chlorine concentration from the 271 wells in the data set will fall between those two values inclusive.

```r
quantile(phat8, probs = c(0.025,0.975))
```

```
##       2.5%     97.5%
##   54.92034 105.20690
```

d) The bootstrap estimated of the bias is -0.1399, this represents -0.010 of the standard error

```r
chlorine <- subset(Bangladesh, select = Chlorine, subset = !is.na(Chlorine), drop = T)
bias <- mean(chlorine) - mean(phat8)
bias / sd(phat8)
```

```
## [1] -0.01083329
```

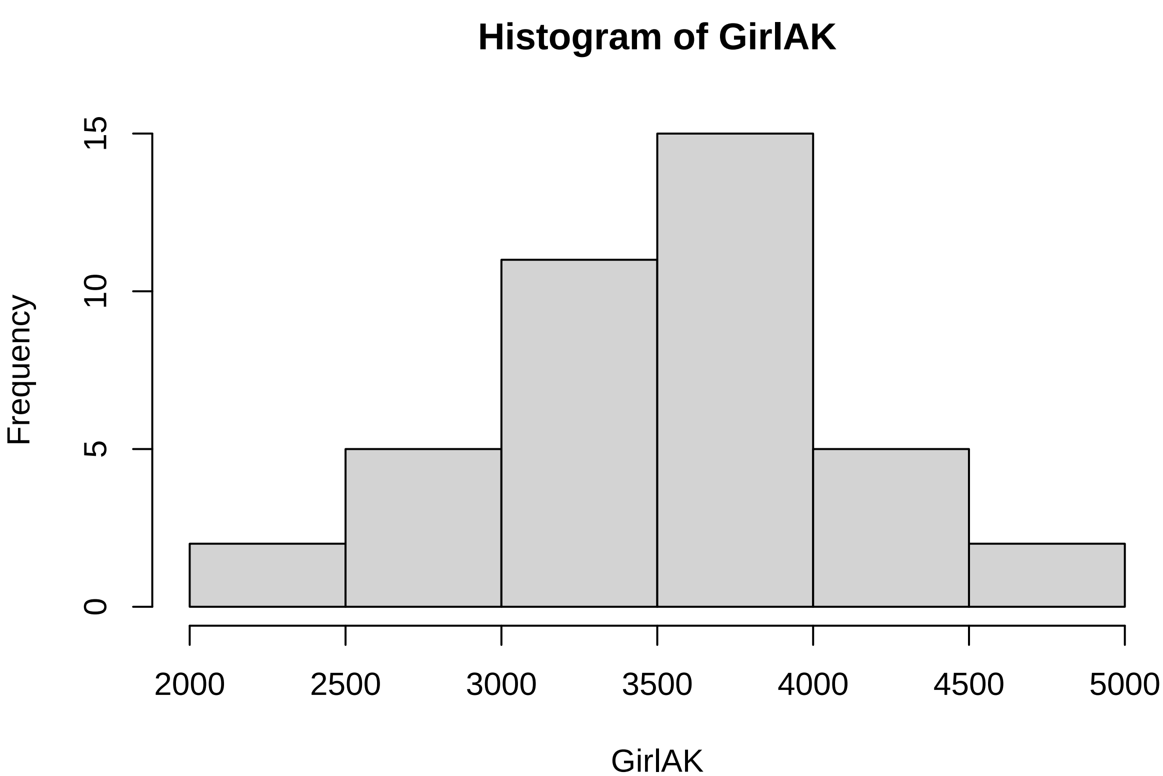
Q9. (Exercise 5.17)

a) Both look normally distributed, GirlAK has a min of 2182, 1st Q of 3558, Median of 3516, Mean of 3516, 3rd Q of 3926, and Max of 4592. GirlWY has a min of 2212, 1st Q of 2934, Median of 3278, Mean of 3208, 3rd Q of 3515, and Max of 3995

```r
Girls2004 <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Girls2004.csv")
GirlWY <- Girls2004 %>% filter (State == "WY") %>% pull(Weight) %>% na.omit()
GirlAK <- Girls2004 %>% filter (State == "AK") %>% pull(Weight) %>% na.omit()
summary(GirlAK)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2182    3170    3558    3516    3926    4592
```

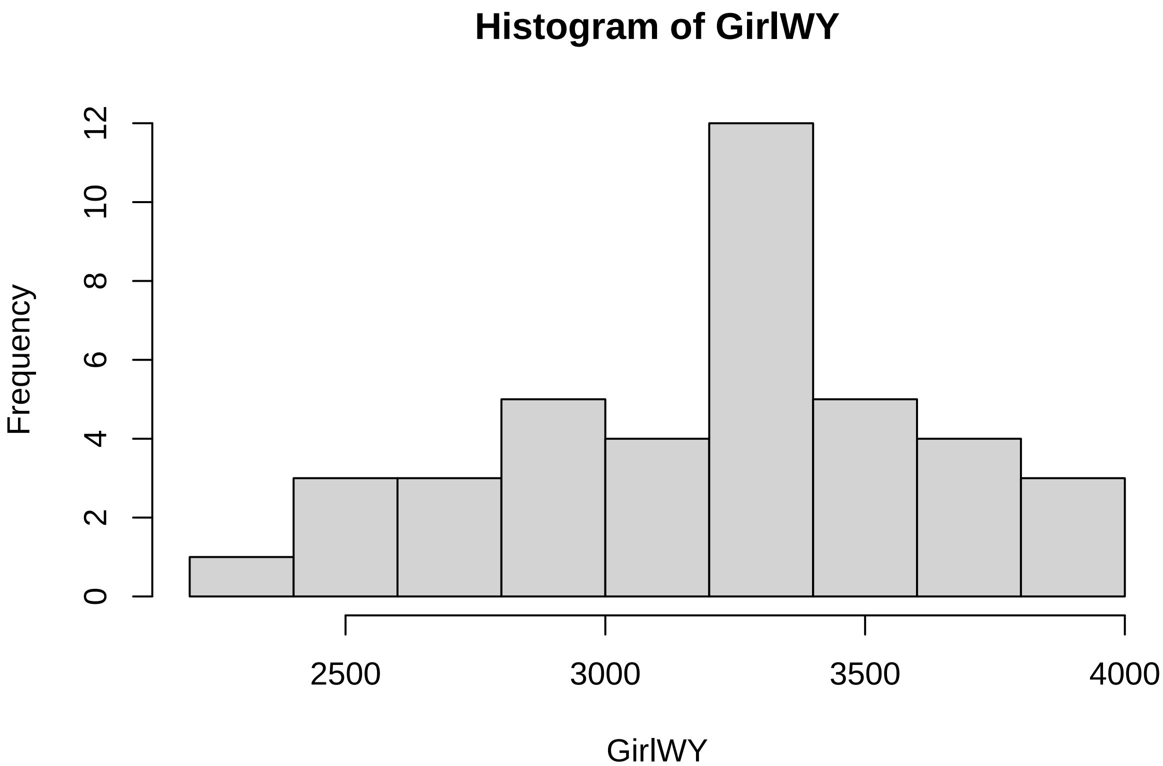
```r
hist(GirlAK)
```




```r
summary(GirlWY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2212    2934    3278    3208    3515    3995
```

```r
hist(GirlWY)
```



b) We get a 95% confidence interval between -529.65313 and -85.54687. What this means is that we are 95% confident that the average weight of a baby girl born in WY is 529.65313 to 85.54687 less than the average weight of a baby girl born in Arkansas.

```r
set.seed(69420)
N <- 10^4
phat9 <- numeric(N)
n <- nrow(Bangladesh)
for (i in 1:N)
{
  WYboot <- sample (GirlWY, replace = TRUE)
  AKboot <- sample (GirlAK, replace = TRUE)
  phat9[i] <- mean(WYboot) - mean(AKboot)
}
quantile(phat9, probs = c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -529.65313  -85.54687
```

c) We get an estimate bias of 0.5158875, which represents a 0.004596454 fraction of the bootstrap standard error

```r
true9 <- mean(GirlWY) - mean(GirlAK)
bias9 <- mean(phat9) - true9
bias9
```

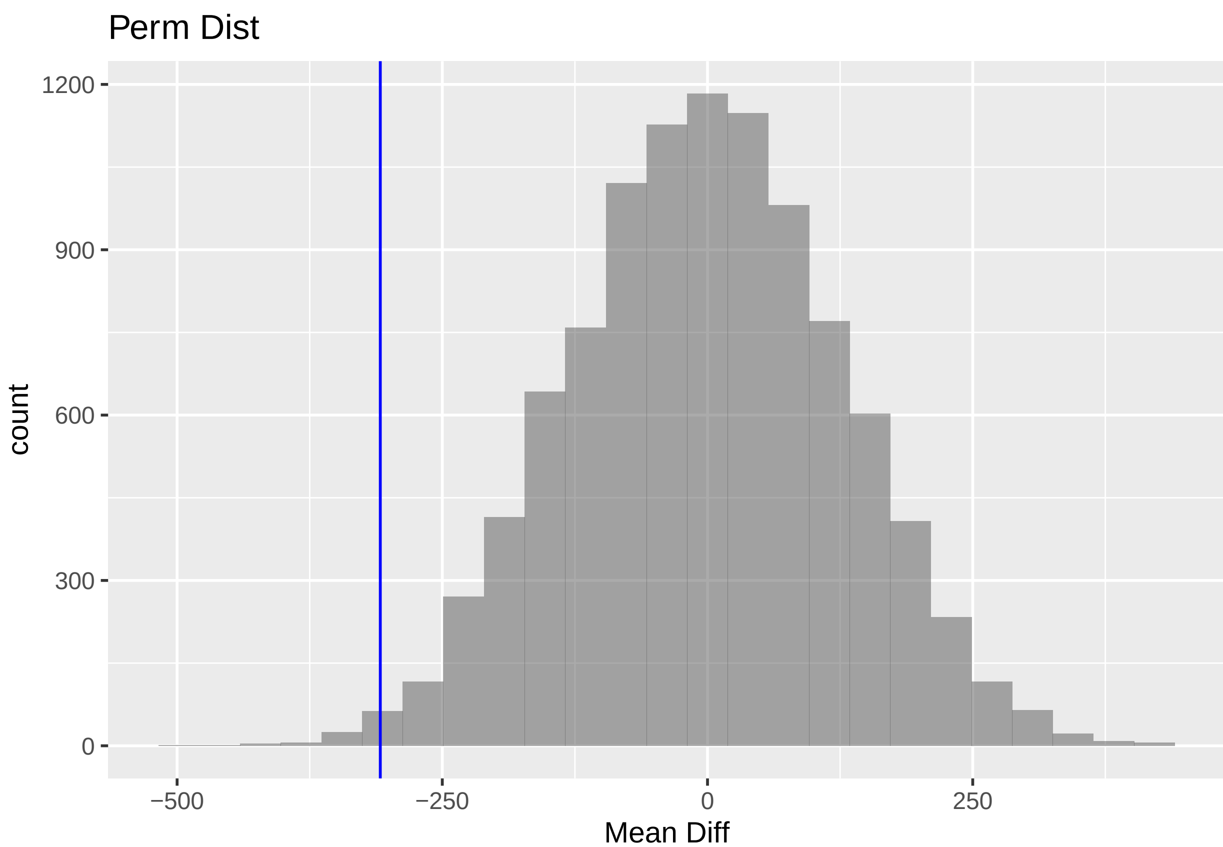
```
## [1] 0.5158875
```

```
bias9 / sd(phat9)
```

```
## [1] 0.004596454
```

d) Our null hypothesis is that there is no weight difference between baby girls born in Arkansas and Wyoming. Our alternative hypothesis is that there is a difference. We get a p-value of 0.01239876 which is small enough to indicate a statisticaly discernible difference between the mean baby girl weight in Wypming and in Arkansas.

```
set.seed(69420)
N <- 10^4
result9 <- numeric(N)
n <- nrow(Bangladesh)
for (i in 1:N)
{
  index <- sample(nrow(Girls2004), size = 40, replace = TRUE)
  result9[i] <- mean(Girls2004$Weight[index], trim = 0.25) - mean(Girls2004$Weight[-index], trim = 0.25
}
gf_histogram(~result9, title = "Perm Dist", xlab = "Mean Diff") %>% gf_vline(xintercept = ~true9, color
```




```
2*(sum(result9 <= true9) + 1)/(N+1)
```

```
## [1] 0.01239876
```

e) Our conclusion holds for the baby girls born in Syoming and Arkansas in 2004 where the gestation period is at least 37 weeks and where there are no twins.