# Homework #3

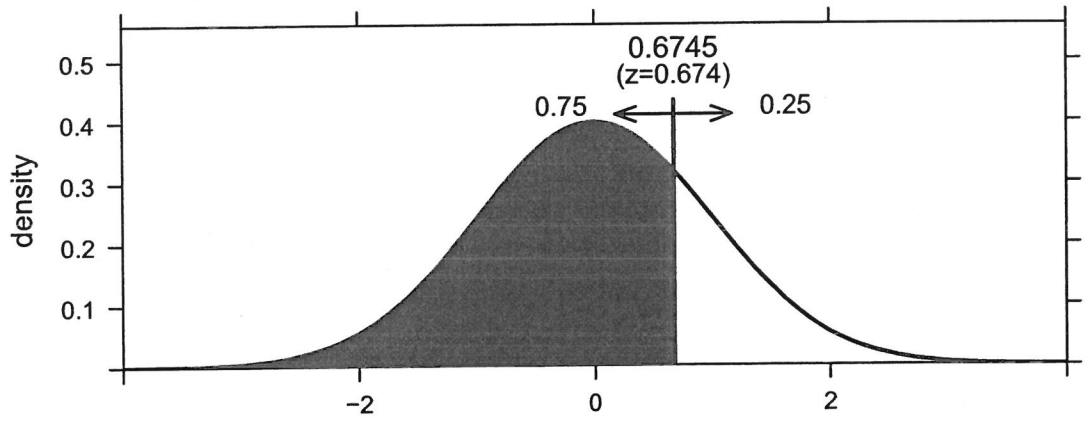
*Derek Sonderegger*

8. Using the Standard Normal Table or the table functions in R, find $z$ that makes the following statements true.

   a) $P(Z < z) = .75$ *For this,* $z = 0.674$ *works.*

```
mosaic::xqnorm(.75)
```

```
##  P(X <= 0.674489750196082) = 0.75
##  P(X >  0.674489750196082) = 0.25
```

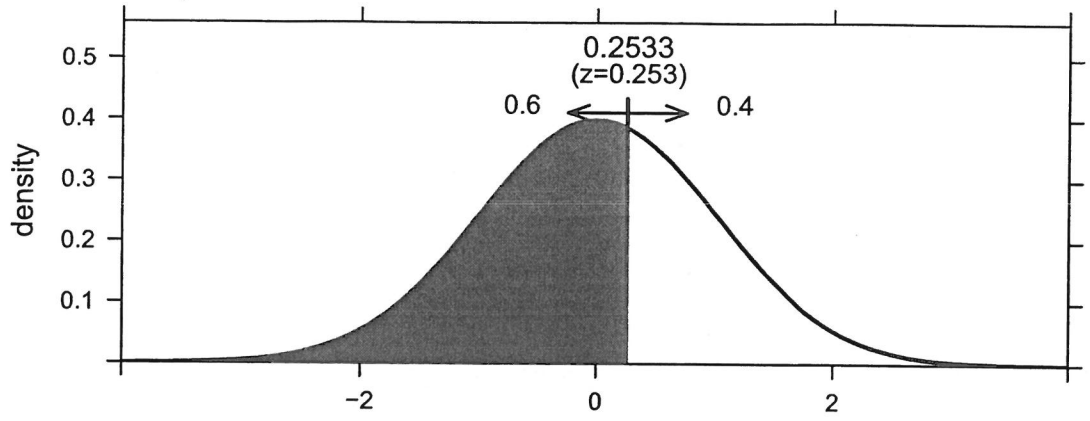



```
## [1] 0.6744898
```

   b) $P(Z > z) = .4$ *For this,* $z = 0.2533$ *works.*

```
mosaic::xqnorm(.6)
```

```
##  P(X <= 0.2533471031358) = 0.6
##  P(X >  0.2533471031358) = 0.4
```




```
## [1] 0.2533471
```

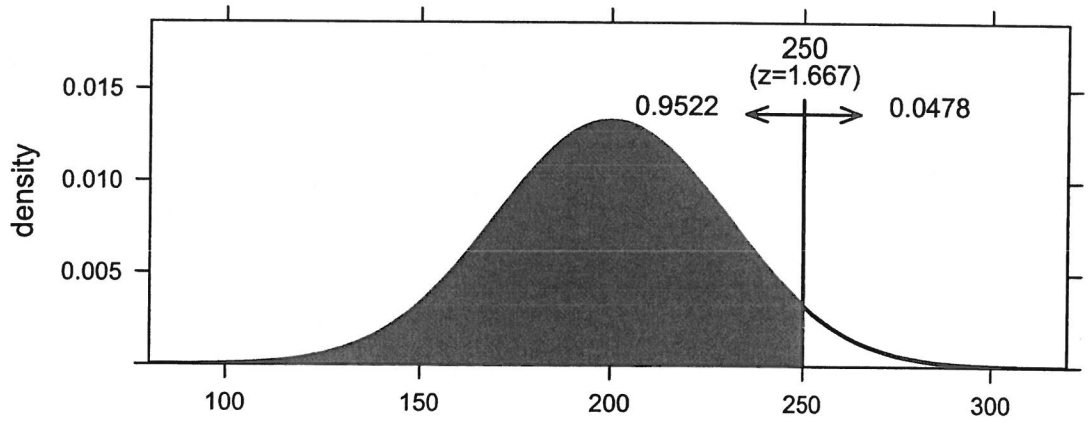
9. The amount of dry kibble that I feed my cats each morning can be well approximated by a normal distribution with mean $\mu = 200$ grams and standard deviation $\sigma = 30$ grams.

   a) What is the probability that I fed my cats more than 250 grams of kibble this morning?

$$\begin{aligned} P(X > 250) &= P\left(\frac{X-\mu}{\sigma} > \frac{250-200}{30}\right) \\ &= P(Z > 1.66667) \\ &= 0.0478 \end{aligned}$$

```
mosaic::xpnorm(250, mean=200, sd=30)
```

```
##
## If X ~ N(200, 30), then
##
##  P(X <= 250) = P(Z <= 1.666667) = 0.9522096
##  P(X >  250) = P(Z >  1.666667) = 0.04779035
```



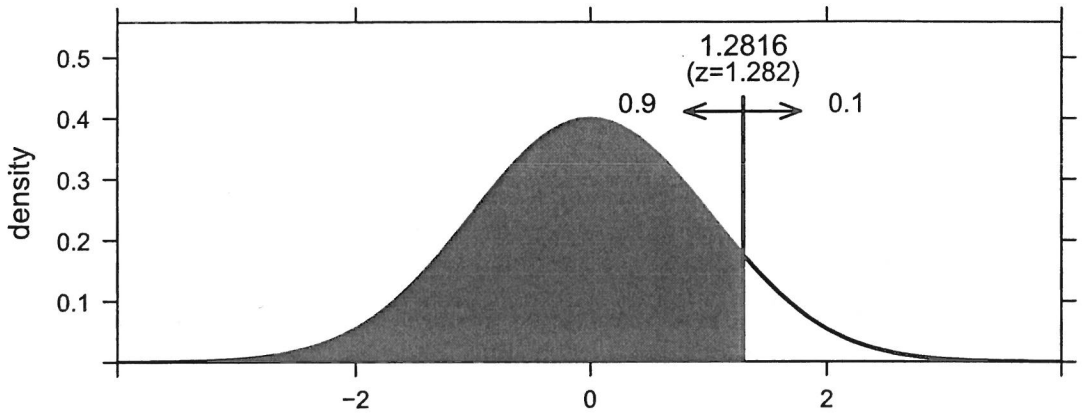

```
## [1] 0.9522096
```

   b) From my cats' perspective, more food is better. How much would I have to feed them for this morning to be among the top 10% of feedings? *We can find the 90th percentile of a standard normal distribution via:*

```
mosaic::xqnorm(.90)
```

```
##  P(X <= 1.2815515655446) = 0.9
```

```
##  P(X >  1.2815515655446) = 0.1
```



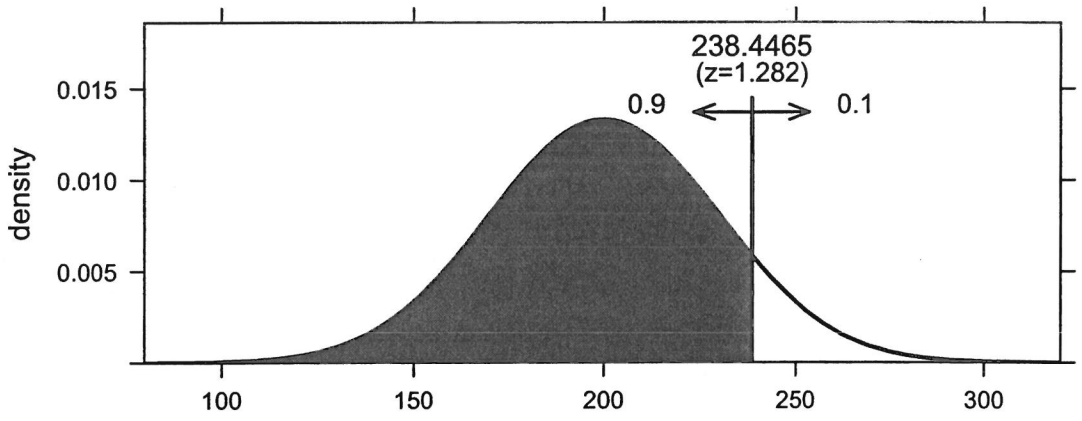

```
## [1] 1.281552
```

*and then plug it back into the standardization equation*

$$
\begin{aligned}
Z &= \frac{X - \mu}{\sigma} \\
1.28 &= \frac{X - 200}{30} \\
1.28\,(30) + 200 &= X \\
238.4 &= X
\end{aligned}
$$

```
mosaic::xqnorm(.90, mean=200, sd=30)
```

```
##  P(X <= 238.446546966338) = 0.9
##  P(X >  238.446546966338) = 0.1
```




```
## [1] 238.4465
```

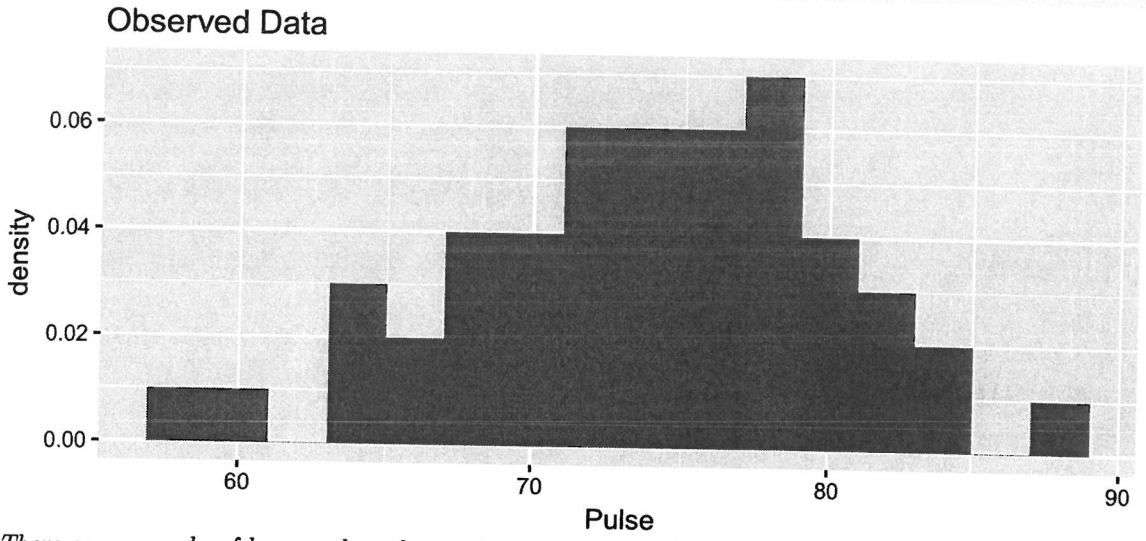
```
# Load all the libraries we'll need
library(mosaic)
library(Lock5Data)
library(ggplot2)
library(dplyr)
```

# Problem 1

Load the dataset `BodyTemp50` from the `Lock5Data` package. This is a dataset of 50 healthy adults. Unfortunately the documentation doesn't give how the data was collected, but for this problem we'll assume that it is a representative sample of healthy US adults. One of the columns of this dataset is the `Pulse` of the $n = 50$ observations, which is the number of heartbeats per minute.

(a) Create a histogram of the observed pulse values.

```
data(BodyTemp50)
ggplot(BodyTemp50, aes(x=Pulse, y=..density..)) +
  geom_histogram(binwidth=2) +
  ggtitle('Observed Data')
```




*There are a couple of lower values, but nothing truly surprising on either the low end or the high end.*

(b) Calculate the sample mean $\bar{x}$ and sample standard deviation $s$ of the pulses.

```
xbar <- BodyTemp50 %>% summarise( mean(Pulse) )
s    <- BodyTemp50 %>% summarise(   sd(Pulse) )
cbind( xbar, s )  # just so they print on one line...
```

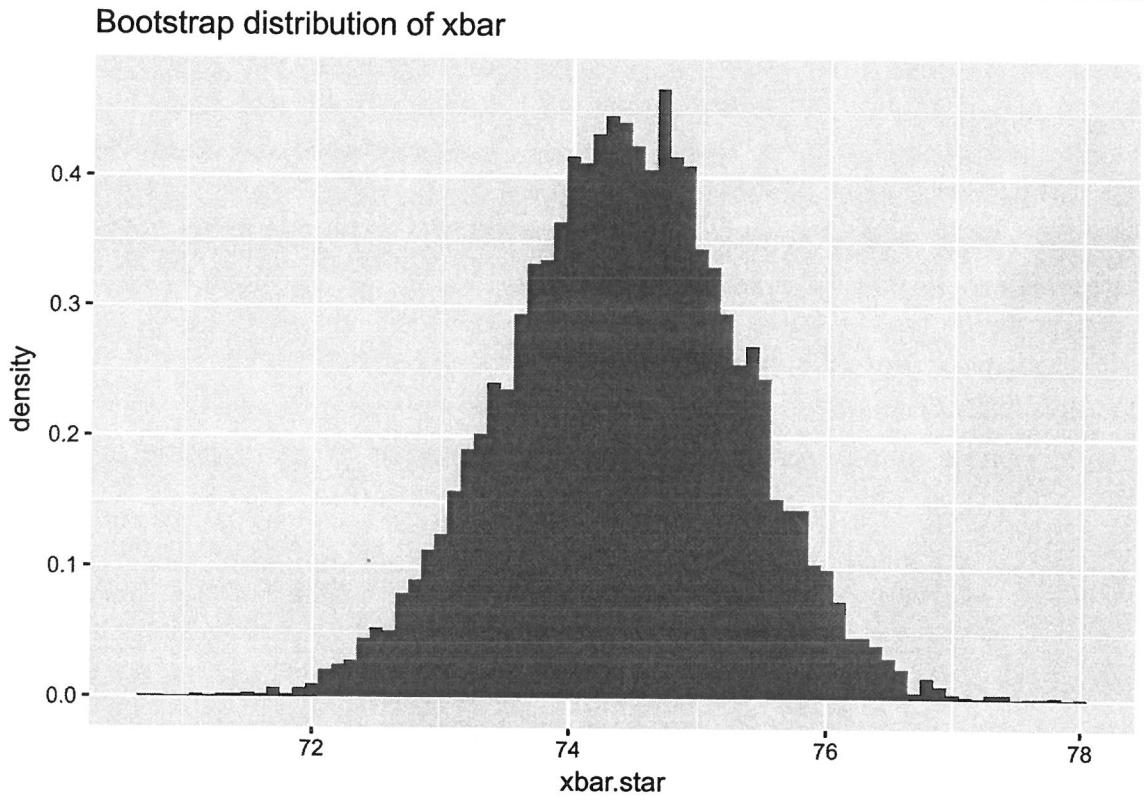
```
##   mean(Pulse) sd(Pulse)
## 1        74.4  6.439673
```

(c) Create a dataset of 10000 bootstrap replicates of $\bar{x}^*$.

```
BootDist <- do(10000) * {
  resample(BodyTemp50) %>% summarise( xbar.star = mean(Pulse) )
}
```

(d) Create a histogram of the bootstrap replicates. Calculate the mean and standard deviation of this distribution.

```
ggplot(BootDist, aes(x=xbar.star, y=..density..)) +
  geom_histogram(binwidth=.1) +
  ggtitle('Bootstrap distribution of xbar')
```




(e) Using the bootstrap replicates, create a 95% confidence interval for $\mu$, the average adult heart rate.

```
quantile( BootDist$xbar.star, probs=c(.025, .975))
```

```
##  2.5% 97.5%
## 72.62 76.14
```

(f) Calculate the standard deviation of your 10000 bootstrap replicates of $\bar{x}$ and we'll call this the Standard Error of $\bar{x}$ and denote it as $\hat{\sigma}_{\bar{x}}$

```
StdErr <- BootDist %>% summarise(sd(xbar.star))
StdErr
```

```
##   sd(xbar.star)
## 1     0.8989337
```

(g) Calculate the interval

$$(\bar{x} - 2\hat{\sigma}_{\bar{x}},\ \bar{x} + 2\hat{\sigma}_{\bar{x}})$$

and comment on its similarity to the interval you calculated in part (e).

```
CI <- cbind(xbar - 2*StdErr,  xbar+ 2*StdErr )
colnames(CI) = c('2.5%', '97.5%')   # pretty column headers...
CI
```

```
##       2.5%    97.5%
## 1 72.60213 76.19787
```

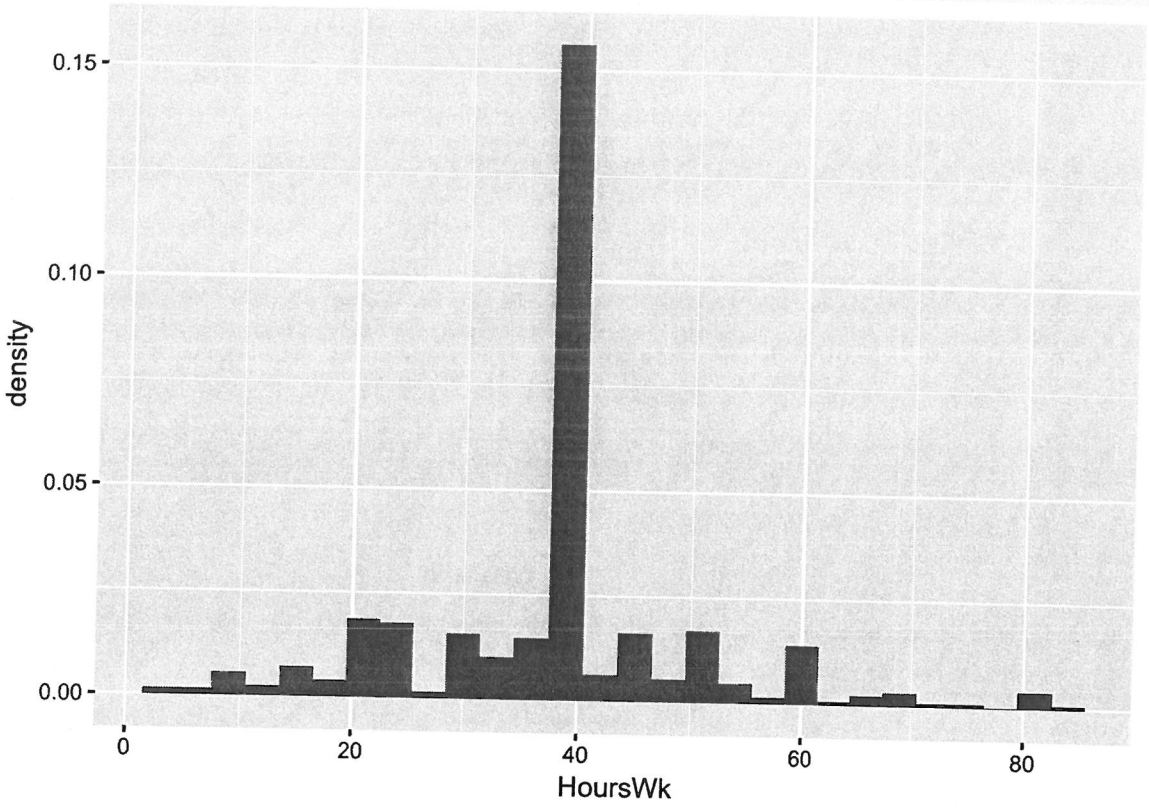
*This is nearly identical to the interval we got when we calculated the quantile based confidence interval.*

# Problem 2

Load the dataset EmployedACS from the Lock5Data package. This is a dataset drawn from American Community Survey results which is conducted monthly by the US Census Bureau and should be representative of US workers. The column `HoursWk` represents the number of hours worked per week.

(a) Create a histogram of the observed hours worked.

```
data(EmployedACS)
ggplot(EmployedACS, aes(x=HoursWk, y=..density..)) +
  geom_histogram(binwidth=3)
```




- This is isn't too surprising. By far the most common number of hours worked is forty hours per week. *

(b) Calculate the sample mean $\bar{x}$ and sample standard deviation $s$ of the worked hours per week.

```
xbar <- EmployedACS %>% summarise( mean(HoursWk) )
s    <- EmployedACS %>% summarise(   sd(HoursWk) )
cbind( xbar, s )  # just so they print on one line...
```

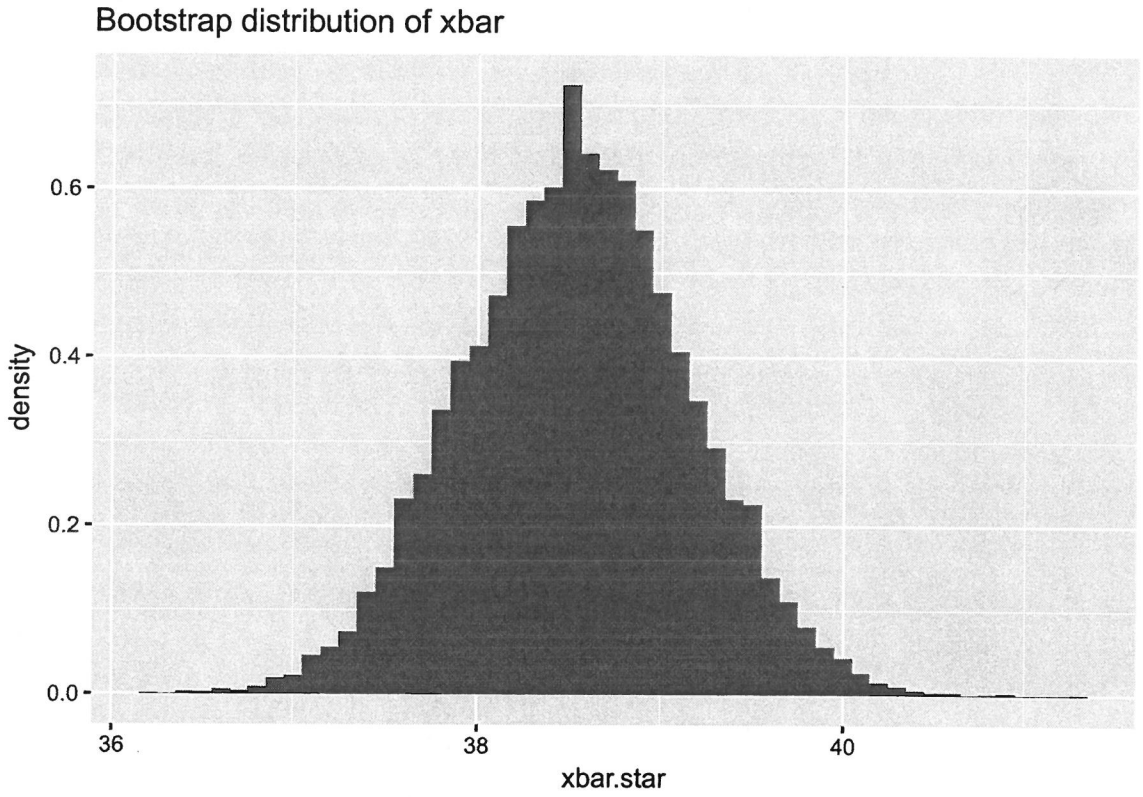
```
##   mean(HoursWk) sd(HoursWk)
## 1      38.53596    12.74373
```

*The sample mean is* $\bar{x} = 38.5$ *and the sample standard deviation is* $s = 12.7$.

```
BootDist <- do(10000) * {
  resample(EmployedACS) %>% summarise( xbar.star = mean(HoursWk) )
}
```

(c) Create a histogram of the bootstrap replicates. Calculate the mean and standard deviation of this distribution.

```
ggplot(BootDist, aes(x=xbar.star, y=..density..)) +
  geom_histogram(binwidth=.1) +
  ggtitle('Bootstrap distribution of xbar')
```




(d) Using the bootstrap replicates, create a 95% confidence interval for $\mu$, the average adult heart rate.

```
quantile( BootDist$xbar.star, probs=c(.025, .975))
```

```
##     2.5%    97.5%
## 37.35731 39.74716
```

(e) Calculate the standard deviation of your 10000 bootstrap replicates of $\bar{x}$ and we'll call this the Standard Error of $\bar{x}$ and denote it as $\hat{\sigma}_{\bar{x}}$

```
StdErr <- BootDist %>% summarise(sd(xbar.star))
StdErr
```

```
##   sd(xbar.star)
## 1     0.6140092
```

(f) Calculate the interval

$$(\bar{x} - 2\hat{\sigma}_{\bar{x}},\ \bar{x} + 2\hat{\sigma}_{\bar{x}})$$

and comment on its similarity to the interval you calculated in part (e).

```
CI <- cbind(xbar - 2*StdErr,  xbar+ 2*StdErr )
colnames(CI) = c('2.5%', '97.5%')   # pretty column headers...
CI
```

```
##       2.5%    97.5%
## 1 37.30794 39.76398
```

*This is nearly identical to the interval we got when we calculated the quantile based confidence interval. Many students used the mean of the bootstrap distribution as the center of this interval which is technically incorrect (and is inconsistent with how we will create this intervals in the next chapter), but that is a relatively small issue because with more bootstrap samples, the mean of the bootstrap samples will eventually be the sample mean.*