Homework 11 Solutions

JMG

2020-05-13

```r
library(resampledata)
library(fastR2)
```

Problem 1

Consider a population that has a normal distribution with mean $\mu = 36$ and standard deviation $\sigma = 8$ (so the variance is $\sigma^2 = 64$).

The following code draws a random sample of size $n = 200$ from this population:

```r
samp_data <- rnorm(200,mean=36,sd=8)
```

For a normal distribution, the sample mean
$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$
is an estimator for the population mean. Thus,

```r
(obs_est <- mean(samp_data))
```

```
## [1] 36.31484
```

provides an estimate for the population mean.

The following code uses 5000 resamples from the sample data to compute the bootstrap distribution for the sample, and returns the bootstrap mean and standard error.

```r
N <- 5000
boot_dist <- do(N) * c(boot_stat=mean(sample(samp_data,replace = TRUE)))
(boot_mean <- mean(boot_dist$boot_stat))
```

```
## [1] 36.29799
```

```r
(boot_se <- sd(boot_dist$boot_stat))
```

```
## [1] 0.5398548
```

a) Interpret the results of the previous code.

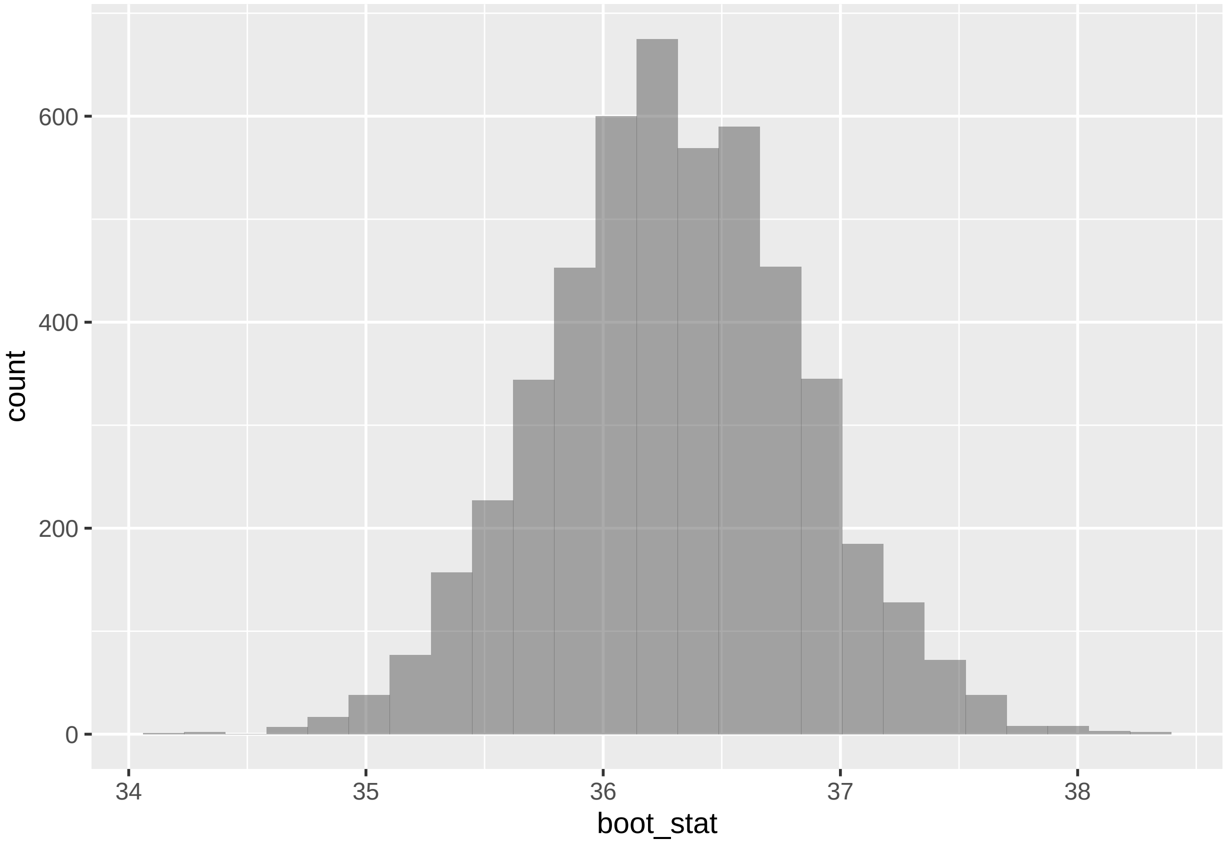
Solution

The mean of the bootstrap distribution 36.2979865 is approximately the sample mean and the standard deviation of bootstrap distribution 0.5398548 approximates the stadard error of the sampling distribution for the mean.

b) Plot a histogram of the bootstrap distribution and describe its characterisitcs.

Solution

```r
boot_dist %>% gf_histogram(~boot_stat)
```



We see that the bootstrap distribution is roughly bell-shaped and is centered approximately at the sample mean 36.314845.

c) Use the bootstrap distribution to obtain an approximate 95% confidence interval for the estimate of the population mean.

Solution

```r
quantile(boot_dist$boot_stat,c(0.025,0.975))
```

```
##     2.5%    97.5% 
## 35.25068 37.37105 
```

Problem 2

Modify the code from problem 1 in order to take a sample of size $n = 150$ from a population that has a normal distribution with mean $\mu = 16$ and standard deviation $\sigma = 6$.

Solution

```r
samp_data <- rnorm(150,mean=16,sd=6)
```

Then,

a) use 10,000 resamples to obtain the bootstrap distribution for your sample,

Solution

```
N <- 10000
boot_dist <- do(N) * c(boot_stat=mean(sample(samp_data,replace = TRUE)))
```

b) compute the bootstrap mean and standard error, and

Solution

```
(boot_mean <- mean(boot_dist$boot_stat))
```

```
## [1] 15.85848
```

```
(boot_se <- sd(boot_dist$boot_stat))
```

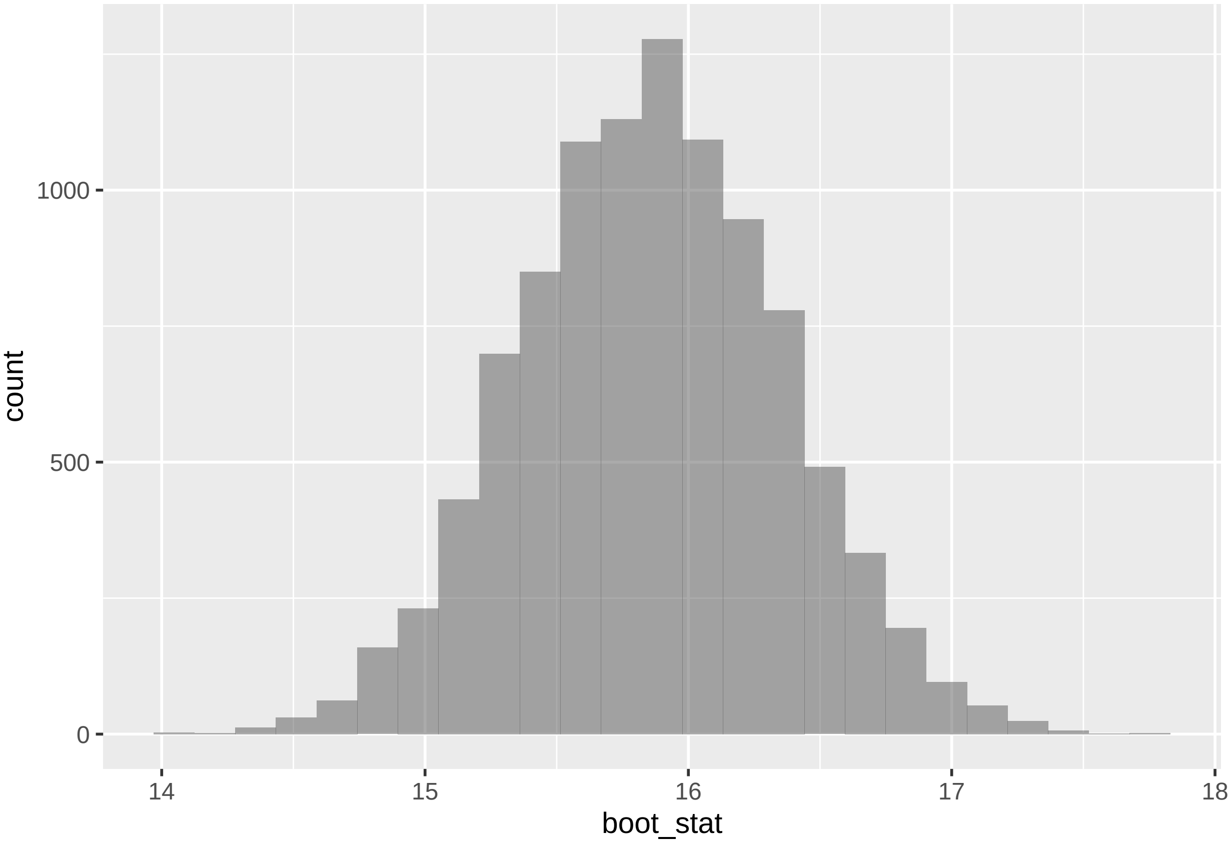
```
## [1] 0.4977154
```

c) repeat parts a, b, and c from problem 1 for the sample data you generated for problem 2.

Solution

The mean of the bootstrap distribution 15.8584783 is approximately the sample mean and the standard deviation of bootstrap distribution 0.4977154 approximates the stadard error of the sampling distribution for the mean.

```
boot_dist %>% gf_histogram(~boot_stat)
```



```
quantile(boot_dist$boot_stat,c(0.025,0.975))
```

```
##      2.5%     97.5%
## 14.88181 16.83997
```

Problem 3

Consider the Bangladesh data set from the resampledata package:

```
head(Bangladesh)
```

```
##   Arsenic Chlorine Cobalt
## 1    2400      6.2   0.42
## 2       6    116.0   0.45
## 3     904     14.8   0.63
## 4     321     35.9   0.68
## 5    1280     18.9   0.58
## 6     151      7.8   0.35
```

This data records levels of three chemicals found in the groundwater of Bangladesh. In this problem we will use the bootstrap to understand the distribution of levels of arsenic (measured in parts per billion (ppb)) in the groundwater. We can extract the vector of arsenic levels as follows:

```
Arsenic <- Bangladesh$Arsenic
```

The US EPA sets an arsenic maximum contaminant level for public water supplies at 10ppb.

a) Compute the mean and standard deviation for the aresnic level recorded in the Arsenic data.

Solution

```
(arsenic_mean <- mean(Arsenic))
```

```
## [1] 125.3199
```

```
(arsenic_sd <- sd(Arsenic))
```

```
## [1] 297.9755
```

b) Use 10,000 resamples to compute the bootstrap distribution for the sample mean of the arsenic data.

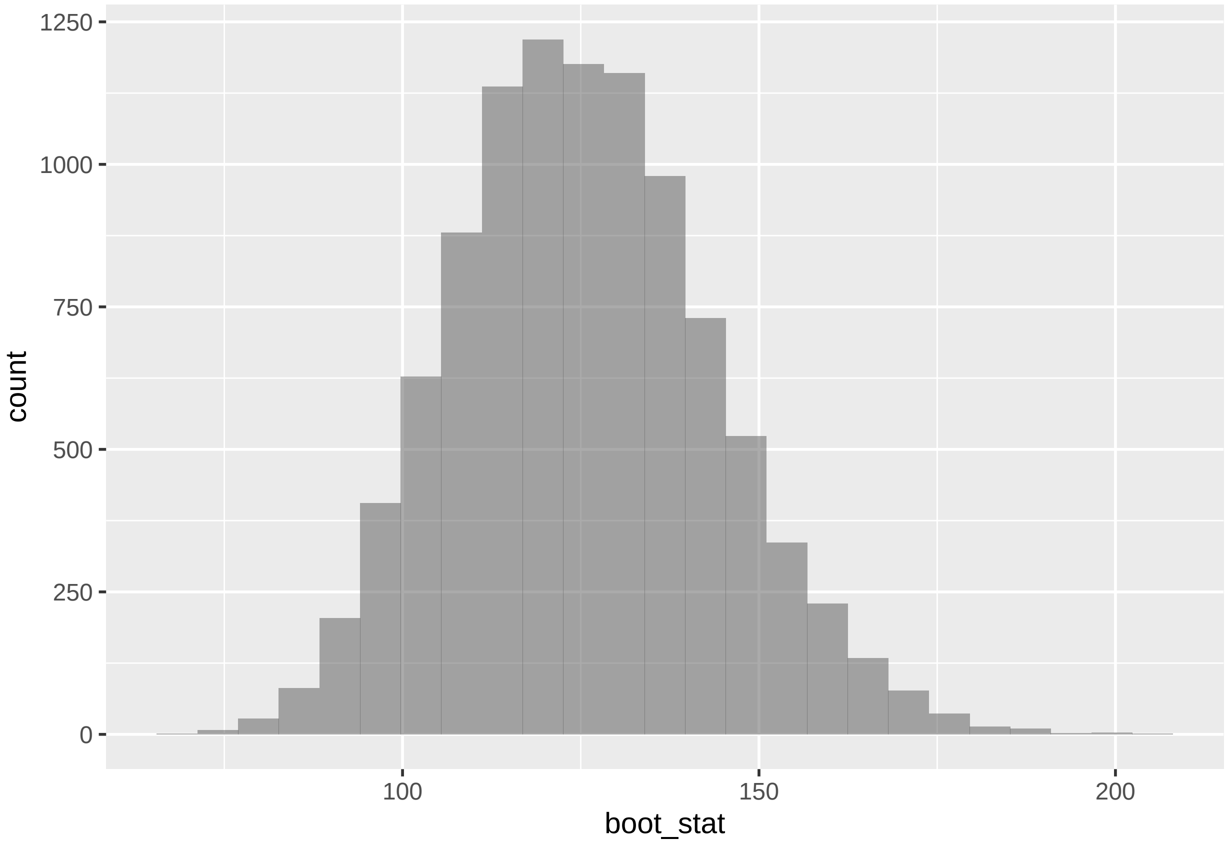
Solution

```
N <- 10000
boot_dist <- do(N) * c(boot_stat=mean(sample(Arsenic,replace = TRUE)))
```

c) Plot a histogram of the bootstrap distribution.

Solution

```
boot_dist %>% gf_histogram(~boot_stat)
```



d) Use the boostrap to find and interpret an approximate 95% confidence interval for the sample mean of the arsenic data.

Solution

```
(arsenic_ci <- quantile(boot_dist$boot_stat,c(0.025,0.975)))
```

```
##      2.5%     97.5% 
##  92.42493 163.63844
```

We can be confident (at the 95% level) that the true mean arsenic level is between 92.4249262 and 163.638441.