Atlas-PS 9

David Atlas

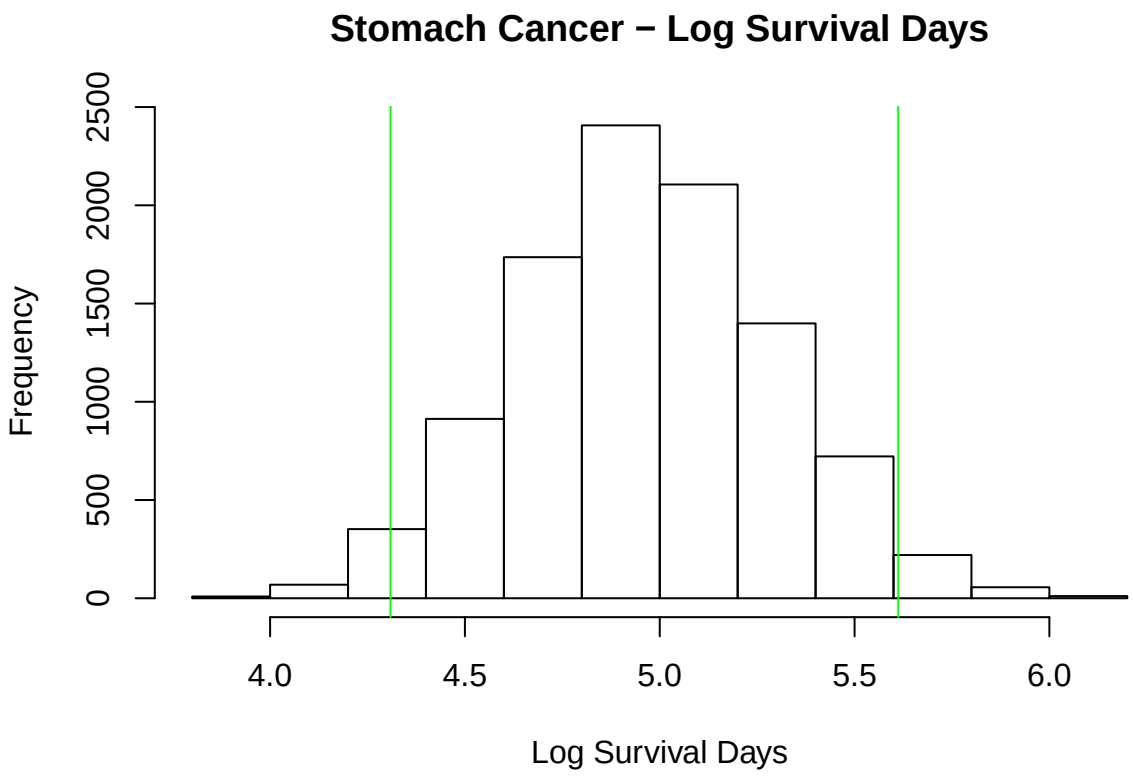
10/21/2018

Problem 1

a)

We find the bootstrapped 95% CI for the mean log survival time below. We plot the histograms with the confidence intervals for the mean.

```
get_bootstraps <- function(data, T, n){
  return(sapply(seq(1, n), function(i) T(sample(data, length(data), replace=TRUE))))
}

get_studentized_percentile <- function(data, alpha){
  return(quantile((data - mean(data)) / sd(data), alpha))
}

get_studentized_ci <- function(data, alpha){
  lwr <- mean(data) - (sd(data) * get_studentized_percentile(data, 1 - (alpha / 2)))
  upr <- mean(data) - (sd(data) * get_studentized_percentile(data, (alpha / 2)))
  return(c(lwr, upr))
}

set.seed(73)
survival.s <- log(c(25, 42, 45, 46, 51, 103, 124, 146, 340, 396, 412, 876, 1112))

n <- 10000
alpha <- .05
bootstraps.s <- get_bootstraps(survival.s, mean, n)
stdev.s <- sd(bootstraps.s)

lwr <- mean(bootstraps.s) - (stdev.s * get_studentized_percentile(bootstraps.s, 1 - (alpha / 2)))
upr <- mean(bootstraps.s) - (stdev.s * get_studentized_percentile(bootstraps.s, (alpha / 2)))
hist(bootstraps.s, main="Stomach Cancer - Log Survival Days", xlab="Log Survival Days")
abline(v=lwr, col='green')
abline(v=upr, col='green')
```



We find the 95% CI for the mean log survival time for stomach cancer to be $[4.309, 5.612]$. To find the bootstrap variance, we simply take the variance of the bootstrapped means:

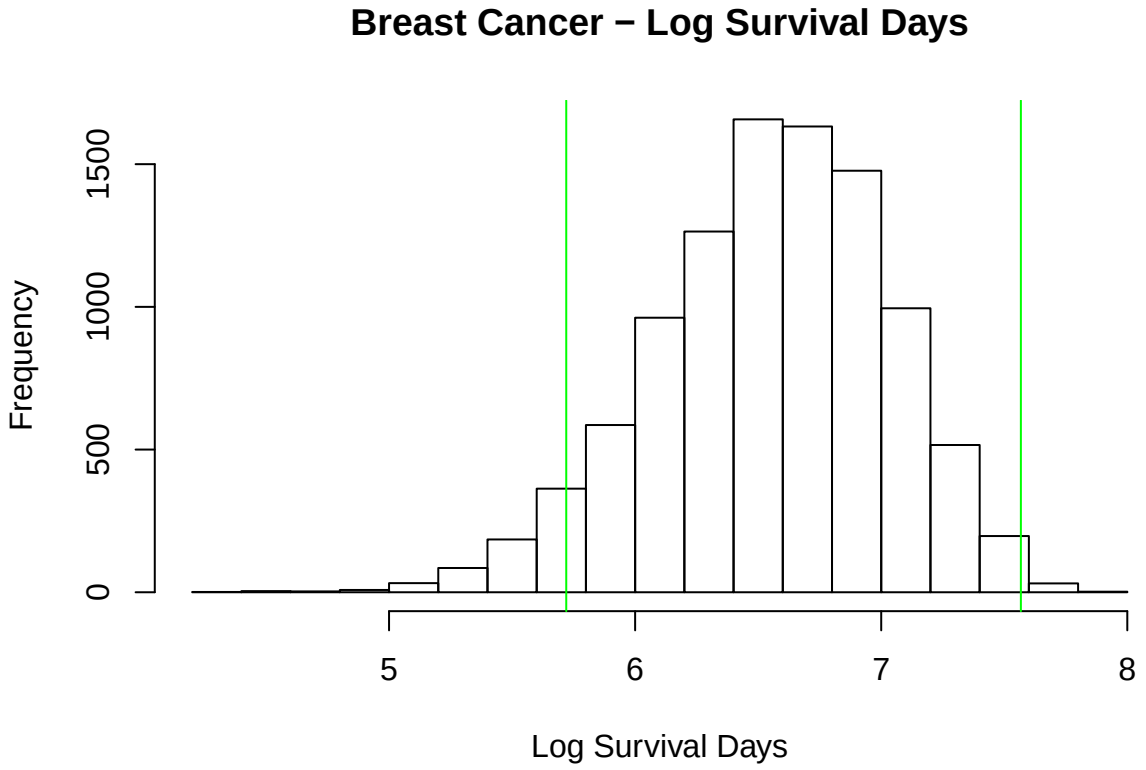
```
var(bootstraps.s)
```

```
## [1] 0.1101809
```

The variance of the estimates is .11.

Next, we repeat the process for breast cancer.

```
set.seed(73)
survival.b <- log(c(24, 40, 719, 727, 791, 1166, 1235, 1581, 1804, 3460, 3808))

# Find the CI for

n <- 10000
alpha <- .05
bootstraps.b <- get_bootstraps(survival.b, mean, n)
stdev.b <- sd(bootstraps.b)

lwr <- mean(bootstraps.b) - (stdev.b * get_studentized_percentile(bootstraps.b, 1 - (alpha / 2)))
upr <- mean(bootstraps.b) - (stdev.b * get_studentized_percentile(bootstraps.b, (alpha / 2)))
hist(bootstraps.b, main="Breast Cancer - Log Survival Days", xlab="Log Survival Days")
abline(v=lwr, col='green')
abline(v=upr, col='green')
```



For breast cancer, we find it to be $[5.720, 7.567]$. We find the variance of the estimates:

```
var(bootstraps.b)
```

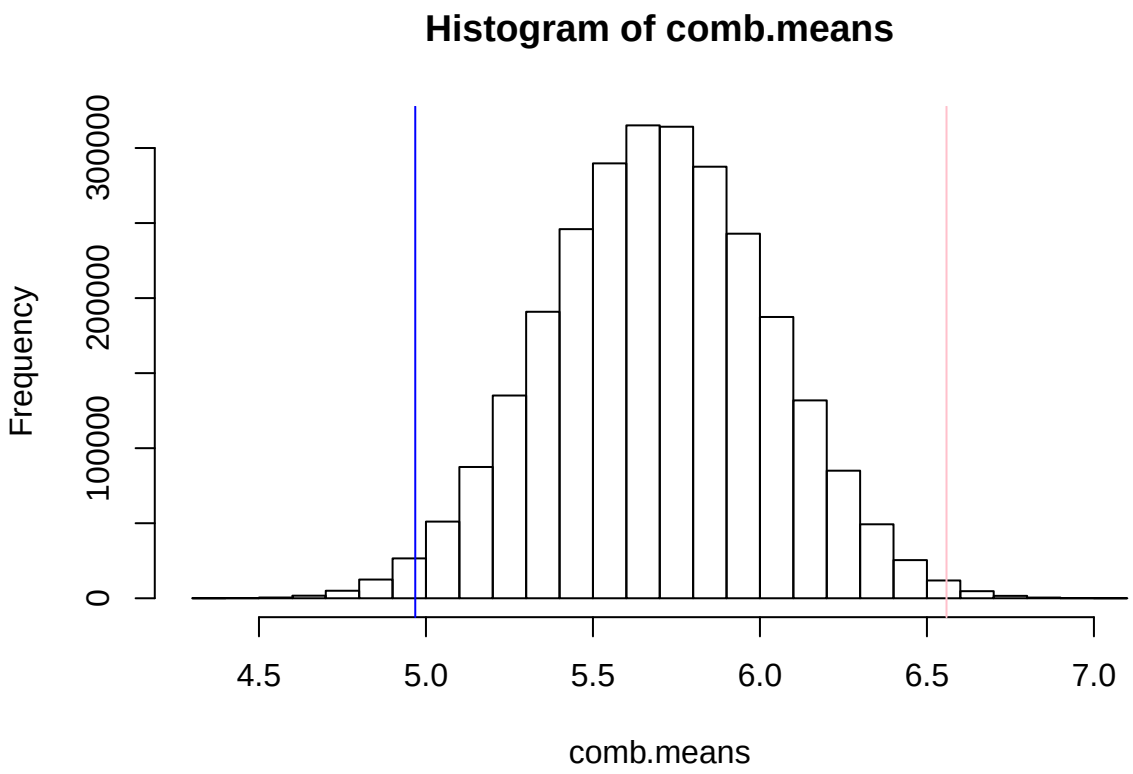
```
## [1] 0.2236246
```

We find it to be larger, at .22.

b)

Next, we do a permutation test. We combine all the observations into one long vector, and create all combinations of 11 observations.

```
# Combine the vectors
survival <- c(survival.s, survival.b)

# Get all combinations
survival.comb <- combn(survival, 12, simplify=T)

# Get the mean of each combination
comb.means <- colMeans(survival.comb)

# Plot the means and show the calculated numbers
hist(comb.means)
abline(v=mean(survival.b), col='pink')
abline(v=mean(survival.s), col='blue')
```




```
# Print the p values
mean(mean(survival.b) > comb.means)
```

```
## [1] 0.9960631
```

```
mean(mean(survival.s) > comb.means)
```

```
## [1] 0.01324073
```

We can reject with 98% confidence the null hypothesis that the mean breast cancer survival time is the same as the mean stomach cancer survival time.

c)

Next, we use the percentile method to create a 95% confidence interval for the mean breast cancer survival time. We do this on both the log data and the data on original scale. We then compare the exponentiated results of the former to the latter.

```
exp(quantile(bootstraps.b, c(.025, .975)))
```

```
##      2.5%     97.5%
##  254.1661 1612.3977
```

```
quantile(get_bootstraps(exp(survival.b), mean, n), c(.025, .975))
```

```
##      2.5%     97.5%
##  751.6318 2123.2864
```

We see that the intervals are wildly different when we compare them. We exponentiate the intervals found in part a and compare as well.

```
lwr <- mean(bootstraps.b) - (stdev.b * get_studentized_percentile(bootstraps.b, 1 - (alpha / 2)))
upr <- mean(bootstraps.b) - (stdev.b * get_studentized_percentile(bootstraps.b, (alpha / 2)))
exp(c(lwr, upr))
```

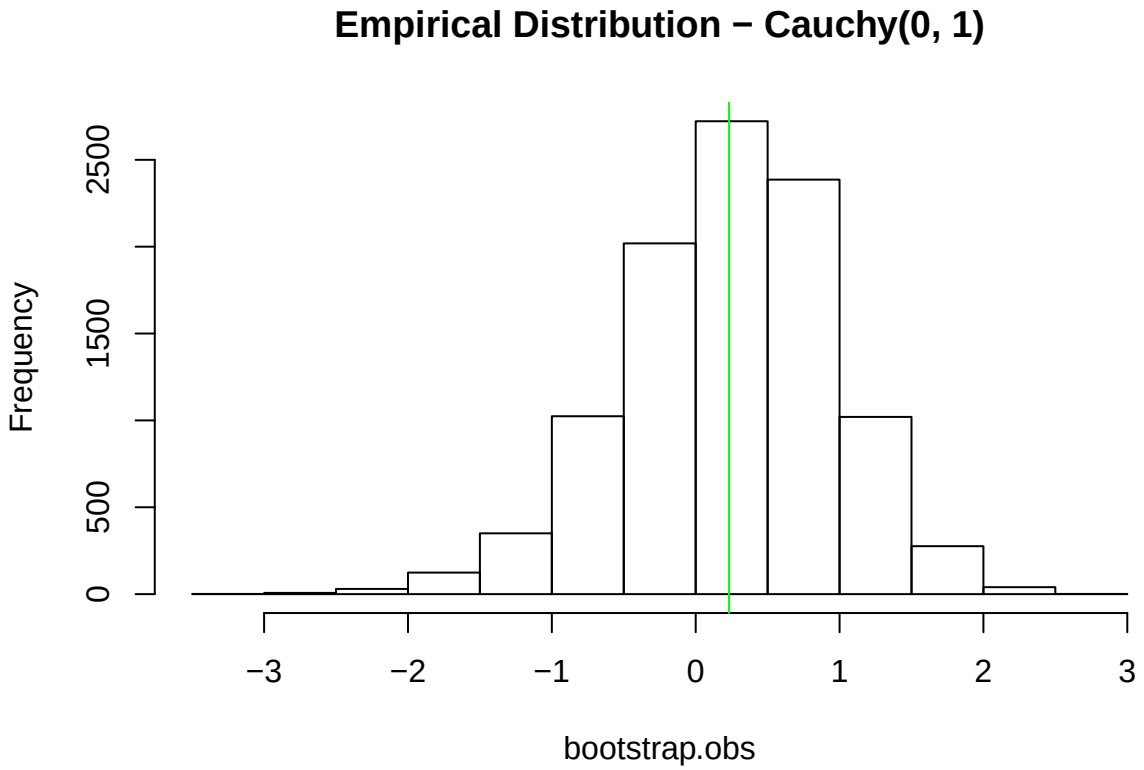
```
##     97.5%      2.5%
##  304.9126 1934.3267
```

We see that this interval is different, although much closer to the exponentiated log bootstrap than to the bootstrap on the original data.

Problem 2

We first tackle the problem of estimating the mean of a Cauchy distribution. We generate an empirical distribution and look at the mean of the mean estimate (in green).

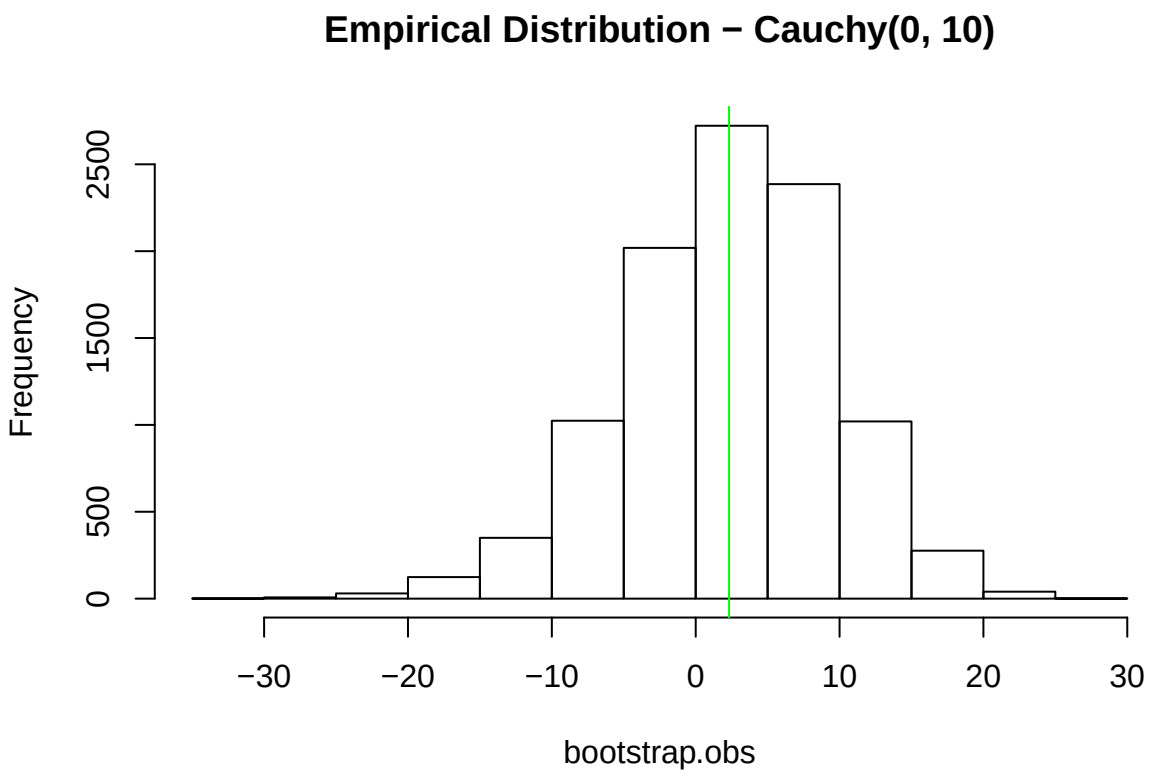
```
set.seed(73)
n0 <- 1000
n <- 10000
obs <- rcauchy(n=n0, location=0, scale=1)
bootstrap.obs <- get_bootstraps(obs, mean, n)

hist(bootstrap.obs, main="Empirical Distribution - Cauchy(0, 1)")
abline(v=mean(bootstrap.obs), col='green')
```



We adjust the parameters to get more skewed data and plot the histogram with the mean in green.

```
set.seed(73)
n0 <- 1000
n <- 10000
```

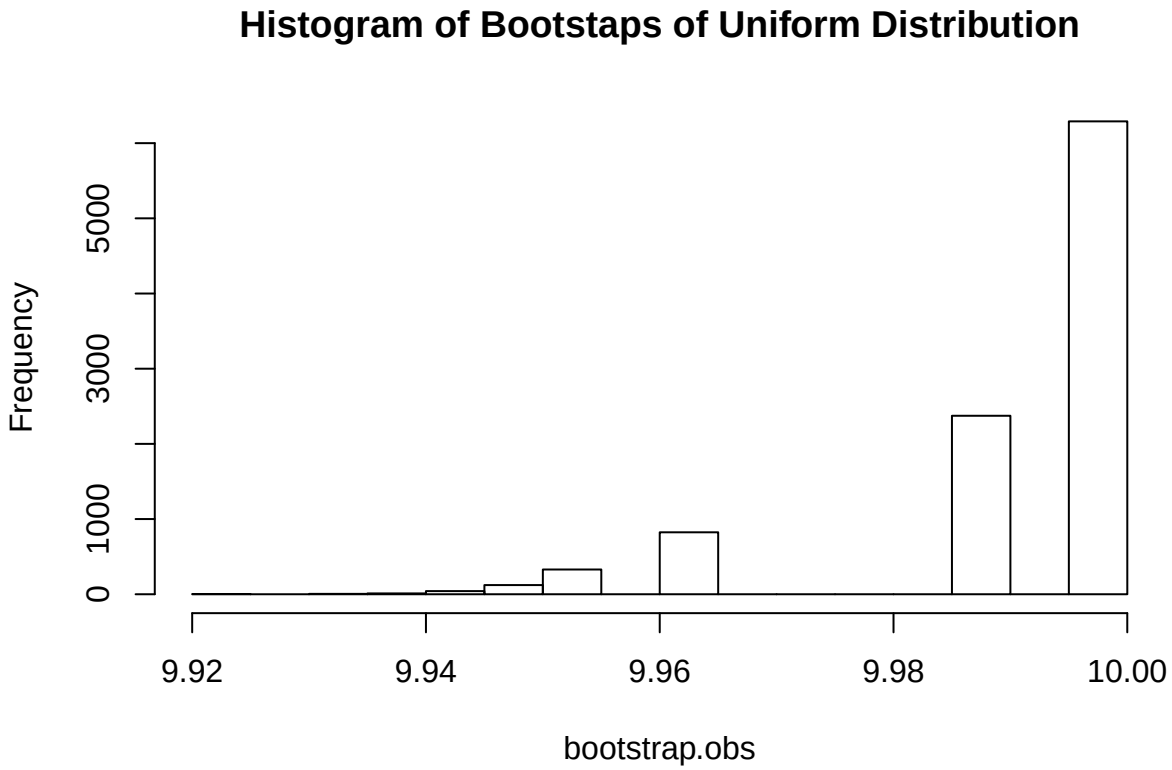
```r
scale0 <- 10
obs <- rcauchy(n=n0, location=0, scale=scale0)
bootstrap.obs <- get_bootstraps(obs, mean, n)

hist(bootstrap.obs, main=paste0("Empirical Distribution - Cauchy(0, ", scale0, ")"))
abline(v=mean(bootstrap.obs), col='green')
```



In both cases, our estimates are not so accurate. The location parameter of a Cauchy distribution is not the expected value, so this is somewhat expected.

Next, we estimate the upper bound θ of a uniform distribution $U(0, \theta)$ by generating an empirical distribution and bootstrapping the maximum.

```r
obs <- runif(n0, min = 0, max=10)
bootstrap.obs <- get_bootstraps(obs, max, n)
hist(bootstrap.obs, main="Histogram of Bootstaps of Uniform Distribution")
```




```
max(bootstrap.obs)
```

```
## [1] 9.995855
```

This bootstrapped estimate for the maximum is very close to the correct estimate. However, no 10 is drawn in the random sample, so getting the degree of error is tough.

Problem 3

a)

First, we define $E_{\hat{P}}(Y) = \sum Y_i \hat{P}(Y_i) = \bar{Y}$, as the empirical distribution is discrete. By the bootstrap principle, we can equate the distribution of $\bar{Y}^*$, so $E_{\hat{P}}(Y^*) = \sum Y_i^* \hat{P}(Y_i^*) = \bar{Y^*}$.

b)

Next, note that $E_P(Y) = \int Y P(Y) dY = \mu$, as defined in the problem. Again, we can apply the bootstrap principle to say that $E_P(Y^*) = \int Y^* P(Y^*) dY^*$, which has the same distribution as the empirical distribution. Therefore, it will also be μ.

Problem 4

a)

```
set.seed(73)
data <- rnorm(n=100)
mean(data)
```

```
## [1] 0.06920712
```

The empirical mean from the 100 data points is .069.

b)

Next, we create 10 bootstrapped datasets and find the sample mean.

```
bootstraps.data <- get_bootstraps(data, mean, 10)
mean(bootstraps.data)
```

```
## [1] 0.09002107
```

The bootstrap mean is .0708.

```
mean(bootstraps.data) - mean(data)
```

```
## [1] 0.02081395
```

The bootstrap bias is .0016.

```
var(bootstraps.data)
```

```
## [1] 0.01815203
```

The variance of the bootstrapped means is .0066.

c)

Next, we use the balanced bootstrap to repeat the experiment.

```
balanced_bootstrapped <- function(data, T, n){
  data.balanced <- sample(rep(data, n), size=length(data) * n, replace=FALSE)
  sapply(split(data.balanced, ceiling(seq_along(data.balanced) / n)), T)
}

mean(balanced_bootstrapped(data, mean, 10))
```

```
## [1] 0.06920712
```

The balanced bootstrap mean is .069. The bootstrap bias is zero, by definition of the balanced bootstrap.

```
var(balanced_bootstrapped(data, mean, 10))
```

```
## [1] 0.1891444
```

The variance of the bootstraps is .134. ## d) The bias is smaller under the balanced bootstrap, as it is zero by definition. However, the variance is higher than under the standard bootstrap.