Lab 4: Interval Estimation and the Bootstrap

Graded Out of 35 Points

Jennifer Bradley

Invalid Date

We will be using a variety of new packages today. Make sure you install each one using `install.packages("package name")`.

```r
library(palmerpenguins)
library(vegan)
```

```
Loading required package: permute
```

```
Loading required package: lattice
```

```r
library(MASS)
library(bcaboot)

#Let's toss out rows with NA values for body mass
penguins <- penguins[!is.na(penguins$body_mass_g),]
```

1. Confidence Intervals

Problem 1.1 (5 Points)

For a given dataset x, I have estimated that $\hat{\mu} = 3$ and that the upper and lower 95% CIs for this estimate are $C_L = 2$ and $C_U = 4$. I wrote the following statements about this analysis in the first draft of a paper I am writing. Based on only the information given, determine if each statement is necessarily true, and revise all false statements to be true (1 Point each):

A. There is a probability of 0.95 that the true mean falls in the interval between 2 and 4.

B. I am 100% certain that the sample mean, $\hat{\mu}$, falls in the interval between 2 and 4.

C. If I were to sample the parent population 100 times to generate 100 new datasets $\{x_1, ..., x_{100}\}$ with sample means $\{\hat{\mu}_1, ..., \hat{\mu}_{100}\}$, each with their own set of confidence intervals, I would expect about 95% of those confidence intervals to contain the true mean μ of the parent population.

D. The sampling error of $\hat{\mu}$ describes the spread (standard deviation) of the sampled data.

E. I am confident about how I calculated the 95% CIs because the sample mean will always be normally distributed regardless of the distribution x was generated from according to the central limit theorem.

Problem 1.1 Answers and Revised Statements

A. FALSE
We are 95% confident that $\hat{\mu}$ falls in the interval between 2 and 4.
B. TRUE
C. IDK
D. TRUE
E. IDK –> DEPENDS ON NUMBER OF SAMPLES? IG 100 IS ENOUGH?

Problem 1.2 (5 Points)

A. (2 Points) For a given random variable X with unknown true mean $\bar{X} = \mu$ and true variance σ^2, we have calculated a sample mean $\bar{x} = \hat{\mu}$ from a set of n observations x. To quantify our confidence in our estimate, we would like to estimate the sample standard error of the mean:

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$$

Write a function to calculate the sample standard error from a vector of data x.

- You may want to use the functions `sd` and `length` here to make your life easier.

```r
sample_standard_error <- function(data_x){
  sse = (sd(data_x))/sqrt(length(data_x))
  return(sse)
}
```

B. (2 Points) Given an estimate for $\hat{\sigma}_{\bar{X}}$, we can estimate the 95% CI for μ, assuming normality of X, as:

$$\hat{\mu} \pm t_{0.975,\,\mathrm{df}=n-1}\frac{S}{\sqrt{n}}$$

where $t_{0.975,\,\text{df}=n-1}$ is the t quantile function with $n-1$ degrees of freedom for probability 0.975. Write a new function that will calculate the 95% CIs of μ using this equation and your function from Part A.

- The t quantile function in R is `qt`

```r
CI_95 <- function(data_x, mu){
  CI <- numeric(length = 2)
  CI[1] <- mu - qt(0.975, df = (length(data_x) - 1)) *
    sample_standard_error(data_x)
  CI[2] <- mu + qt(0.975, df = (length(data_x) - 1)) *
    sample_standard_error(data_x)
  return(CI)
}
```

C. (1 Point) Use your function from Part B to estimate the 95% CIs around the sample mean for the Palmer Penguins body mass data for Chinstrap penguins.

```r
cs_penguins <- subset(penguins, penguins$species=="Chinstrap")
cs_penguins_mu <- mean(cs_penguins$body_mass_g)

CI_95(cs_penguins$body_mass_g, cs_penguins_mu)
```

```
[1] 3640.059 3826.117
```

95% Confidence Interval for Chinstrap Penguin Body Mass:

$[3640.059, 3826.1117]$

2. The Bootstrap

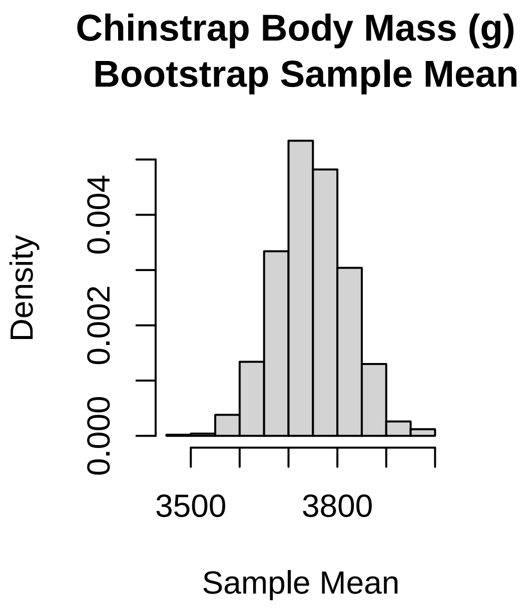
For all problems in this section we will use the Palmer Penguins body mass data for Chinstrap penguins.

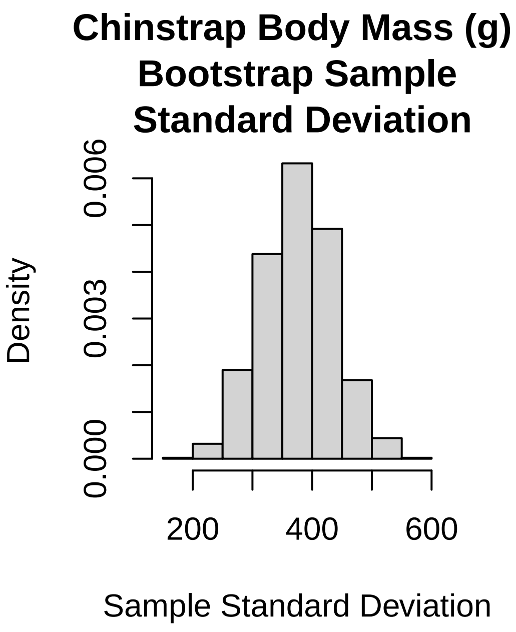
Problem 2.1 (5 Points)

Write two functions that take as input a vector of data x and generate (1) a bootstrap sample of the sample mean and (2) a bootstrap sample of the sample standard deviation. These functions should take as arguments a vector of data as well as the number of bootstrap replicates to run. Test your functions by generating 1000 bootstrap samples of the sample mean and standard deviation for the Palmer Penguins body mass data for Chinstrap penguins. Plot the resulting distributions and denote your actual sample mean and standard deviation in these plots using the `abline` function to draw a vertical rule.

- The `sample` function is your friend here.

```r
## bootsrap sample of the sample mean

bootstrap_sample_mean <- function(data_x, replicates){
  sample_means <- c()
  sample1 <- sample(data_x, 50, replace = TRUE) ##taking first sample

  for (sample in 1:replicates) { ##sampling from first sample
    sample_means[sample] <- mean(sample(sample1, 30, replace = TRUE))
  }
  hist(sample_means, xlab = "Sample Mean", main = "Chinstrap Body Mass (g) \n Bootstrap Sampl
}

## boostrap sample of the sample standard deviation

bootstrap_sample_sd <- function(data_x, replicates){
  sample_sds <- c()
  sample1 <- sample(data_x, 50, replace = TRUE) ##taking first sample

  for (sample in 1:replicates) { ##sampling from first sample
    sample_sds[sample] <- sd(sample(sample1, 30, replace = TRUE))
  }
  hist(sample_sds, xlab = "Sample Standard Deviation", main = "Chinstrap Body Mass (g) \n Boo
}

## testing
par(mfrow=c(1,2))
bootstrap_sample_mean(cs_penguins$body_mass_g, 1000)
bootstrap_sample_sd(cs_penguins$body_mass_g, 1000)
```





Problem 2.2 (5 Points)

Generate 95% bootstrap confidence intervals for the sample mean and sample standard deviation for the Palmer Penguins body mass data for Chinstrap penguins using the "percentile method" (the `quantile` function is your friend here). Then, use the function `bcajack` in the `bcaboot` package to generate 95% bootstrap confidence intervals using the "bias corrected and acceletated (BCa)" bootstrap method (you may want to use option `verbose=F`). Compare the intervals obtained by these two methods as well as your result from Problem 1.2 Part C. Do these various methods seem to agree or disagree?

3. Unit I Review

Problem 3.1 (10 Points)

Today we are going to analyze tree count data from [Barro Colorado Island](https://stri.si.edu/facility/barro-colorado) provided with the `vegan` package, a comprehensive community ecology package in R. This dataset describes the number of individuals per 1-hectare plot for 225 species of tree over 50 plots. You can load the data with:

```
data(BCI)
```

Choose one tree species (one column of the dataframe) to work with for this problem. The choice is up to you, but I suggest choosing a species with a decent number of observations. I will use the species *Faramea occidentalis* in my solution set for this problem (you are welcome to also use this species).

Using what you have learned so far:

A. (2 Point) Plot the empirical distribution of your data and its normal Q-Q plot.

B. (3 Point) Estimate the mean number of trees per plot, the 95% CIs of this mean based on standard error (assuming normality), and the 95% bootstrap CIs of this mean (specify the type of bootstrap you choose to apply). Also estimate the bootstrap bias of the mean.

C. (3 Points) Fit two distributions to this data using the MLE approach: (1) a Poisson distribution and (2) a negative binomial distribution. Plot your fitted distribution against your data and report your parameter estimates and standard errors from your maximum likelihood fit.

   - For this exercise, use the `fitdistr` function from the `MASS` package. Recall that you can use the command `?fitdistr` in the console to look up the help file for any function. For this problem, let the function choose the starting parameter values for optimization for you (you can omit the `start` argument).
   - For plotting, the functions `dpois` and `dnbinom` will be useful.
   - Make sure your plots have appropriate x and y axis limits so you can fully view your fitted distributions.

D. (2 Point) Discuss whether the Poisson or negative binomial fits the data better (justify based on plots **and** likelihoods).

Problem 3.2 (5 Points)

Consider a discrete random variable X with PDF:

X	-2	-1	0	1
$f(X)$	1/2	1/4	1/8	1/16

Do the following (1 Point each):

A. Plot the PDF and CDF of X

B. Calculate $E[X]$ and $Var(X)$

C. Plot the sampling distribution of $\hat{\mu}$ where $n = 30$ by repeatedly drawing samples from X and estimate the lower 0.025 and upper 0.975 quantiles for this distribution

    - The `prob` argument in the `sample` function may be useful here

D. Draw a set of $n = 30$ observations from X and calculate the 95% bootstrap CI for $\hat{\mu}$ from this sample. Compare to your result from Part C.

E. Which has greater probability: (1) drawing four observations in a row from X that all have a value of -2, or (2) drawing a single observation from X with a value of 1?

4. List of All Problems (Above, 35 Points)

- 1.1(5 Points)
- 1.2 (5 Points)
- 2.1 (5Points)
- 2.2 (5 Points)
- 3.1 (10 Points)
- 3.2 (5 Points)