Module 2 Assignment 2

Ellen Bledsoe

2025-03-17

Assessing Accuracy

Assignment Details

Purpose

The goal of this assignment is to use R to calculate and understand accuracy of estimates through standard error and confidence intervals.

Task

Write R code to successfully answer each question below or write text to successfully answer the question.

Criteria for Success

- Code is within the provided code chunks
- Code chunks run without errors
- Code produces the correct result
  - Code attempts will get half points
  - Code that produces the correct answer will receive full points
- Text answers correctly address the question asked

Due Date

Oct 22 before lab

Assignment Questions

Each question is worth 1 point unless otherwise noted.

Set-up:

For our assignment today, we will be using data from my PhD field site in southeastern Arizona in the Chiricahuas. In fact, we will be using some of the first data I ever collected for the project! To learn more about the Portal Project, check out the project's website. Researchers have been collecting data at this site monthly since 1977!

What you need to know for today is that this is small mammal (rodent) data.

The site has 24 plots, although we will only be using 18 of them today (6 of them are rodent removal plots, so they aren't very helpful in estimating rodent densities!).

As such, `n` will be 18 for this assignment. We do not have a `N`.

The first thing we need to do is download the data. My PhD lab group actually created our own R package for using the cleaned up data, so we are going to load that package, called `portalr`.

```
# you can ignore the message that pops up when you run this!
library(portalr)
```

Now that we have loaded the packages we need, we are going to download some rodent abundance data. This `abundance()` function downloads the entire dataset, so it might take a few minutes.

Let's take a look at what the dataset looks like.

```
head(abund)
```

```
##   period treatment plot BA DM DO DS NA OL OT PB PE PF PH PI PL PM PP RF RM RO
## 1     27   control    1  0  0  0  0  0  0  1  0  0  0  0  0  0  0  1  0  0  0
## 2     27   control    2  0  2  0  0  0  0  2  0  0  1  0  0  0  0  0  0  0  0
## 3     27   control    4  0  1  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 4     27   control    6  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 5     27   control    8  0  0  0  1  0  0  0  0  0  0  0  0  0  0  1  0  0  0
## 6     27   control    9  0  1  0  2  0  1  0  0  0  0  0  0  0  0  0  0  0  0
##   SF SH SO
## 1  0  0  0
## 2  0  0  0
## 3  0  0  0
## 4  0  0  0
## 5  0  0  0
## 6  0  0  0
```

We want to choose data only from a certain data collection period and only for certain species. To do this, we are going to filter the data to select trapping period "440" (the 440th time the site was trapped–this was back in 2015. We are over 500 trapping sessions now!) and get rid of the rodent removal plots.

We are also going to select only the columns that are relevant to us.

For this assignment, we are going to be calculating values for two different species of pocket mice: *Chaetodipus penicillatus* and *Chaetodipus baileyi*. In the Portal data, these species are abbreviated as `PP` and `PB`, respectively.

When we choose which columns we keep, we will keep the period, plot, and the PP and PB columns. These columns have the number of each species caught on each plot that trapping session.

```
# filter period 440, exclude "removal" treatment type, and select "PP" and "PB" column
abund440 <- abund[abund$period == 440 & abund$treatment != "removal",
                  c("period", "plot", "PP", "PB")]
```

The `abund440` object you just created by running the code chunk above is the data frame that we will use for the rest of the assignment.

Estimating the Mean

The first bit of this assignment is very similar to Module 2 Assignment 1, so if you're feeling stuck with the code, you might want to check that assignment out.

1. First things first, we want to create an object for our sample size. We don't know `N` in this scenario, but we do know `n`. Create the object `n`, which is the number of sample units that make up our sample. This will be the same for both species.

   To do this, use the `nrow()` function, which counts up the number of rows in a dataframe. The argument for `nrow()` is the name of the dataframe.

```
n <- nrow(abund440)
n
```

```
## [1] 18
```

Chaetodipus pencicillatus (PP) Data

We are going to start by calculating some population parameter estimates and measures of uncertainty for the PPs.

Calculating Estimates of Population Parameters

2. Create an object called `PP_mean` that contains the sample mean for the PPs

```
PP_mean <- mean(abund440$PP)
PP_mean
```

```
## [1] 11.05556
```

3. Create an object called `PP_var` which contains the sample variance. Hint: use the `var()` function.

```
PP_var <- var(abund440$PP)
PP_var
```

```
## [1] 15.70261
```

Calculating Standard Error for the Mean Reminder: because we don't know what `N` is, we don't need to use the population correction factor when calculating our standard error and whatnot.

4. Calculate the variance of the estimate of the mean for the PPs.

```
PP_var_ybar <- PP_var / n
PP_var_ybar
```

```
## [1] 0.8723675
```

5. Calculate the SE of the estimate of the mean.

```
PP_SE <- sqrt(PP_var_ybar)
PP_SE
```

```
## [1] 0.9340061
```

Calculating Confidence Intervals for the Mean Hooray, we've now calculated the sample mean, sample variance, and the standard error! Let's calculate a 95% confidence interval for our sample mean.

6. Create an object called `alpha95` to represent our alpha value for a 95% confidence interval. Because this is a constant, we can assign a number to it (take a look back at lecture notes).

```
alpha95 <- 0.05
alpha95
```

```
## [1] 0.05
```

7. Create an object called `df` for degrees of freedom. Degrees of freedom in this case are equal to the number of sample units in our sample (`n`) minus 1. Use the object `n` in your calculation.

```
df <- n-1
df
```

```
## [1] 17
```

We are going to use a function called `qt()` to find the t-value that we need. Assuming you've created the objects `alpha95` and `df` above, running this code chunk will give us our t-value, or critical value, for 95% CI.

```
t95 <- qt(1-(alpha95)/2, df)
t95
```

```
## [1] 2.109816
```

8. Using `t95` and the standard error that we calculated earlier, create an object called `half_width` which contains the half-width value for the 95% CI.

```
halfwidth_95 <- t95 * PP_SE
halfwidth_95
```

```
## [1] 1.970581
```

9. Calculate the upper and lower limits of the 95% confidence intervals. Use objects you've already created to calculate it.

```
upperCI_95 <- PP_mean + halfwidth_95
lowerCI_95 <- PP_mean - halfwidth_95

upperCI_95
```

```
## [1] 13.02614
```

```
lowerCI_95
```

```
## [1] 9.084975
```

10. What happens when we create 90% CI for the data? Let's calculate 90% CI. (3 points total)

    a. Input the correct value for `alpha90` and then run the code chunk to get the new t-value, `t90`.

```
alpha90 <- 0.10
alpha90
```

```
## [1] 0.1
```

```
t90 <- qt(1-(alpha90)/2, df)
t90
```

```
## [1] 1.739607
```

b. Now, using 't90', calculate the 90% CI. and compare to the 95% CI. Which one is wider? Why?

```
halfwidth_90 <- t90 * PP_SE

upperCI_90 <- PP_mean + halfwidth_90
lowerCI_90 <- PP_mean - halfwidth_90

upperCI_90
```

```
## [1] 12.68036
```

```
lowerCI_90
```

```
## [1] 9.430752
```

c. Compare the 90% CI to the 95% CI. Which one is wider? Why would expect that to be the case?

Answer:

Answer: 95% CI are wider; they are wider because we are willing to accept less uncertainty, so the range of values that must be included needs to be broader to ensure that we have a higher percentage (95 vs 90) of having the true parameter value included in the range.

Chaetodipus baileyi (PB) Data

Let's do the same thing as above for our *Chaetodipus baileyi* (PB) data.

Calculating Estimates of Population Parameters

11. Create an object called `PB_mean` that contains the sample mean.

```
PB_mean <- mean(abund440$PB)
PB_mean
```

```
## [1] 0.6666667
```

12. Create an object called `PB_var` which contains the sample variance. Hint: use the `var()` function.

```
PB_var <- var(abund440$PB)
PB_var
```

```
## [1] 1.647059
```

Calculating Standard Error for the Mean

13. Start by calculating the variance of the estimate of the mean

```
PB_var_ybar <- PB_var / n
PB_var_ybar
```

```
## [1] 0.09150327
```

14. Calculate the SE of the estimate of the mean.

```
PB_SE <- sqrt(PB_var_ybar)
PB_SE
```

```
## [1] 0.3024951
```

Calculating Confidence Intervals

15. Create 95% CI for the PB data. You should already have the t-value in your environment, so you don't need to go through calculating that again.

```
halfwidth_95_PB <- t95 * PB_SE

upperCI_95_PB <- PB_mean + halfwidth_95_PB
lowerCI_95_PB <- PB_mean - halfwidth_95_PB

upperCI_95_PB
```

```
## [1] 1.304875
```

```
lowerCI_95_PB
```

```
## [1] 0.02845785
```

16. Compare the ranges of the 95% CI for the PP and the PB data.

Which estimate of the mean seems to be more accurate? What are you using to make that determination? What value(s) is driving the difference you are seeing?

Answer: PB seems more accurate because the range of the confidence intervals is narrower than those for PPs. This is because the SE is lower, leading to smaller CIs. [Note: this is actually being driven by the differences in standard deviation since the sample sizes are the same, but they don't need to say that to get credit]

17. Right now, we can't calculate an estimate of the population total (abundance, represented by "tau hat") because we do not have a value for `N`.

    Let's say that we did have a value for `N` and wanted to calculate 95% confidence intervals for the total number of PPs at the site.

    What additional values (ones we haven't already calculated in the assignment) would we need to calculate to get the confidence intervals for our estimate of the abundance? Look back through the equations. Make sure you clearly state *all* of the values we would need to calculate.

Answer: First, we would need to calculate the estimate for the total number of PP (tau_hat)[using the mean we already calculated and N].

[Note: We generally use the same t-value for this calculation because the degrees of freedom don't change, but if students say we would need a new t-value based on the different degrees of freedom, don't penalize them for that]

To calculate the standard error for tau_hat, which would be used to calculate the half-width which would then allow us to calculate the upper and lower bounds of the confidence interval, we first need to calculate the variance of the estimate of tau hat. You can either use N and the variance of the estimate of the mean OR N, n, and sample variance to do that. We then take the square root, which gives us the standard error of the estimate of the total, which we then multiply by the t-value to get the half-width.