Module 3 Assignment 1

Ellen Bledsoe

2022-10-11

Simple Random Sampling – Rocks!

Assignment Details

Purpose

The goal of this assignment is to use R to calculate and understand accuracy of estimates.

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 is commented with brief descriptions of what the code does
- Code chunks run without errors
- Code produces the correct result
- Text answers correctly address the question asked

Due Date

March 13 by 2pm MST

Assignment Questions

Short Answer

Think back to how we sampled the rock data. Answer the following question in terms of our data (aka, "big square," "little squares," "dark rocks," etc.)

1. Define the *sample* standard deviation of the mean in terms of our data (not the equation but what we use to calculate standard deviation, what it tells us, and what it is an estimate of).

   Answer: the sample standard deviation is a measure of how much variation there is in the number of rocks counted in each little square. The sample sd is calculated from the 8 sample units (little squares) that we counted and is an estimate of *sigma*, the population standard deviation (big square).

2. Define the standard error of the estimate of the mean (again, in terms of our data and what it tells us).

   Answer: The SE of the estimate of the mean tells us how good our estimate of the mean (# of rocks per little square) is (AKA how much uncertainty there is in our estimate). It is calculated from the sample standard deviation divided by the square root of the number of sample units (little squares sampled).

3. How do sample standard deviation and standard error differ? (If you've explained that above, reiterate it here.)

   Answer: The sample standard deviation tells us something inherit about the population–the amount of variation in the number of dark rocks per little square. On the other hand, the standard error tells us how much uncertainty we have in our estimate of the number of rocks in each little square.

4. What would happen to the sample standard deviation if we sampled 20 sampling units (little squares) instead of only 8? Would it increase, decrease, or stay about the same? How about the standard error? Explain your rationale for both.

   Answer: The sample standard deviation estimate would stay about the same because it is an estimate of the amount of variation in the population, something that is a property of the population and won't change with increased sampling. The standard error, however, would decrease. That is because we now have more information about our population parameter that we are estimating (mean), so we have less uncertainty in the estimate (AKA smaller SE).

Calculations

Each question is worth 1 point unless otherwise noted.

Set-up:

(a) Get your (or your group's) simple random sampling rock data.
(b) Make sure it has been saved as a .csv file. If it hasn't, you will need to do so using "Save As".
(c) In RStudio Cloud, using the "Upload" button in the lower right-hand panel, upload your .csv file.

Load the tidyverse. For this assignment, we will only be using this to read in our .csv file.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

5. Use the `read_csv()` function from the tidyverse to read in your rocks file. Save this data in an object called `rocks`.

   Below, replace the file name with the file name of your rock data.

```r
rocks <- read_csv("../data_raw/example_data.csv") %>% drop_na()
```

```
## Rows: 8 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): sample_unit, abund_dark_stones
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

6. Create the object N, which is the number of units in your *sampling frame*.

```r
N <- 64
```

7. Create the object `n`, which is your *sample size,* the number of sample units that make up your sample. To do this, use the `nrow()` function

   (Hint: if you don't remember how to use the `nrow()` function, run `?nrow`)

```r
n <- nrow(rocks)
```

Estimating the Mean

One way we can calculate values is to call the column that we want and then save that value to an object.

For example, let's say we wanted to calculate the median of our dark stone abundance data. We could run the following line of code: `rocks_median <- median(rocks$abund_dark_stones)`

Typically, I prefer using the `tidyverse` functions, but for performing calculations like we are doing in this assignment, working with vectors (as opposed to data frames) is easier; the code above does that for us.

8. Create an object called `rocks_mean` that contains the sample mean (AKA the mean number of dark stones per sample unit). Follow the code example above.

```r
rocks_mean <- mean(rocks$abund_dark_stones)
```

9. Create objects called `rocks_var` and `rocks_sd` which contain the sample variance and sample standard deviation, respectively. Hint: use `var()` and `sd()` functions.

```r
rocks_var <- var(rocks$abund_dark_stones)
rocks_sd <- sd(rocks$abund_dark_stones)
```

Calculating Standard Error for the Mean

R doesn't have a function for calculating the standard error (SE), so let's calculate it ourselves using the equation from class.

10. First, do you remember that finite population correction factor? We use that to modify SE when we know what N is. Let's create an object called `pop_correction` using that formula: (N-n)/N

    (Hint: use the objects we've already created (`N`, `n`; don't enter any numbers directly)

```
pop_correction <- (N-n)/N
```

11. Why are we using the finite population correction factor in our formulas?

    Answer: We know the total number of sample units (64), so we need to adjust for that.

12. Using the objects you've already created and the formula (check the lecture slides from this module!), create an object called `rocks_var_ybar` which contains the variance of the *estimate* of the mean.

    Remember, this is different from the sample variance that you've already calculated, but that sample variance is included...

```
rocks_var_ybar <- pop_correction * (rocks_var/n)
```

13. Calculate the SE of the estimate of the mean by taking the square root of the `rocks_var_ybar` value. Call this object `rocks_SE`.

```
rocks_SE <- sqrt(rocks_var_ybar)
```

Estimating the Population (and Accuracy of the Estimate)

We've calculated standard error (SE) for the estimate of the mean above. Now, let's calculate them for the estimate of the population abundance (tau hat)

14. First, let's calculate the total number of rocks in the population. Call this `rocks_total`. Use objects you've already created.

```
rocks_total <- N * rocks_mean
```

15. Now calculate the variance of the estimate of the population total, `rocks_var_total`.

```
rocks_var_total <- N * (N-n) * (rocks_var/n)
```

16. Calculate the SE of the estimate of the population total, `rocks_SE_total`.

```
rocks_SE_total <- sqrt(rocks_var_total)
```

Bringing It All Together

17. Write a short paragraph (3-5 sentences) putting the values that we've just calculated into context for our dark rocks sampling adventure. Be sure to answer the following questions. (3 points)

- According to your calculations, approximately how many dark rocks were there in your big square? How confident are you in that estimate?
- What is your estimate of density of dark rocks (average # dark rocks per little square)? How much variation do you estimate is there in the number of dark rocks per little square? How confident are you in your estimate of the density?

Answer:

- number of dark rocks in big square = `rocks_total`
- confidence is estimate of number of dark rocks total = `rocks_SE_total` or `rocks_var_total`
- density estimate = `rocks_mean`
- variation in density = `rocks_var` or `rocks_sd`
- confidence in density estimate = `rocks_var_ybar` or `rocks_SE`