Module 2 Assignment 3

Ellen Bledsoe

2022-10-18

Stratified Random Sampling

Assignment Details

Purpose

The goal of this assignment is to understand, apply, and interpret a stratified random sampling method and calculations.

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

March 27 before lab

Assignment Questions

We have been asked to go out and sample for grasshoppers. The location we are sampling has 3 different vegetation types: riparian, shrubland, and grassland. We have decided to use a stratified random sample design for our survey.

1. In 1-2 sentences, explain why we have chosen stratified random sampling as apposed to simple or systematic? What benefit does it provide us?

   Answer:

Set-up

Load the `tidyverse` package. We won't be using it *too* much this time around, but we need it a couple times.

```
library(tidyverse)
```

Part 1: Allocation of Sample Units

Let's say we have a sampling frame with 1000 sample units. We can sample 100 of those units. We have our 3 strata which we have determined via vegetation type (we will called them A, B, and C for simplicity for the rest of the assignment).

How do we decide how many sampling units to sample from each stratum?

2. Create an object, `N`, which represents the total number of sample units in our *sampling frame*. We also need to create `n`, which is the total number of sample units in our *sample*.

```
N <- 1000
n <- 100
```

We have determined that Stratum A has 200 sampling units, Stratum B has 500 sampling units, and Stratum C has 300 sampling units.

3. Add the correct values to the lines of code below to create objects for future use.

```
N_A <- 200
N_B <- 500
N_C <- 300
```

Due to a pilot study done last year, we have an estimate of the amount of variation (standard dev.) for each stratum.

```
sd_A <- 9
sd_B <- 4
sd_C <- 7
```

We also have an estimate of the time-cost (hours) it takes to survey each unit.

```
cost_A <- 1.5
cost_B <- 2
cost_C <- 4
```

Optimal Allocation We want to use the optimal allocation equation to calculate how many sample units each stratum should have. Let's do this incrementally.

4. Start by calculating the numerators (N_i * s_i / square root of c_i) for each stratum. Remember to use objects you've created rather than typing out specific numbers into the equation.

```
opt_allo_A <- N_A * sd_A / sqrt(cost_A)
opt_allo_B <- N_B * sd_B / sqrt(cost_B)
opt_allo_C <- N_C * sd_C / sqrt(cost_C)
```

5. Calculate the denominator of the equation by adding all of the values above together.

```
opt_allo_total <- opt_allo_A + opt_allo_B + opt_allo_C
```

6. Calculate how many sampling units we should optimally sample in each stratum. Use the `round()` function to round to the nearest whole number.

```
n_A <- n * (opt_allo_A / opt_allo_total) # 37
n_A <- round(n_A)
# OR #
n_B <- round(n * (opt_allo_B / opt_allo_total)) # 36
n_C <- round(n * (opt_allo_C / opt_allo_total)) # 27
```

7. If we did not have the pilot study telling us about the variation of the populations in each strata, we could have allocated very differently:

   - stratum A = 24
   - stratum B = 53
   - stratum C = 23

   Why is this? What would this allocation have meant for the precision of our estimates?

   Answer:

Part II: Stratified Random Sampling

Data Hooray! We have gone out and sampled the number of grasshoppers per sample unit, allocating our sample units optimally (including variation in population).

Now we want to calculate our population parameters:

- mean grasshoppers per sampling unit (`y_bar`)
- total grasshoppers in the population (`tau_hat`)

Let's load our "grasshoppers.csv" data into our workspace using the `read_csv()` function and save the data frame as an object called `grasshoppers`.

```
grasshoppers <- read_csv("../data_raw/grasshoppers.csv")
```

```
## New names:
## Rows: 100 Columns: 3
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): stratum dbl (2): ...1, abund
## 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.
## * `` -> `...1`
```

We can explore the data frame to get an idea of what each row and column represent.

```
str(grasshoppers)
```

```
## spc_tbl_ [100 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1   : num [1:100] 1 2 3 4 5 6 7 8 9 10 ...
##  $ stratum: chr [1:100] "A" "A" "A" "A" ...
##  $ abund  : num [1:100] 34 26 32 29 35 38 38 28 31 24 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_double(),
##   ..   stratum = col_character(),
##   ..   abund = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Run the following code chunk. What do these numbers represent?

(Hint: do these numbers match your values from question 7? If not, you probably want to revisit your calculations up in the previous section!)

```r
grasshoppers %>%
  group_by(stratum) %>%
  count(stratum)
```

```
## # A tibble: 3 x 2
## # Groups:   stratum [3]
##   stratum     n
##   <chr>   <int>
## 1 A          37
## 2 B          36
## 3 C          27
```

Based on what I've taught you in R thus far, the easiest way to do our calculations will be to have three different data frames, one for each stratum.

```
stratumA <- grasshoppers %>% filter(stratum == "A")
stratumB <- grasshoppers %>% filter(stratum == "B")
stratumC <- grasshoppers %>% filter(stratum == "C")
```

That said, if you're feeling ambitious and want to *try* to use the `tidyverse` and a lot of `group_by()`, `mutate()`, and `summarise()` functions, be my guest! To be clear, though, that is neither required nor expected.

Calculations NOTE: Open up your script from Module 2, Assignment 2—it's about to come in handy! I would also recommend opening up the lecture slides.

8. Before we begin calculating, stop for a moment to reflect and predict. Take a minute to skim through the lecture slides and remind yourself of the process of how we go about calculating the *overall* population mean and total abundance when we use stratified sampling. Describe (either in sentences or in bullet points) how you think we will proceed.

   Note: This question in graded for completion only, not for accuracy.

   Answer:

Step 1: Calculatings Sample Statistics (per stratum) For our first step, we will calculate for each stratum the mean number of grasshoppers, the total number of grasshoppers, and the sample variances.

9. Use the code from Module 2 Assignment 2 to calculate the following for each stratum:

   a. mean grasshoppers per stratum (`y_bar`) (1 point)

```
# take the mean of the `abund` column
mean_A <- mean(stratumA$abund)
mean_B <- mean(stratumB$abund)
mean_C <- mean(stratumC$abund)
```

   b. total grasshoppers per stratum (`tau_hat`) (1 point)

```
tau_hat_A <- N_A * mean_A
tau_hat_B <- N_B * mean_B
tau_hat_C <- N_C * mean_C
```

   c. sample variance per stratum (`var` or s^2) (1 point)

```
# similar to (a) but calculating the variance instead of the mean
var_A <- var(stratumA$abund)
var_B <- var(stratumB$abund)
var_C <- var(stratumC$abund)
```

Step 2: Calculating Uncertainty (per stratum) Our next step is to calculate the uncertainty in our estimates for each stratum.

10. To do this, calculate the following:

    a. variance of the estimate of the mean per stratum (var_y_bar). Remember that we need to use the finite population correction factor and to use only objects already created, not plain numbers! (2 points)

```
pop_correction_A <- (N_A-n_A)/N_A
pop_correction_B <- (N_B-n_B)/N_B
pop_correction_C <- (N_C-n_C)/N_C

var_y_bar_A <- pop_correction_A * (var_A / n_A)
var_y_bar_B <- pop_correction_B * (var_B / n_B)
var_y_bar_C <- pop_correction_C * (var_C / n_C)
```

    b. variance of the estimate of the total abundance (var_tau_hat). You might need to go fishing for this equation in one of the earlier powerpoints. (1 point)

```
var_tau_hat_A <- N_A^2 * var_y_bar_A
var_tau_hat_B <- N_B^2 * var_y_bar_B
var_tau_hat_C <- N_C^2 * var_y_bar_C
```

Step 3: Calculating Total Abundance Statistics Congratulations! The hardest parts of the calculation are over (thank goodness, right?).

Next, we will calculate the total abundance and the variance of the total abundance.

11. Calculate the following:

    a. total abundance (1 point)

```
tau_hat <- tau_hat_A + tau_hat_B + tau_hat_C
```

    b. variance of the total abundance estimate (1 point)

```
var_tau_hat <- var_tau_hat_A + var_tau_hat_B + var_tau_hat_C
```

Step 4: Calculating Population Means, etc. Finally, we can now calculate the overall population mean and variance of the estimate of the overall population.

12. Calculate the following:

    a. overall population mean (1 point)

```
y_bar <- tau_hat / N
```

    b. variance of the estimate of the overall population mean (1 point)

```
var_y_bar <- var_tau_hat / N^2
```

Conclusion Ok, that was a lot of calculating! Let's summarize and make sure we know what just happened. We were ultimately able to calculate estimates the parameters we were interested in: the population mean and the population total along with the amount of uncertainty in both of those estimates.

12. Based on our calculations above, answer the following questions with the name of the R object and its corresponding value for each calculation. (2 points)

    a. the number of grasshoppers per sample unit: `y_bar`, 36.12
    b. total number of grasshoppers in the population: `tau_hat`, 36109
    c. measure of uncertainty in the # of grasshoppers per sample unit: `var_y_bar`, 0.357
    d. measure of uncertainty in the total # of grasshoppers: `var_tau_hat`, 357411.9