Lab 3: Partial Key

Group Member Names - here

Lab Overview

All students attending class in the group can turn in a single document with each participants name. Students not attending class will need to complete their own lab.

You have been hired to consult on a project by the Montana Department of Natural Resources. The goal is to collect data on the abundance of the Western meadowlark across a study area. Your sampling frame consists of two hundred plots of land and your budget allows sampling of twenty plots. There are four terrain types within the study area:

- desert (40 plots)
- forest (40 plots)
- wetland (20 plots)
- prairie (100 plots)

The dataset can be obtained with the following R code.

contained in the file: 'birdsurvey.csv' which can be accessed on D2L.

```r
set.seed(09232019)
birds <- read.csv('http://math.montana.edu/ahoegh/teaching/stat446/birdsurvey.csv', header = T)
population_total <- birds %>% summarize(t = sum(bird.counts)) %>% select(t) %>% pull()
```

1.

a. (4 points)

Take a SRS of size 20 and compute an approximate sampling distribution (by repeated samples) of the population total.

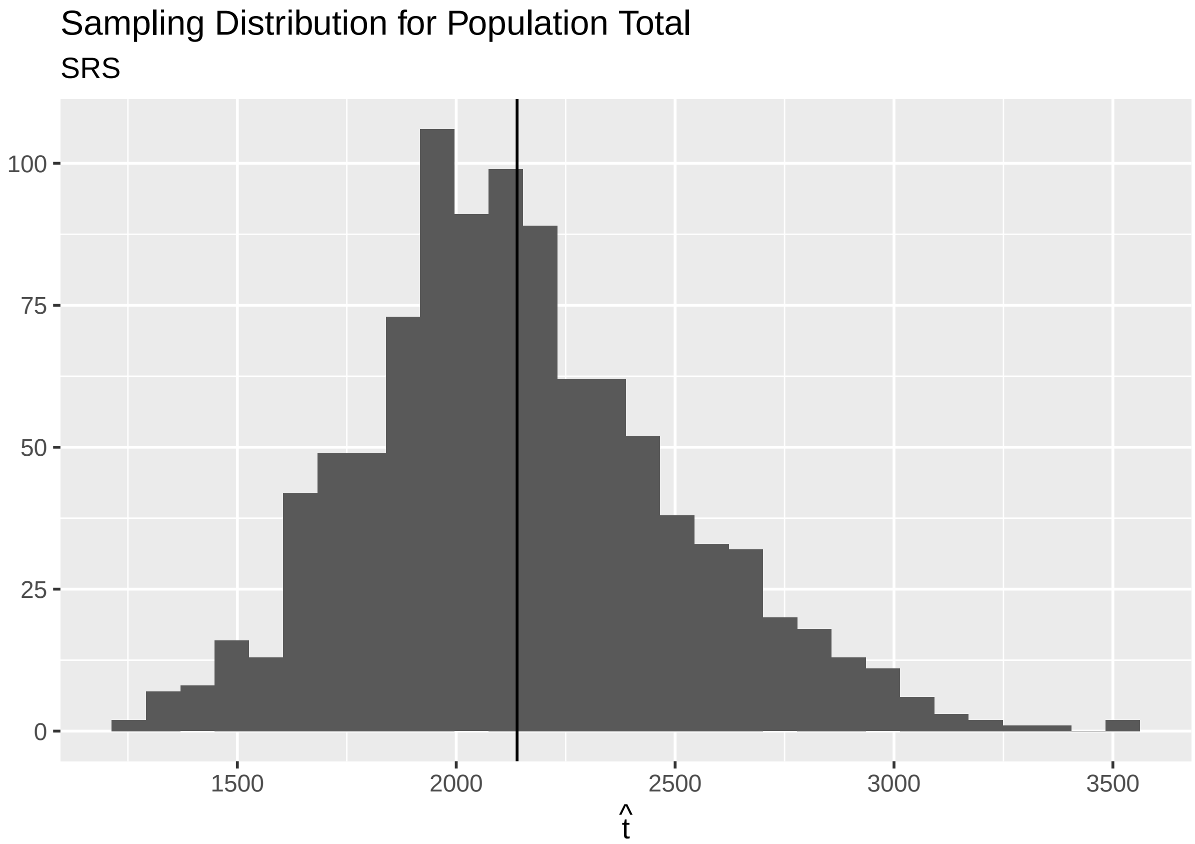
```r
samples_srs <- replicate(1000, sample(birds$bird.counts, size = 20)) %>% colMeans() * 200

# or equivalently with a loop
samples_srs <- rep(0, 1000)
for (iter in 1:1000){
  samples_srs[iter] <- mean(sample(birds$bird.counts, size = 20)) * 200
}
```

b. (3 points)

Plot the approximate sampling distribution and the true population total.

```r
tibble(val = samples_srs, label ='samples') %>% ggplot(aes(x = val)) + geom_histogram(bins=30) +
  xlab(expression(hat(t))) + ylab('') + ggtitle('Sampling Distribution for Population Total',
  subtitle = "SRS") + geom_vline(xintercept = population_total)
```



c. (3 points)

Compute the MSE of the estimator.

```
mse_srs <- mean((samples_srs - population_total)^2)
```

The MSE of the estimator is 1.31209×10^5. Often the square root of the MSE (the rMSE) is more intuitive. This value is 362

2.

a. (4 points)

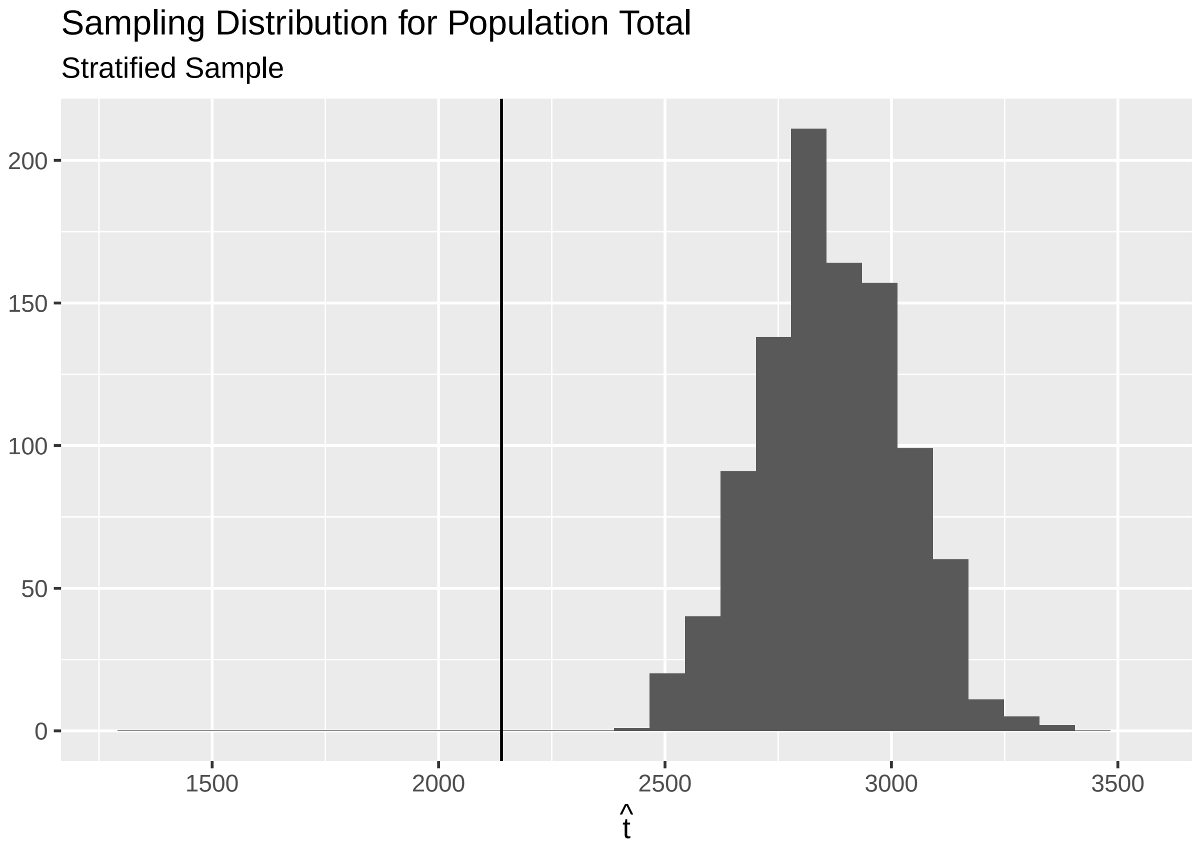
Take a stratified random sample where five samples are drawn from each terrain type. Compute the MSE of this estimator.

```
#Hint this will take one sample
samples_strat <- rep(0, 1000)
for (iter in 1:1000){
  samples_strat[iter] <- birds %>% group_by(terrain) %>% sample_n(5) %>% ungroup() %>% select(bird.coun
}
```

b. (3 points)

Plot the approximate sampling distribution and the true population total.

```
tibble(val = samples_strat, label ='samples') %>% ggplot(aes(x = val)) + geom_histogram(bins=30) +
  xlab(expression(hat(t))) + ylab('') + ggtitle('Sampling Distribution for Population Total',
  subtitle = 'Stratified Sample') + geom_vline(xintercept = population_total) + xlim(range(samples_srs)
```



c. (3 points)

Compute the MSE of the estimator.

```
mse_strat <- mean((samples_strat - population_total)^2)
```

The mse of the stratified sample is 5.52565×10^5.

3. (5 points)

a. (4 points)

What method was most effective in terms of MSE? What shortcomings did the other methods have?

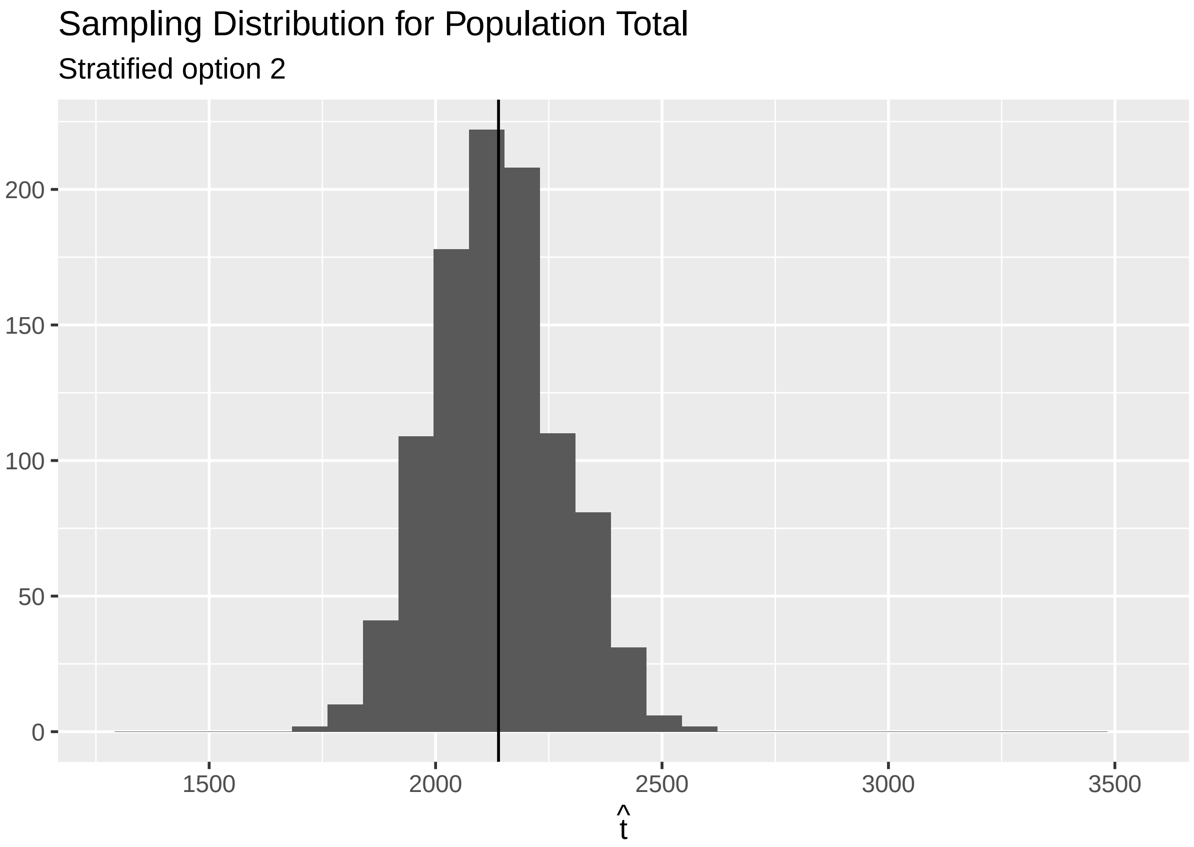
The SRS was much more effective with a mse of 1.31209×10^5*, whereas the stratified sample resulted in an mse of* 5.52565×10^5*. The shortcoming with the stratified sample was that it was not representative and did not have the appropriate weighting to account for the sampling approach.*

b. (4 points)

Suppose another option was to use a stratified sample where 10 % of the sampling units in each strata are selected. Do you think this will perform better or worse than the option in part 2, why?

```
samples_strat2 <- rep(0, 1000)
for (iter in 1:1000){
  samples_strat2[iter] <- birds %>% group_by(terrain) %>% sample_frac(.1) %>% ungroup() %>%
    select(bird.counts) %>% summarize(ybar = mean(bird.counts)) %>% pull() * 200
}

tibble(val = samples_strat2, label ='samples') %>% ggplot(aes(x = val)) + geom_histogram(bins=30) +
  xlab(expression(hat(t))) + ylab('') + ggtitle('Sampling Distribution for Population Total',
  subtitle = 'Stratified option 2') + geom_vline(xintercept = population_total) + xlim(range(samples_srs
```



```
mse_strat2 <- mean((samples_strat2 - population_total)^2)
```

The approach taking 10% of the samples from each group performs substantially better. This is because we are taking a representative sample of each group. Alternatively, the sample previously selected with the other approach could also be weighted – more on that later.

```
knitr::kable(tibble(method = c('srs','stratified 1', 'stratified 2'),
      mse = c(round(mse_srs), round(mse_strat), round(mse_strat2))))
```

method	mse
srs	131209
stratified 1	552565
stratified 2	19657