Sampling Project #3

Stefano Fochesatto, Thomas House, Emily Richmond

Estimating the Total Number of Pages in the Q-Z section on the Fifth Floor of the Rasmussen Library

Stratified Two-Stage Cluster Sampling

On first inspection of the study area our group realized that the population of books was highly variable with whole aisle with no books so we knew that we would have to stratify. A sampling plan was proposed where we could cluster each strata by bookcase, and inside each bookcase cluster we would cluster the pages by books and record their total, i.e. perform a two-stage cluster sample. To avoid having to count the total number of books in the whole study area we used the two-stage cluster sample, unbiased estimator for the total in each strata. To make the sampling inside of each bookcase cluster easier for our group we devised a 1 in 10 systematic sample. Since the length of a book has no relation with the author's last name we can assume that the population is randomly ordered and treat the sample the same as a SRS.\

Sampling Plan

First we can stratify the population of bookcases by sparsity. In each strata we will perform a two-stage cluster sample. In the first stage we take an SRS of the bookcases. In the second stage we take a 1 in 10 systematic sample and compute the total number of pages for each strata. Note that this means that we will need to count the total number of books in each bookcase that we sample.\ Recall the stratified sample estimator for the total of a population,

$$\hat{\tau}_{total} = \sum_{i=1}^{K} \hat{\tau}_i$$

$$Var(\hat{\tau}_{total}) = \sum_{i=1}^{K} Var(\hat{\tau}_i)$$

Here there would be N strata split by bookcase sparsity. N_i is bookcase totals in the ith strata. n_i is number of sampled bookcases in the ith strata. $\hat{\tau}_i$ is the total pages per strata computed by the two-stage cluster sample estimator. $Var(\hat{\tau}_i)$ is the variance from that estimator.

In the two-stage cluster sample we let N be the total number of bookcases,n is the number of sampled bookcases, M_i is the total number of books in the ith bookcase, m_i is the sample of books in the ith bookcase, $\bar{y}_i$ is the mean pages per book, and $\hat{s}_i^2$ is the variance of $\bar{y}_i$.

$$\hat{\tau}_i = \frac{N}{n} \sum M_i \bar{y}_i$$

Here is the variance estimator. Note we are substituting $n/\sum M_i$ since we don't know M the total number of books.

$$V(\hat{\tau}_i) = M^2 \left(\frac{N-n}{N} \left(\frac{N}{M} \right)^2 \frac{MSE}{n} + \frac{1}{nN} \left(\frac{N}{M} \right)^2 \sum M_i^2 \frac{M_i - m_i}{M_i} \frac{\hat{s}_i^2}{m_i} \right),$$

Simplifying we get,

$$V(\hat{\tau}_i) = (N-n)N \frac{MSE}{n} + \frac{N}{n} \sum M_i^2 \frac{M_i - m_i}{M_i} \frac{\hat{s}_i^2}{m_i}.$$

Where the MSE is equal to,

$$MSE = \frac{\sum(M_i\bar{y}_i - \frac{\hat{\tau}_i}{N})^2}{n-1}.$$

Strata Determination and Bookcase Sampling

On our initial survey of the study area we counted the total number of bookcases in each column and decided that we would split the bookcases into 4 strata: Strata 1 having zero books, Strata 2 having around one-thirds of the bookcase full, Strata 3 having around two-thirds of the bookcase full, and Strata 4 being almost entirely full of books. The initial count of bookcases was used to devise a system for identifying each bookcase via coordinates with the origin at the northwestern most point of the Q-Z Block. The following figure describes how the coordinate system works,

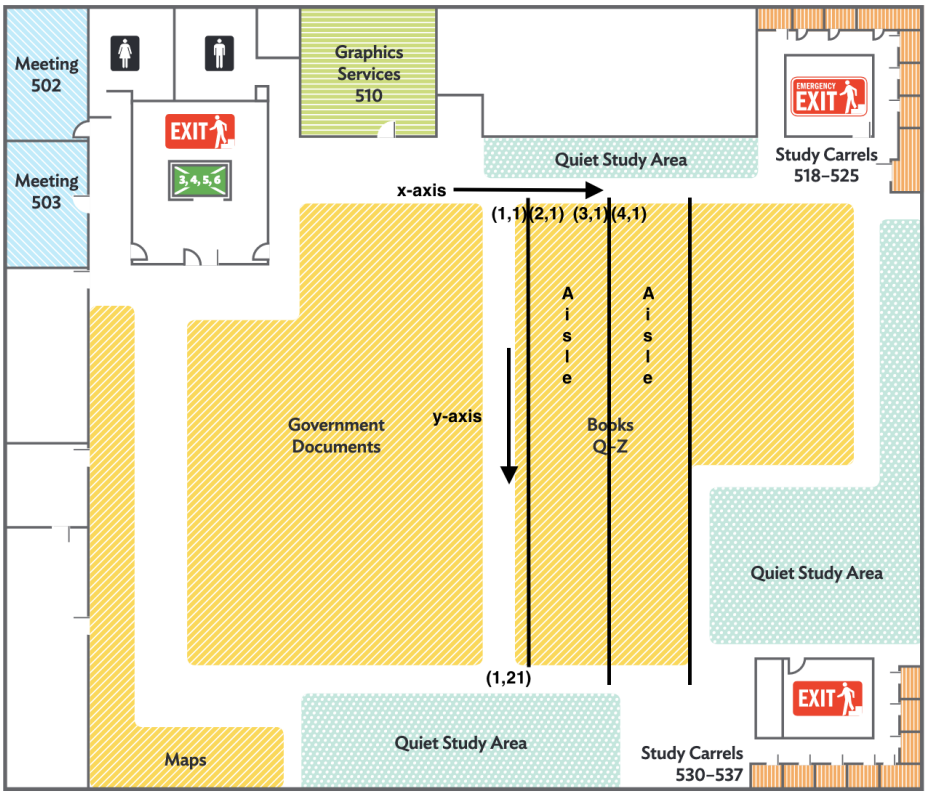

Figure 1: Coordinate System for Mapping Bookcases

This system allowed us to generate a random sample of bookcases, as well as making it easier to rank each bookcase in order to stratify. The following code shows how the dataframe containing the coordinates was made,

```
## Bookcase totals by aisle from west to east
BookcasesByAisle <- c(21, 21, 26, 26, 26, 26,
                      26, 26, 26, 26, 21, 21,
                      26, 26, 22, 22, 15, 15,
                      15, 15, 13, 13, 15, 15,
                      15, 15, 12, 12)

## Generate Bookcase Coordinates
```

```
BookcaseCoordinatex <- vector()
BookcaseCoordinatey <- vector()
for (i in 1:length(BookcasesByAisle)){
  for (j in 1:BookcasesByAisle[i]){
    BookcaseCoordinatex <- append(BookcaseCoordinatex, i)
    BookcaseCoordinatey <- append(BookcaseCoordinatey, j)
  }
}


df <- data.frame(BookcaseCoordinatex, BookcaseCoordinatey)
head(df)
```

```
##   BookcaseCoordinatex BookcaseCoordinatey
## 1                   1                   1
## 2                   1                   2
## 3                   1                   3
## 4                   1                   4
## 5                   1                   5
## 6                   1                   6
```

After generating the coordinates we were able to rank each bookcase by walking down the aisles, filling in a value for rank for each coordinate pair. This data is was then imported into an r dataframe and subsetted by rank to produce our bookcase samples. We decided that every group member would sample 1 bookcase from each strata. The following code shows our samples were taken.

```
## Data input has all bookcases represented as x,y
## coordinate pair, along with their strata assignment(rank)
## Rank 1 is for empty
## Rank 2 is for low
## Rank 3 is for med
## Rank 4 is for high
BookcaseRankings<- read.csv('LibraryBookcaseRankings.csv')
head(BookcaseRankings)
```

```
##   X Y Rank
## 1 1 1    1
## 2 1 2    1
## 3 1 3    1
## 4 1 4    3
## 5 1 5    4
## 6 1 6    4
```

```
## Subsetting the dataframe by ranking.
Strata_1 = BookcaseRankings[BookcaseRankings$Rank == 1,]
Strata_2 = BookcaseRankings[BookcaseRankings$Rank == 2,]
Strata_3 = BookcaseRankings[BookcaseRankings$Rank == 3,]
Strata_4 = BookcaseRankings[BookcaseRankings$Rank == 4,]

## Setting Seed to reproduce samples
set.seed(1)
### Sampling relevant dataframes.
## Pulling Index.
Strata_2_SampleIndex <- sample(nrow(Strata_2), size = 3)
Strata_3_SampleIndex <- sample(nrow(Strata_3), size = 3)
Strata_4_SampleIndex <- sample(nrow(Strata_4), size = 3)
```

```
## Assigning Row Values.
Strata_2_Sample <- Strata_2[Strata_2_SampleIndex,]
Strata_3_Sample <- Strata_3[Strata_3_SampleIndex,]
Strata_4_Sample <- Strata_4[Strata_4_SampleIndex,]

Stefano <- rbind(Strata_2_Sample[1,],Strata_3_Sample[1,], Strata_4_Sample[1,] )
Emily <- rbind(Strata_2_Sample[2,],Strata_3_Sample[2,], Strata_4_Sample[2,] )
Thomas <- rbind(Strata_2_Sample[3,],Strata_3_Sample[3,], Strata_4_Sample[3,] )

Stefano
```

```
##      X  Y Rank
## 456 21  8    2
## 305 13 13    3
## 42   2 21    4
```

```
Emily
```

```
##      X  Y Rank
## 26   2  5    2
## 494 24  5    3
## 544 27 10    4
```

```
Thomas
```

```
##      X  Y Rank
## 237 10 13    2
## 130  6 10    3
## 123  6  3    4
```

Computing the Estimator

Below is a function that computes the two-stage cluster sample total and variance based on the unbiased total estimator.

```
TwoStageClusterSample <- function(N,M,y){
  ## This function takes a Total number of first stage
  ## clusters N(number of bookcases),
  ## Population total inside the sampled second stage
  ## cluster M(Booktotal inside sampled bookcases),
  ## and the second stage sample y, as a dataframe
  ## with books as rows and columns as bookcase samples.

  ## Extracting number of samples n
  n = length(y)
  ## Computing yHats for each sample
  yHats = colMeans(y, na.rm = TRUE);
  ## Computing the Tau estimator
  TauHat = N*sum(M*yHats)/n;
  ## Computing the variance for each sample
  s_squared = diag(var(y,  na.rm = TRUE))
  ## Pulling number of samples taken, at each bookcase.
  m = colSums(!is.na(y))
  ## Computing the MSE
  MSE = sum((M*yHats - (TauHat/N))^2) / (n-1);
```

```
  ## Computing The variance
  Cluster2ClusterVar = ((N-n)/N)*(N^2)*(MSE/n)
  InsideClusterVar = (1/(n*N))*(N^2)*sum(M^2*((M - m)/M)*(s_squared/m))
  TotalVar = Cluster2ClusterVar + InsideClusterVar

  # Returning the results as a list.
  RList <- list('tHat' = TauHat, 's_squared' = TotalVar)
  return(RList)
}
```

After the data was collected we collated it inside the 'SampledData.csv' file, where the first line after the header was designated for the total books in the bookcase. The following script reads the data from the 'SampledData.csv' which contains data from our whole group and computes the total estimator.

```
################### READING IN DATA ##########################
SamplingData <- read.csv('SampledData.csv')
head(SamplingData)
```

```
##   Strata.2 Strata.2.1 Strata.2.2 Strata.3 Strata.3.1 Strata.3.2 Strata.4
## 1      112        107        128      185        177        169      189
## 2      118        392        386      264        448        613      297
## 3      167        300        637      158        902        352      528
## 4      271        267         81      288         71        736      356
## 5      247        220        144       68        483        285      506
## 6       51        160        109       88        404        202      377
##   Strata.4.1 Strata.4.2
## 1        238        150
## 2        180        642
## 3        352        166
## 4        268        386
## 5        495        288
## 6        106        258
```

```
## Assigning Bookcase Totals
N <- c(nrow(Strata_2),nrow(Strata_3),nrow(Strata_4))
## Assigning Book Totals
M <- SamplingData[1,]
## Removing Bookcase Totals From DataFrame
SamplingData <- SamplingData[2:nrow(SamplingData),]
## Subset by Strata
Strata_2_Data = SamplingData[,1:3]
Strata_3_Data = SamplingData[,4:6]
Strata_4_Data = SamplingData[,7:9]

################### COMPUTING ESTIMATOR ######################
## Compute 2-Stage Cluster Sample Estimators
TwoStage_2 <- TwoStageClusterSample(N[1], M[1:3], Strata_2_Data)
TwoStage_3 <- TwoStageClusterSample(N[2], M[4:6], Strata_3_Data)
TwoStage_4 <- TwoStageClusterSample(N[3], M[7:9], Strata_4_Data)

## Compute Stratified Estimator.
Est.Tau <- sum(c(TwoStage_2$tHat, TwoStage_3$tHat, TwoStage_4$tHat))
Est.SE <- sqrt(sum(c(TwoStage_2$s_squared, TwoStage_3$s_squared, TwoStage_4$s_squared)))
CI95 <- c(Est.Tau + 2*Est.SE, Est.Tau - 2*Est.SE)
```

```
#> Est.Tau
#[1] 25945611
#> Est.SE
#[1] 1823375
#> CI95
#[1] 29592360 22298862
```

Conclusion

Our group found that there were approximately 25,945,611 million pages with a 95 percent confidence interval of (29,592,360 , 22,298,862). In the middle of the project we discovered that the reason a whole section of the study area was empty was because library staff was in the middle of moving books to the east side of the study area. That is why there is a whole section of bookcases ranked 1 followed by another section ranked 4. Luckily it didn't affect the samples that we took, but if it had we likely would have had to re-stratify the bookcases and sample again. Our group also explored the possibility of post-stratification, by finding the proportions of each strata through a SRS of bookcases on the entire study area. We decided against this since, it turns out ranking all the bookcases isn't very difficult. Had the move affected our samples, post-stratification might have been a good way to resample.