STA322 Project 1

Sherry Huang, Eddie Yang

March 8, 2019

Introduction

This study aims to examine the number and cost of required textbooks for Natural Sciences courses at Duke University. Specifically, the average and total cost of required books for courses under 700 level are estimated assuming they are bought new at the Duke textbook store. This report includes details of the study design, procedures of data collection, and analysis of estimation results using the `survey` package in R.

Study Design and Data Collection

According to the Trinity College of Arts and Sciences webpage, Natural Sciences at Duke University include the following nine departments: Biology, Chemistry, Computer Science, Evolutionary Anthropology, Mathematics, Physics, Psychology, Neuroscience, and Statistical Science. Given this existing intuitive grouping of courses, stratified sampling is the most suitable sampling method in this case. Each department is a stratum which may have similar textbook requirements and price levels for its courses. Courses were randomly sampled from each of the nine departments, with sample sizes for each department proportionally allocated based on stratum size and population size.

Using the DukeHub webiste as our source of information for course listings, we first recorded all course numbers of unique courses under 700 level and their total counts for each Natural Sciences department. Each course that has multiple lecture sections or complementary lab sessions was counted only once; each course that has combined sections between undergraduate and graudate students was also counted only once for the undergraduate section. Courses under one Natural Sciences department that are cross listed in other Natural Sciences department were counted towards the first department that the courses were listed under (the counting process took place in an alphabetic order based on the name of the department). Courses offered abroad were excluded, but two Biology courses partially taught in the Marine Lab were kept. There are in total 247 unique courses across the nine departments, and the specific breakdown is shown in the table below.

To ensure the abundance and representativeness of our sample, we decided on a sample size of 100 courses. Sample size (n_h) for each of the stratum was calculated so that it is proportional to its stratum size (n) in relation to the total population size (N) $\frac{n_h}{n} = \frac{N_h}{N}$. Specific sample sizes, shown in the table below, were used to randomly sample course numbers without replacement from all departments. The number of required books and the total cost of buying these required books new at the Duke store were then collected manually from the Duke Textbook Store website for each of these selected courses. Textbooks that were shown as "currently unavailable" online were included in the number of books counted for the course but were considered as having no price. These information can be found in the accompanying data file "SampleData_SH&EY.csv".

Table 1: Number of Natural Sciences Courses by Department

Stratum	StratumSize	SampleSize
Biology	38	15
Chemistry	23	9
Computer Science	30	12
Evolutionary Anthropology	12	5
Mathematics	50	20
Neuroscience	22	9
Physics	22	9
Psychology	31	13

Stratum	StratumSize	SampleSize
Statistical Science	19	8
Total	247	100

Data Analysis and Results

Do we want to add anything about exploratory data analysis?

Upon completion of sample data collection, the dataset was loaded and prepared for the R `survey` package, which was used to make finite population inference for this study. The employment of stratification and proportional allocation allows us to calculate the survey weights for each individual course, by dividing total course counts for each department by its respective sample size. A survey design can thus be created with these survey weights along with departments as strata, and course counts for each department as the finite population correction parameter.

Based on our study design, the total cost of required books for Natural Sciences at Duke is estimated to be \$12482 (SE = 1736), with a 95% confidence interval of (9079,21, 15884.11). The average cost of required books per course for Natural Sciences is estimated to be \$50.53 (SE = 7.03), with a 95% confidence interval of (36.76, 64.31). The average number of required books per course for Natural Sciences is estimated to be 0.458 (SE = 0.0572), with a 95% confidence interval of (0.346, 0.570). It is understandable that not every course requires the purchase of a new textbook.

Discussion

Although the use of stratification and proportional alloscation is intuitive and appropriate for the similarities shared between courses listed under the same department, there are various limitations to the study design. First of all, numbers of textbook required and prices may vary based on the level and difficulty of the courses within the same department, and this is not accounted for by randomly sampling courses within a department. Secondly, the courses included in the sample may not be representative enough for students studying Natural Sciences at Duke, since some courses, such as entry level courses, are more likely to be taken by large groups of students. It may be beneficial to assign higher weights to these classes than classes that fewer students may choose. We also noticed that most of the courses sampled do not have any required textbooks. This could be reflective of the reality, but could also be caused by the outdated or limited information on the Duke Textbook Store website. Textbooks that are required by professors but not entered into the system are thus neglected in our analysis. Those who do have count and cost data on the website tend to have higher prices, which could a potential source of bias for our inference. It is also uncertain that, for crosslisted courses, whether or not required textbooks will be listed only under a specific department, or whether graduate sections may need more textbooks than the undergraduate sections.

Code

```
# data collection
bio = c(89,154,157,201,202,203,205,207,209,212,221,223,250,255,304,321,329,347,348,364,365,369,415,420,4
chem = c(89,101,201,202,210,295,301,302,311,401,410,420,493,494,496,518,533,535,536,538,590,611,630)
compsci = c(94,101,116,201,216,230,249,250,270,290,307,308,310,316,330,334,342,350,356,520,524,527,534,5
evanth = c(101,220,231,260,285,330,333,344,363,520,530,561)
math = c(75,89,105,106,111,112,181,190,202,212,216,218,221,222,230,238,240,290,342,353,356,361,375,390,4
physics = c(133,139,141,142,151,153,161,162,175,264,271,305,346,361,363,465,505,549,566,567,590.01,590.0
psych = c(89.01,89.02,101,102,103,104,105,201,202,203,212,213,222,230,236,240,250,256,321,323,324,368,39
neuro = c(101,103,150,212,252,267,280,282,290,301,333,352,353,355,360,366,376,378,499,500,515,595)
stats = c(101,102,111,130,199,210,322,323,360,450,470,502,531,532,581,602,613,650,663)

depNames= c("bio", "chem", "compsci", "evanth", "math", "physics", "psych", "neuro", "stats")
```

```r
numDep = length(depNames)

allCourses = list() # list of lists to store all the courses
for (i in 1:numDep) {
  allCourses[[length(allCourses) + 1]] = get(depNames[i])
}

N = 0 # total number of courses to be computed

for (i in 1:numDep) { # 9 strata
  assign(paste("N", i, sep = ""), length(allCourses[[i]]))
  N = N + length(allCourses[[i]])
}
#N
# 247 courses in total

set.seed(322)
n = 100 # set sample size to be 100
department = c() # array to store department of all the sampled courses
course = c() # array to store course numbers of all the sampled courses

# sample courses from each department
for (i in 1:numDep) {
  assign(paste("n", i, sep = ""),
         round(get(paste("N", i, sep = ""))*n/N, 0))
  # make sample size within each stratum proportional to stratum size
  sample = sample(allCourses[[i]], get(paste("n", i, sep = "")), replace = FALSE)
  assign(paste("s", depNames[i], sep = ""), sample)
  department = c(department, rep(depNames[i], get(paste("n", i, sep = ""))))
  course = c(course, sample)
}

dat = data.frame(cbind(department, course))
dat$booknum = rep(0, 100)
dat$cost = rep(0,100)
# generate sample data file
#write.csv(dat, file = "sampledata.csv")
```

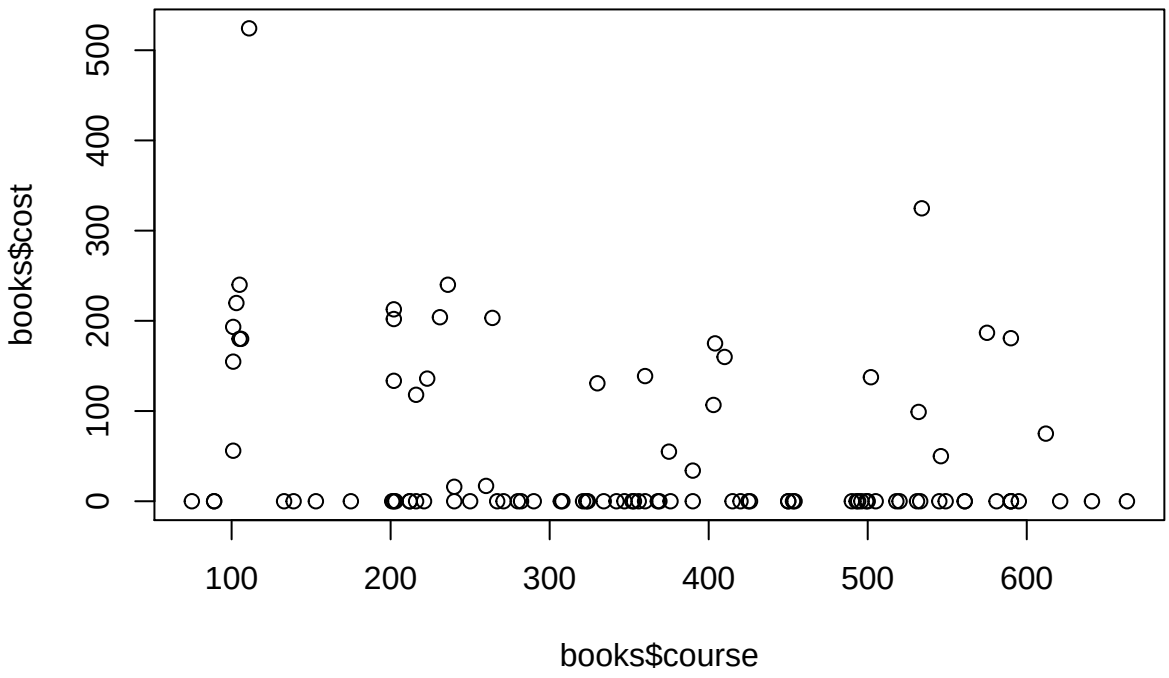
```r
# sample data is collected and entered manually
books = read.csv("SampleData_SH&EY.csv")

# exploratory data analysis
summary(books$cost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   50.84   80.96  524.25
```

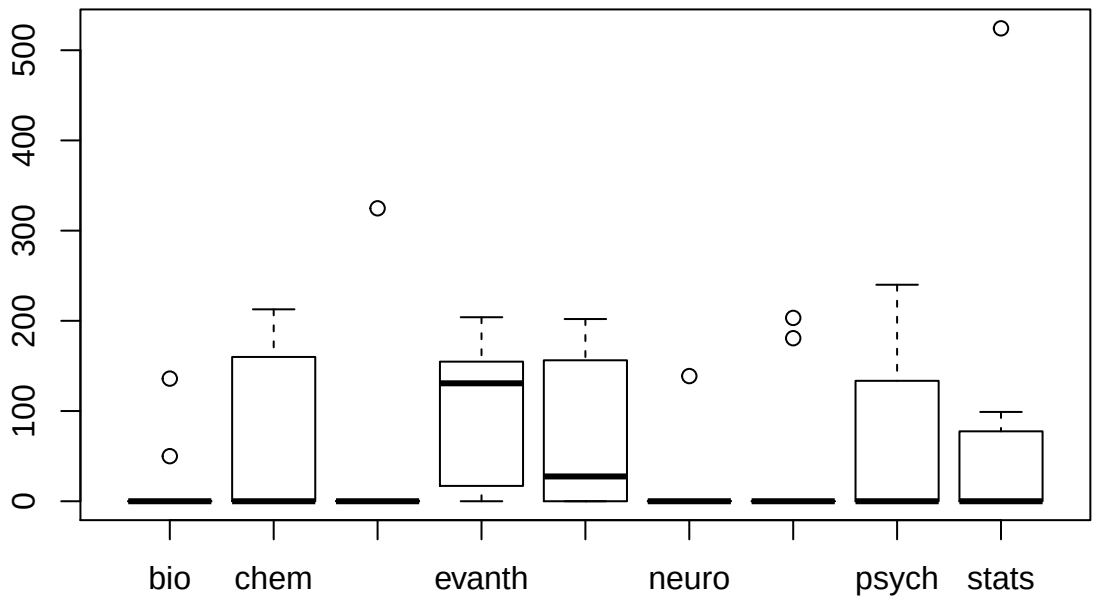
```r
summary(books$booknum)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.46    1.00    4.00
```

```r
plot(books$course, books$cost)
```




```
plot(books$department, books$cost)
```




```
library(survey)

#set weights to Nh/nh
books$weights = c(rep(38/15,15), rep(23/9,9), rep(30/12,12),
                  rep(12/5,5), rep(50/20,20), rep(22/9,9),
                  rep(31/13,13), rep(22/9,9), rep(19/8,8))

#set fpc to Nh
books$fpc = c(rep(38,15), rep(23,9), rep(30,12), rep(12,5),
              rep(50,20), rep(22,9), rep(31,13), rep(22,9), rep(19,8))

#create survey design
desBooks = svydesign(~1, strata = ~department, weights = ~weights, fpc = ~fpc, data = books)

#total cost of required books
svytotal(~cost, desBooks)
```

```
##      total   SE
## cost 12482 1736
```

```
confint(svytotal(~cost, desBooks), level = .95)
```

```
##         2.5 %   97.5 %
## cost 9079.206 15884.11
```

```
#average number of required books
svymean(~booknum, desBooks)
```

```
##            mean     SE
## booknum 0.45804 0.0573
```

```
confint(svymean(~booknum, desBooks), level = .95)
```

```
##             2.5 %    97.5 %
## booknum 0.3457552 0.5703215
```

```
#average cost per course
svymean(~cost, desBooks)
```

```
##        mean     SE
## cost 50.533 7.0282
```

```
confint(svymean(~cost, desBooks), level = .95)
```

```
##         2.5 %   97.5 %
## cost 36.75792 64.30814
```