Duke Textbook Project Report

Mingxuan Yang, Jiawei Chen

03/03/2020

Introduction

In this project, we designed and implemented a survey to explore the expected expenditures on required textbooks for natural science students at Duke, if they buy them new at the Duke textbook store.

The information we could easily acquire from the website is:
- the total number of courses for each natural science department.

The information that takes considerable resources (time) to collect, and thus needs to be sampled, is:
- the number of textbooks required for every single course within natural science departments;
- the costs of textbooks as new purchases for each course within natural science departments.

Bearing this in mind, we aim to answer the following questions:
- What is the total cost of required books in the natural sciences at Duke?
- What is the average number of required books per course in the natural sciences at Duke?
- What is the average of the (total) cost of required books per course in the natural sciences at Duke?

Note that for this project we only included courses numbered 699 and lower for each department, excluding any independent study courses. If a course has multiple lecture sections, only the first one that appears on the textbook store website counts.

The rest of the report is organized in the following way: the *Survey Design* section offers a detailed explanation of our survey design, defines some of terms we used when sampling and the data collection process; the *Results and Discussion* section presents answers for the three questions above, quantifies the accuracy of our estimates and comments on those results; the *Appendix* section displays codes used for this project, which are written in *R* 3.6.1.

Survey Design

Since our population can be naturally divided into 9 departments, we decided to use a stratified sampling method with 9 strata: department Biology, Chemistry, Computer Science, Evolutionary Anthropology, Mathematics, Physics, Psychology, Neuroscience and Statistical Science.

One big advantage of such stratified sampling design is it offering a comprehensive view of our population and thus tend to yield estimates with high accuracy. Also, since our data collection method is merely clicking on the bookstrore website, incorporating further clustering sampling design does not bring much convenience for collecting data but has the potential of driving variance up. Thus, we stick to the stratified sampling with proportional allocation.

In the following paragraph, we define some terms for our design.

Definitions

$N:$ the total population size, i.e. the number of courses across natural science departments;

$n:$ the sample size, i.e. the number of courses we sampled across natural science departments. We chose n = 60;

$h:\ h = 1, ..., 9,$ the index of stratum;

$i:\ i = 1, ...N,$ the index of courses;

$N_h:$ the population size of stratum h;

$n_h:$ the sample size taken from stratum h;

number_of_textbook$_i:$ the number of textbooks required for course i;

cost_of_textbook$_i:$ the costs of textbooks required for course i;

Data collection

In this part we briefly describe how we collected our data. We decided to sample 60 courses from the frame and used a stratified sampling with proportional allocation, i.e. we ensured that $\frac{n_h}{N_h} \approx \frac{n}{N}$ for each stratum h, which resulted in picking $10, 4, 6, 3, 11, 5, 11, 5,$ and 5 courses in those 9 departments respectively. This process was achieved with the help of `sample` function in *R*.

Having settled the course indices we decided to sample, we stored them, together with weights and Finite Population Correction Factor, in `textbook.csv` file and went to the bookstore website to obtain relevant information (number of books and costs). With the collected information about book quantities and prices, we updated the dataset into `textbook_new.csv`.

Table 1 below shows the specific population and sample sizes for each stratum, as well as the total sizes.

Table 1: Sample

	N_h	n_h
Biology	47	10
Chemistry	20	4
Computer Science	29	6
Evolutionary Anthropology	15	3
Mathematics	52	11
Physics	27	5
Psychology	55	11
Neuroscience	26	5
Statistical Science	24	5
Total	295	60

Results & Discussion

In this section, we are going to answer the three questions and make some comments. Note that these results are calculated using the stratified sampling design described above with R package `survey`, version 3.37. The dataset used is `textbook_new.csv`, and the random seed for sampling course indices is 1.

a) What is the total cost of required books in the natural sciences at Duke?

Table 2: Estimated total costs of textbooks

	total cost	standard error	upr	lwr
cost_of_textbook	13145.93	2623.78	8003.42	18288.44

As we can see from the **Table 2** above, the total cost of textbooks in natural science departments is estimated as about 13145.93 dollars. While this seems to be a huge figure, we need to bear in mind that there are about 300 courses in total across natural sciences.

The standard error is about 2623.78 dollars, which is a little bit big but still acceptable. One reason for this big figure could be the fact that most courses in natural science departments actually require no textbooks, but for those which require, the costs are usually high since books are indeed expensive in the US. This large variation may therefore lead to a high SE.

The resulting 95% confidence interval is (8003.42, 18288.44), which is wide because of the considerable SE.

b) What is the average number of required books per course in the natural sciences at Duke?

Table 3: Estimated avg. number of required books

	avg. no. of books	standard error	upr	lwr
number_of_textbook	0.34	0.06	0.24	0.45

As is displayed in **Table 3**, the average number of textbooks required per course in natural sciences is estimated as about 0.34. This has proved our statement that most courses at Duke University don't require any textbooks. This phenomenon may result from Duke's Green Classroom Certification. For the faculty members, impact on the environment needs to be considered during course delivery. Therefore, more online reading and electronic resources are used, which greatly reduces the average number of required textbooks per course.

The standard error is about 0.06, and the corresponding 95% confidence interval is (0.24, 0.45).

c) What is the average of the (total) cost of required books per course in the natural sciences at Duke?

Table 4: Estimated avg. costs of books

	avg. costs of books	standard error	upr	lwr
cost_of_textbook	44.56	8.89	27.13	61.99

Table 4 displays the estimated average costs of textbooks required per course in natural sciences, which is about 44.56 dollars. It indicates that a Duke student who takes courses in natural science departments should expect to spend 44.56 dollars on required textbooks if the student buys them new at the Duke textbook store. This seems to be a relatively low amount. According to Greenfield Community College in Massachusetts, "the national average of course material cost is 153 dollars per course", which is more than three times the cost from our survey. Based on these survey results, online material has greatly lightened the financial burden of Duke students.

The standard error is about 8.89, which is acceptable. The corresponding 95% confidence interval for estimated average costs of textbooks is thus (27.13, 61.99).

Appendix

Code for sampling

```
library(survey)
library(tidyverse)
library(knitr)

biology <- c(154,157,190,201,202,203,205,207,209,212,215,221,223,250,255,273,278,293,304,
319,326,329,335,347,348,364,365,368,369,375,391,415,420,425,427,432,450,452,453,454,490,
491,493,495,546,556,557,566,571,665,668) # copied & pasted from bookstore website; same
for below

biology_independent <- c(293,391,491,493) # acquired from duke public page course introduction;
same for below

chem <- c(101,201,202,210,295,301,302,311,393,394,401,410,420,493,494,496,511,531,533,
535,536,590,611,630)

chem_independent <- c(393,394,493,494)

compsci <- c(94,101,102,201,216,223,230,249,250,290,307,308,316,323,330,342,350,356,370,
391,394,512,520,524,527,553,561,571,590,630,650)

compsci_independent <- c(391,394)

evanth <- c(101,220,230,231,260,285,330,333,344,359,363,391,393,561,570,580,585)

evanth_independent <- c(391,393)

math <- c(105,106,111,112,181,191,202,212,216,218,221,222,228,230,238,240,260,304,340,
342,353,356,361,375,391,392,393,394,401,403,411,421,431,451,453,477,491,492,493,494,502,
531,532,545,557,563,565,575,577,582,'590-02',602,605,612,621,627,633,641,653, '690-05','690-70')

math_independent <- c(191,391,392,393,394,491,492,493,494)

physics <- c(133,137,139,141,142,151,153,161,162,175,264,271,305,346,361,363,364,365,
415,465,491,493,495,505,513,549,566,567,590,621)

physics_independent <- c(491,493,495)

psy <- c(101,102,103,104,105,106,141,201,202,203,208,212,213,221,222,230,236,240,250,
252,255,256,257,274,276,278,282,290,305,309,313,318,321,334,335,340,353,355,368,376,
392,394,425,426,435,436,490,492,494,496,499,500,561,603,610,611,671,686,690)

psy_independent <- c(392,394,492,494)

neurosci <- c(101,104,150,211,212,223,241,260,267,282,289,301,340,353,355,360,366,376,
391,392,427,493,494,495,496,499,500,503,504,567,595,686)

neurosci_independent <- c(391,392,493,494,495,496)

sta <- c(101,102,111,199,210,230,231,240,250,291,322,323,360,391,393,470,493,522,532,
561,581,602,612,613,641,642,663,693)

sta_independent <- c(391,393,493,693)

lst1 <- list(biology, chem, compsci, evanth, math, physics, psy, neurosci, sta) # list of
raw data from online - need to remove independent study courses

lst_ind <- list(biology_independent, chem_independent, compsci_independent, evanth_independent,
math_independent, physics_independent, psy_independent, neurosci_independent, sta_independent)
# list of independent study courses
```

```
N_h <- unlist(map2(lst1, lst_ind, ~length(.x) - length(.y))) # vector of N_h for each
stratum h

N <- sum(N_h) # total eligible no. of records as the sampling frame

n <- 60 # chosen sample size

n_h <- unlist(map(N_h, ~round(.x/N*n))) # vector of n_h for each stratum h

lst_N <- map2(lst1, lst_ind, ~.x[-which(.x %in% .y)]) # sampling frame

set.seed(1)

course <- map2(lst_N, n_h, ~sample(.x, .y)) # sampled courses

dep <- unlist(map2(c('biology', 'chem', 'compsci', 'evanth', 'math', 'physics', 'psy',
'neurosci', 'sta'),n_h, ~rep(.x, .y)))

weight <- unlist(map2(N_h, n_h, ~rep(.x/.y, .y)))

fpc <- unlist(map2(N_h, n_h, ~rep(.x, .y)))

data <- data.frame(course = unlist(course), department = dep, weights = weight, fpc =
fpc)

write_csv(data, "textbook.csv")

textbook <- read.csv('textbook_new.csv', header = TRUE) # collected textbook information
from online into textbook.csv to form textbook_new.csv

textbook_survey <- svydesign(~1, strata = ~department, weights = ~weights, fpc =~ fpc,
data = textbook)

a = svytotal(~cost_of_textbook, textbook_survey)

ac = confint(svytotal(~cost_of_textbook, textbook_survey), level = 0.95)

b = svymean(~number_of_textbook, textbook_survey)

bc = confint(svymean(~number_of_textbook, textbook_survey), level = 0.95)

c = svymean(~cost_of_textbook, textbook_survey)

cc = confint(svymean(~cost_of_textbook, textbook_survey), level = 0.95)
```