STAT 260 - Short Report 3

Conner Taylor, Parker Johnson

3/3/2022

Introduction

Most courses at Carleton are given descriptions beyond their name in order to better clarify their content, suggested to be at most 100 words. However, we were curious as to if this limit was met on average. In this study, we estimated the average word count description for Carleton courses, and additionally, the proportion of courses in the Academic Catalog that were offered this school year.

Methodology

For this design, we applied a two-stage cluster sample sampling design. This design seemed appropriate for our study, because cluster sampling by the departments then sampling certain pages within departments was convenient and time-effective as opposed to searching for individual courses spread out across multiple sub-pages. Additionally, courses only appear in their respective department, so the secondary sampling units are non-overlapping.

The population of primary sampling units (clusters) was the total number of departments found in the 2021-22 Academic Catalog excluding any that simply linked to a different department, which was 43, and the secondary sampling units were the courses within each department. Our measurements were the word counts of course descriptions and whether the course was offered or not. Our sampling frame was all 43 departments.

In order to figure out how large each SRS should be– of the departments and of the courses within each department–, a pilot study of two departments with two courses each was conducted. We assigned each department a number based on the alphabetical list on the 2021-22 Academic Catalog, and then got two departments from a random sample: Asian Studies and Cognitive Science. We then counted the number of courses on the page by copying the list, pasting it into Google Docs, and then observing if all instances of the subject's four-letter name (e.g. ASSN for Asian Studies) were included just in the course name and not a course's description or requirements, and then got the count directly from Docs' "Find" command. The two courses sampled were done in a similar manner, with each course being assigned a number corresponding to the order of courses on the department's page, and then the sampled course's description was pasted into Docs and the words were counted via the "Word count" command. A binary yes-no variable was also noted for whether or not the course was offered this year or not. 400-level Integrative Exercises were not counted as courses, and the course name, professor, requirements, and the course being offered were not included as descriptions. Courses that were not 6 credit courses were still included in the study.

From the pilot sample, we figured that it took twice the time to count the number of words in the course description, about 2 minutes, than it was to count the number of courses per department, which took about 1 minute, at least with what we had sampled for our pilot study. In order to calculate the optimal course sample size, we had to assume that the departments all contained the same number of courses. With this assumption and costs, we ended up with an optimal course sample size of 1, and from a budget of spending a maximum of 90 minutes with no initial time cost, an optimal department sample size of 30. We then used the same process as in the pilot study above for the actual study.

Results

The results from each of the 30 sampled courses can be seen in Figure 1. The largest number of words in a course description was 117 words, which was for AFST 180, an Africana Studies class. The smallest number of words in a course description was 5 words, which was for BIOL 225, a Biology class. A table of the results can be found at the end of the appendix.

We found that the estimated word count for all course descriptions was 85.076 words with a standard error of 6.433 words. We are 95% confident that the true mean word count for all course descriptions at Carleton is between 71.917 and 98.234 words, and thus on average, course descriptions are under 100 words. The estimated proportion of courses offered in 2021-2022 was 0.614, with a standard error of 0.128. We are 95% confident that the true proportion of courses offered in 2021-2022 is between 0.352 and 0.876.

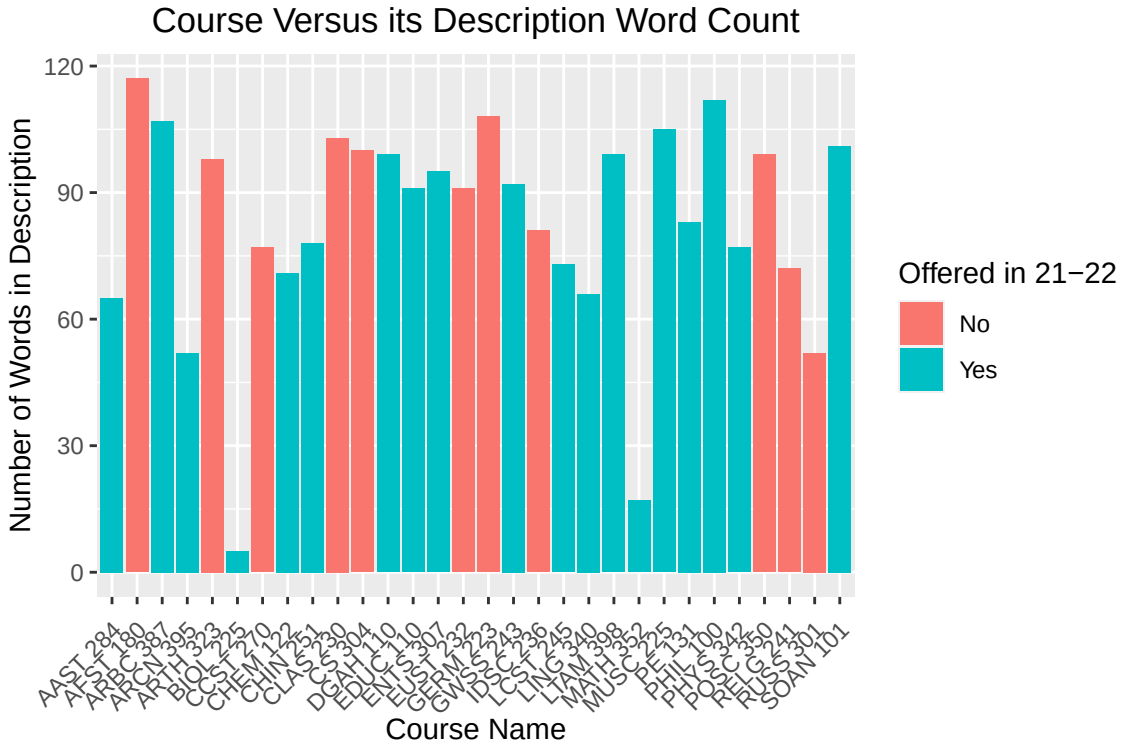

Figure 1: The sampled data from each of the 30 courses

Discussion

One issue was that we only sampled one course per department, partially due to our pilot sample having within-department course word counts very close to each other, and partially due to our pilot sample including departments whose list of courses were very simple to count with our process. This resulted in our c_1 value being less than our c_2 value, even though when we did the actual study this stopped being accurate. With the former, we ended up with an R_a^2 value very close to 1 from our pilot sample, suggesting a lot of within-cluster homogeneity within each department. This led not just to our optimal course sample size even being 1, but instead around .25 and having been rounded up to 1. Thus, it would have been worth trying maybe even just a slightly larger pilot study to have maybe caught one or both of these discrepancies and thus resulting in a value for m_opt greater than 1.

Alternatively, our value of 1 for the number of courses to sample within each department was fixed. This also decreased the precision of our study, because some departments such as Music contained well over 100 courses, yet only 1 Music course was sampled. Even with another study having a value of m_opt greater than

1, it would still result in departments with fewer courses having a greater effect on the estimate, as there is a greater weight towards them.

Even with the specifics of our methodology, there was also still a chance for measurement error. We could have had a point in which we didn't notice a four-letter name in a description and counted more courses than there were in a particular department, or failed to avoid copy-pasting something superfluous like an Integrative Exercise course or a prerequisite in a course's description. Additionally, we could have miscounted which course specifically was our sampled course, as that was only done via counting down the page.

Additionally, since our $DEff$ for both estimators was greater than 1, our precision was less than a standard SRS. While our SE was about one-fourth the range of our CI, our SE itself was rather large, leading to, especially for the proportion of courses offered, a very large and thus unhelpful CI. For the confidence interval, the skew was approximated to be very close to zero, and so the minimum sample size suggested by Sugden et. al was 28, which means the Central Limit Theorem and normality assumption applies with the chosen sample size of 30.

Technical Appendix

The first step in our sampling design is to find the cluster sample size m_i within each department. We did this using a pilot sample to calculate an estimated value for R_a^2. Using the average value of M from the pilot sample and estimated costs of $c_1 = 1$, and $c_2 = 2$, we used the optimal SSU sample size formula $m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$ to find the number of courses per department to sample. We then used this value for m_opt to find the optimal number of departments to sample using the formula $n_{opt} = \frac{C-c_0}{c_1+c_2 m_{opt}}$, where $C = 90$ was our timed budget in minutes, and c_0 was the initial cost, which was 0 minutes.

```
set.seed(2553517)
N <- 43 #total number of departments
pilot_samp <- sample(1:N, 2, replace = FALSE)
#generate two values for departments to use in a pilot sample
pilot_M6 <- sample(1:pilot_samp[1], 2, replace = FALSE)
pilot_M11 <- sample(1:pilot_samp[2], 2, replace = FALSE)
#generate two values each for courses to use in a pilot sample
coursedesc_pilot <- read_csv("coursedesc_pilot.csv")
```

```
## Rows: 4 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): deptname, coursename, offered2122
## dbl (4): deptnum, M_i, coursenum, wordcount
##
## 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.
```

```
c_1 <- 1
c_2 <- 2
#cost values in minutes based on pilot sample
pilot_lm <- lm(wordcount~deptname, data = coursedesc_pilot)
R2a_pilot <- summary(pilot_lm)$adj.r.squared
R2a_pilot
```

```
## [1] 0.8913969
```

```
#adjusted R-squared value from pilot sample
M <- (9+11)/2
#take average of M_i from pilot sample for M in m_opt calculation
m_opt <- sqrt((c_1*M*(N-1)*(1-R2a_pilot)) / (c_2*(N*M-1)*R2a_pilot))
```

```r
m_opt #calculate m_opt
```

```
## [1] 0.2442117
```

```r
m_opt <- ceiling(m_opt)
m_opt #round m_opt to nearest whole number, in this case up to 1
```

```
## [1] 1
```

```r
C <- 90 #max cost valued to be around 90 minutes
c_0 <- 0 #no initial time value
n_opt <- (C - c_0) / (c_1 + c_2*m_opt)
n_opt #calculate n_opt in terms of cost
```

```
## [1] 30
```

```r
samp <- sample(1:N, n_opt, replace = FALSE)
samp
```

```
##  [1] 32  3  9 25 35 23 40 34 13 19 10 36  2 27  6  5 18 14  8 12  4 39  7 29 33
## [26] 42 30 22 38 31
```

```r
#apparently I get a different list depending on which computer I use

coursedesc <- read_csv("coursedesc.csv")
```

```
## New names:
## * `` -> ...1
## Rows: 30 Columns: 8-- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): deptname
## dbl (4): ...1, deptnum, M_i, coursenum
## lgl (3): coursename, wordcount, offered2122
## 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.
```

```r
for (i in 1:nrow(coursedesc)) {
  coursedesc$coursenum[i] <-
    sample(1:coursedesc$M_i[i], 1)
}
#unlike above, this was the same for both my laptop and desktop

#The 30 samples and their course names, word count, and offered or not
coursedesc$coursename <- c("IDSC 236", "PHIL 100", "CCST 270", "EUST 232",
                           "CLAS 230", "LCST 245", "AAST 284", "CHIN 251",
                           "ENTS 307", "DGAH 110", "POSC 350", "CHEM 122",
                           "CS 304", "MATH 352", "LING 340", "ARTH 323",
                           "SOAN 101", "BIOL 225", "PHYS 342", "GWSS 243",
                           "RELG 241", "ARCN 395", "GERM 223", "LTAM 398",
                           "MUSC 225", "EDUC 110", "PE 131", "AFST 180",
                           "RUSS 301", "ARBC 387")
coursedesc$wordcount <- c(81, 112, 77, 91, 103, 73, 65, 78, 95, 99, 99, 71,
                          100, 17, 66, 98, 101, 5, 77, 92, 72, 52, 108, 99, 105,
                          91, 83, 117, 52, 107)
coursedesc$offered2122 <- c("No", "Yes", "No", "No", "No", "Yes", "Yes", "Yes",
                            "Yes", "Yes", "No", "Yes", "No", "Yes", "Yes", "No",
                            "Yes", "Yes", "Yes", "Yes", "No", "Yes", "No",
```

```
                                "Yes", "Yes", "Yes", "Yes", "No", "No", "Yes")
```

We then checked to make sure that normality assumptions were met with this sample size. Since our sample size was greater than the calculated recommended sample size, the normality assumption was presumed to have been met.

```
coursedesc$skew1 <- (coursedesc$wordcount - mean(coursedesc$wordcount))^3
#calculate the summand for the skew calculation
skew <- (sum(coursedesc$skew1))/(N*sd(coursedesc$wordcount))^3 #approximate skew
28+25*(skew)^2
```

```
## [1] 28.00001
```

```
#calculate min sample size from Sugden et. al equation from p. 49 of text
```

We then used the survey package to calculate our estimated word count mean and proportion, and their corresponding standard errors and confidence intervals.

```
coursedesc$m_i <- 1
coursedesc$wts <- coursedesc$M_i / coursedesc$m_i
course_design <- svydesign(id = ~deptname + coursename,
  weights = ~wts,
  data = coursedesc) #Creating the survey design object

#Estimated mean and confidence intervals for the course description word count
svymean(~wordcount, course_design, deff=T)
```

```
##              mean     SE   DEff
## wordcount 85.0756 6.4333 1.6802
```

```
confint(svymean(~wordcount, course_design, deff=T), df=degf(course_design))
```

```
##              2.5 %  97.5 %
## wordcount 71.91796 98.2332
```

```
#Estimated mean and confidence intervals for the proportion of courses offered
#in 2021-2022
svymean(~offered2122 == "Yes", course_design, deff=T)
```

```
##                              mean     SE   DEff
## offered2122 == "Yes"FALSE 0.38621 0.12794 2.0535
## offered2122 == "Yes"TRUE  0.61379 0.12794 2.0535
```

```
confint(svymean(~offered2122 == "Yes", course_design, deff=T), df=degf(course_design))
```

```
##                               2.5 %    97.5 %
## offered2122 == "Yes"FALSE 0.1245538 0.6478715
## offered2122 == "Yes"TRUE  0.3521285 0.8754462
```

Here are the data tables for our pilot sample and sampled data:

Table 1: The dataset for the pilot study.

deptnum	deptname	M_i	coursenum	coursename	wordcount	offered2122
6	Asian Studies	9	1	ASST 100	92	Y
6	Asian Studies	9	2	ASST 130	105	Y
11	Cognitive Science	11	10	CGSC 394	66	N
11	Cognitive Science	11	4	CGSC 232	65	Y

Table 2: The full dataset for the sampled data, excluding skew and sample design variables.

...1	deptnum	deptname	M_i	coursenum	coursename	wordcount	offered2122
1	25	Interdisciplinary Studies	24	16	IDSC 236	81	No
2	34	Philosophy	46	1	PHIL 100	112	Yes
3	13	Cross-Cultural Studies	7	5	CCST 270	77	No
4	19	European Studies	11	7	EUST 232	91	No
5	10	Classics	38	14	CLAS 230	103	No
6	27	Literary and Cultural Studies	5	5	LCST 245	73	Yes
7	6	Asian Studies	9	6	AAST 284	65	Yes
8	5	Asian Lang and Lit	48	13	CHIN 251	78	Yes
9	18	Environmental Studies	16	11	ENTS 307	95	Yes
10	14	Digital Arts and Humanities	2	1	DGAH 110	99	Yes
11	37	Political Science	133	118	POSC 350	99	No
12	8	Chemistry	33	2	CHEM 122	71	Yes
13	12	Computer Science	38	16	CS 304	100	No
14	29	Mathematics and Statistics	40	27	MATH 352	17	Yes
15	26	Linguistics	22	20	LING 340	66	Yes
16	4	Art & Art History	79	38	ARTH 323	98	No
17	41	Sociology and Anthropology	46	1	SOAN 101	101	Yes
18	7	Biology	62	10	BIOL 225	5	Yes
19	36	Physics and Astronomy	39	26	PHYS 342	77	Yes
20	21	GWSS	15	5	GWSS 243	92	Yes
21	39	Religion	72	38	RELG 241	72	No
22	3	Archaeology	6	6	ARCN 395	52	Yes
23	23	German	22	12	GERM 223	108	No
24	28	Latin American Studies	3	3	LTAM 398	99	Yes
25	32	Music	196	112	MUSC 225	105	Yes
26	16	Educational Studies	13	2	EDUC 110	91	Yes
27	35	PEAR	111	27	PE 131	83	Yes
28	1	Africana Studies	12	7	AFST 180	117	No
29	40	Russian	29	21	RUSS 301	52	No
30	31	Middle Eastern Languages	27	20	ARBC 387	107	Yes