How Many Words do College Kids Know?

Elliot Cahn and Colin Pi

May 13, 2019

1. Introduction

The current education system attaches a great importance to vocabulary. For example, before College Board redesigned SAT verbal section in 2016, approximately a quarter of available point corresponded to the sentence completion section in which test takers had to choose the best word or words that completes the sentence. Likewise, a vast knowledge of words is considered as a major component for assessing one's proficiency in the English language. It also plays a pivotal role in college entrance. But how many words do college students actually know? Our report examines the vocabulary size of one college student, Elliot Cahn, by estimating how many words that he knows in Meriam Webster's Collegiate Dictionary, 10th edition.

2. Methodology

In this study, a "known" word was defined as a word that Elliot could meaningfully include in a sentence. In cases where a word had multiple definitions, the word itself was counted only once. If Elliot could meaningfully include the word in a sentence using at least one recognized definition, then the word counted as "known."

To determine how many words Elliot knows, we implemented one-stage cluster sampling with pages as the primary sampling unit and words as the secondary sampling unit. A simple random sample was implemented to determine the pages examined instead of stratified random sampling due to limited time scope and the lack of belief that stratification at the page level would significantly alter our results.

A ratio estimator was used to calculate the proportion of the words that Elliot knows in the dictionary. The total number of words in the dictionary (M_0) is unknown, and the number of words on each page is unequal. Therefore, we cannot use an unbiased estimator to infer the proportion. Also, ratio esimator may produce more precise results than an unbiased estimator since the quantity of words on a page likely corresponds to the number of words known on that page. For estimating the total number of the words, however, an unbiased estimator of the total was used since ratio estimators for small sample sizes may either be biased or not produce as precise results when compared to unbiased estimator.

The total number of pages of the dictionary is 1387, and we chose to sample 30 of the pages for our analysis. Even if it is considerably smaller than the population size, 30 is reasonably high number considering the time constraints and effort of counting all the words within the chosen pages of the dictionary.

3. Results

The summary statistics of our result are displayed in Table 1. The number of words within each page varies significantly, with range stretching 42 words and sample standard error of 10.779

words. This suggests that assuming equal cluster sizes is unreasonable to make in this analysis and indicates that ratio estimation is the most viable way to estimate the proportion.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Words	30	50.500	10.779	27	44	54.8	69
Known	30	19.867	7.427	5	14.5	24	36

Figure 1 illustrates the proportion of words Elliot knows on each page sampled. The illustration suggests that there are considerable differences in the proportions of known words among the pages. In particular, we do not observe any positive association between number of words in a page and number of words we know. The correlation coefficient (ρ) between these two variables is 0.09648897, and their uncorrelatedness is clearly visible in Figure 2 as well.

Figure 1: Number of Words in a Page and Words Elliot Knows

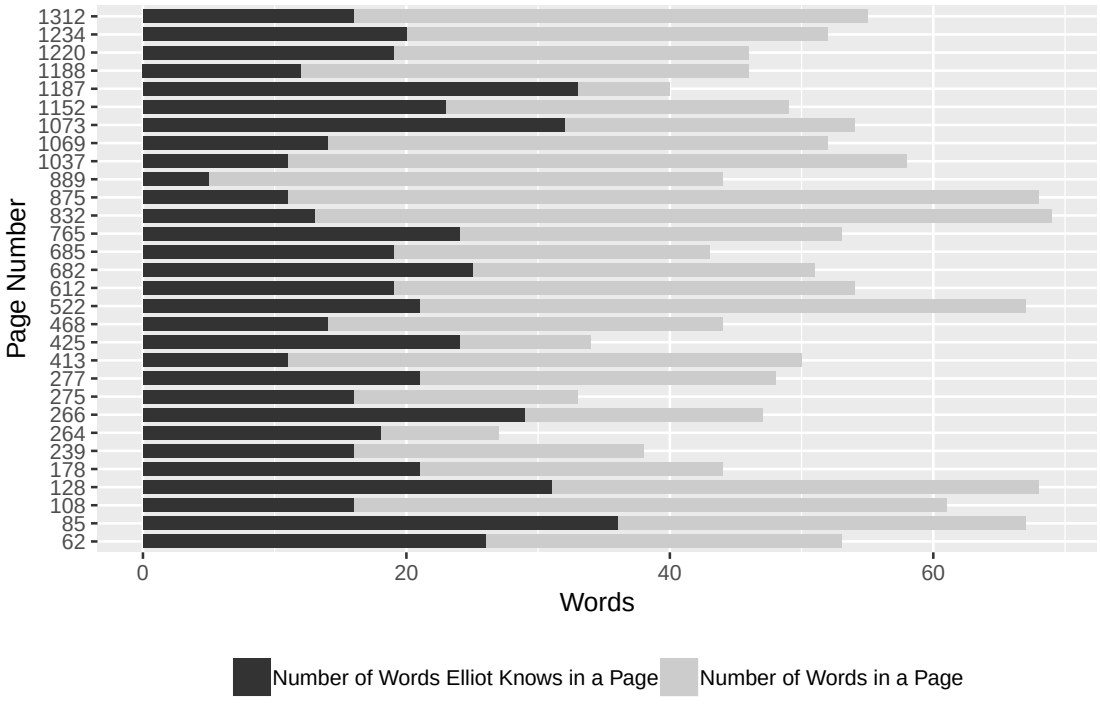
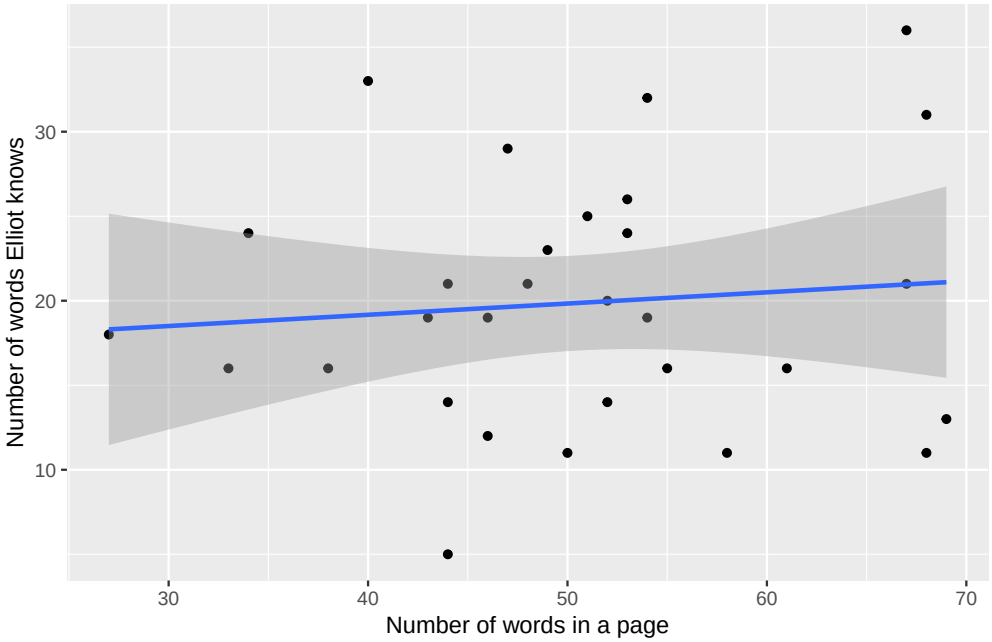


Figure 2: Number of Words in a Page v. Words Elliot Knows



We analyzed the sampled data using the `survey` package and obtained the estimate of 27,376 words for the total number of words Elliot knows in the dictionary, with a standard error of 1848 words. These results suggest with 95 percent confidence that the true total number of know words lies between 23,754 words and 30,998 words. The ratio estimation estimates Elliot knows 39.34 percent of words in the dictionary with a standard error of 2.93 percent. These results suggest with 95 percent confidence that the true proportion of words Elliot knows lies between 33.6 percent and 45.08 percent.

4. Conclusion

Our study estimates that the total number of words Elliot, a college student, knows in Meriam Webster's Collegiate Dictionary, 10th edition, is 27,376 words, which is about 39.34 percent of the words listed in this dictionary. Even accounting for the sampling errors, this finding suggests that his vocabularly covers between 30 to 40 percent of the words in the dictionary.

One interesting observation is the void of a positive correlation between the number of words in each page and the number of words known by Elliot on that page. Even though the prime reason for implementing ratio estimation was accounting for unknown population size (M_0), the absence of any association between these two variables suggests ratio estimation might not produce more precise result than unbiased estimation but simply introduce more bias to the result. This also corroborates that using an unbiased rather than ratio estimation was more appropriate for the analysis.

This study could be improved if information about the number of words in the dictionary were made available. This would have allowed the use of unbiased estimation for proportion and avoid the bias introduced by ratio estimation. Moreover, more precision could have been achieved through stratified random sampling if meaningful classifiers were made available, such as some measurement of difficulty for each word.

5. Appendix

Picking clusters (pages) to sample

```
> N <- 1378
> n <- 30
>
> set.seed(70)
>
> ## Sample 30 pages using SRS
> page <- sample(1:N, n, replace = F)
> knitr::kable(t(page[1:15]))
```

85	1312	875	178	1188	1187	425	682	239	413	612	1234	832	1073	468

```
> knitr::kable(t(page[16:30]))
```

1152	62	264	889	522	275	108	1069	1037	765	277	1220	128	685	266

Dataset

```
> dictionary <- read.csv("Report 2.csv")
> kable(dictionary %>% dplyr::select(1:3))
```

Page	Words	Known
62	53	26
85	67	36
108	61	16
128	68	31
178	44	21
239	38	16
264	27	18
266	47	29
275	33	16
277	48	21
413	50	11
425	34	24
468	44	14
522	67	21
612	54	19
682	51	25
685	43	19
765	53	24
832	69	13
875	68	11

Page	Words	Known
889	44	5
1037	58	11
1069	52	14
1073	54	32
1152	49	23
1187	40	33
1188	46	12
1220	46	19
1234	52	20
1312	55	16

Survey Design

```
> dictionary$N <- N
> dictionary$n <- n
> dictionary$wts <- dictionary$N/dictionary$n
>
> dictionary.des <- svydesign(id = ~1, fpc = ~N, weight = ~wts,
+     data = dictionary)
>
> ## total (unbiased) estimate
> svytotal(~Known, dictionary.des)
      total   SE
Known 27376 1848
> confint(svytotal(~Known, dictionary.des))
        2.5 %   97.5 %
Known 23754.29 30998.24
>
> ## proportion/ratio (biased) estimate
> svyratio(~Known, ~Words, dictionary.des)
Ratio estimator: svyratio.survey.design2(~Known, ~Words, dictionary.des)
Ratios=
          Words
Known 0.3933993
SEs=
           Words
Known 0.02928164
> confint(svyratio(~Known, ~Words, dictionary.des))
                2.5 %    97.5 %
Known/Words 0.3360084 0.4507903
>
> ## correlation coefficient
> cor(dictionary$Words, dictionary$Known)
[1] 0.09648897
```