Report 2- How Many Words Starting with h Does Helen Know

Sammi Sheridan and Helen Moses

2024-02-03

Introduction

Many people wonder if they have a large vocabulary, but typically lack the tools to find out. This question is more difficult to answer than those where Simple Random Sample estimation applies due to the fact that this question has two variability components. To estimate the proportion of words an individual knows, it is necessary to estimate both how many words the individual knows and how many words exist. The estimation tool that allows one to work with two variability components is called Ratio Estimation. In this report, we first utilize Simple Random Sample Estimation to estimate the total number of words starting with an h Helen knows. Second, we utilize Ratio Estimation to estimate the proportion of words starting with an h that Helen knows.

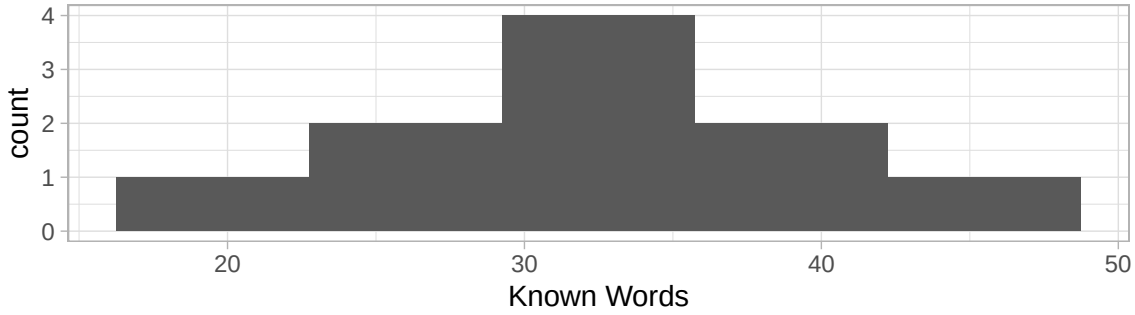
Methodology

To conduct this analysis, we utilized Merriam Webster's Collegiate Dictionary, Tenth edition. This dictionary has N = 52 pages that contain words starting with an h (pages 520-571). Our population and sampling units are pages from the dictionary that contains h words. From our sampling frame of 52 pages, we randomly selected n = 10 pages for our sample. We elected to have a sample size of 10 because it takes a long time to collect the data from each page and the time it takes to sample 10 pages was the most amount of time we were willing to spend on data collection. With more available time, we could have afforded a larger sample size. On each selected page we counted the total number of words and the number of words that Helen knew. The words included in the word counts were those on their own line that were not tabbed. We created a data frame where one variable was the number of words Helen knows on a page and a second variable was the total number of words on the same page. With this data frame, we utilized the survey package to obtain an SRS estimate of the total number of words starting with an h that Helen knows. Additionally, we utilized the survey package to obtain a ratio estimate of the proportion of words starting with an h that Helen knows contained in the Merriam Webster's Collegiate Dictionary, Tenth edition.

Results

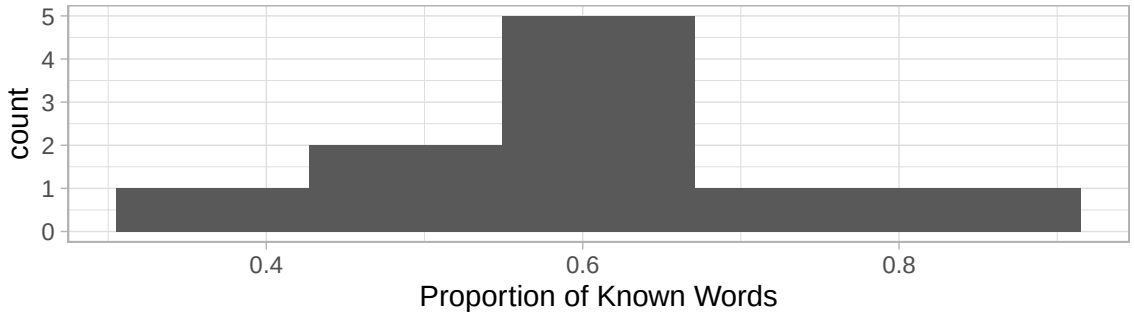
Figure 1 displays the results from our sample for the number of words that Helen knows starting with the letter h per page.

Figure 1: Sample Results of Number of Known Words per Page



The data for the number of words Helen knows starting with the letter h seems to follow a normal distribution with mean 31.20 and standard deviation 7.51. Figure 2 displays the results from our sample for the proportion of words that Helen knows starting with the letter h.

Figure 2: Sample Results of Proportion of Words Known



The data for the proportion of words Helen knows starting with the letter h seems to follow a normal distribution with mean .61 and standard deviation .14. The correlation between the sample results of total words and words known is .738.

By using a SRS sampling design, the estimated total number of words Helen knows starting with the letter h is 1622 with a standard error of 110.99. We are 95% confident that Helen knows between 1371 and 1873 words starting with the letter h. By using a ratio estimation, the estimated percentage of words Helen knows starting with the letter h is 59.77% with standard error .03. We are 95% confident that Helen knows between 52.94% and 66.60% of words starting with the letter h.

Discussion

Our best guess at the total number of words Helen knows starting with the letter h is 1622, and our best guess at the proportion of words Helen knows starting with the letter h is 59.77%. There may be non-sampling errors in our estimates because it may be possible that we did not accurately count the total and known number of words on each page. We decided to count each word that did not have an indent, but we could have excluded words that do start with the letter h in the dictionary. In addition, when Helen was deciding whether she knew a word or not, she may have inaccurately represented her knowledge because she read the definition of the word before deciding if she knew it or not. This may have impacted our measurement because Helen may not have necessarily known the word had she not read the definition first. Ratio estimators are biased, but because our sampling fraction was large (.19) and the correlation was close to 1 (.738), our proportion estimate will have small bias. Still, because our sample size was n=10 pages, our assumption for large sample size (n=30) was not met when estimating the total number of words Helen knows starting with h, meaning our estimations and confidence intervals may be inaccurate for the total estimate.

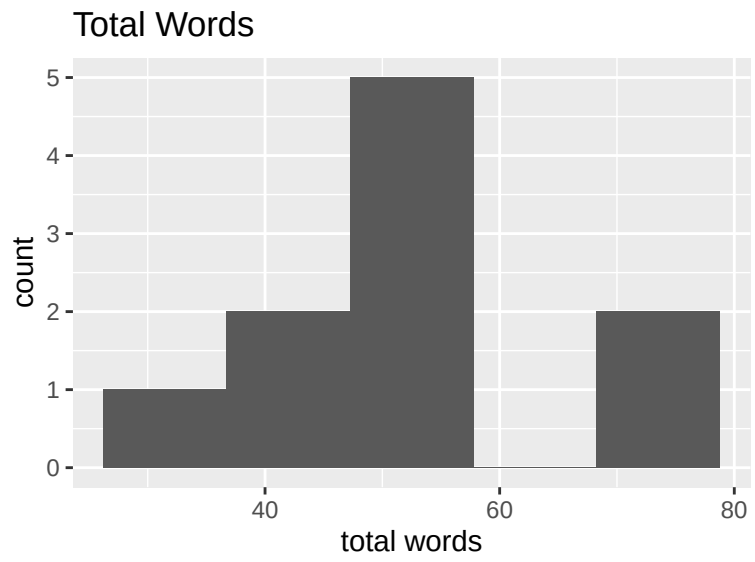
Computational R Appendix

```
set.seed(13)
#Find random sample of 10 pages in the H section
sample(520:571,10,replace=FALSE)
```

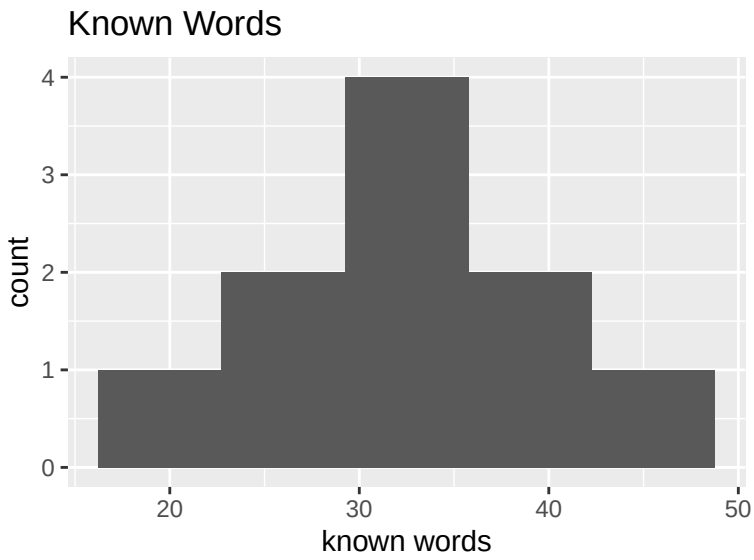
```
##  [1] 543 522 556 529 532 525 541 523 548 536
```

```
#Find total words per page and known words per page
pagenumber<-c(543, 522, 556, 529 ,532, 525, 541, 523, 548, 536)
totalwords<-c(57,56,28,70,42,39,55,69,50,56)
knownwords<-c(32,37,25,38,17,30,25,43,31,34)

data<-as.data.frame(cbind(pagenumber,totalwords,knownwords))
```

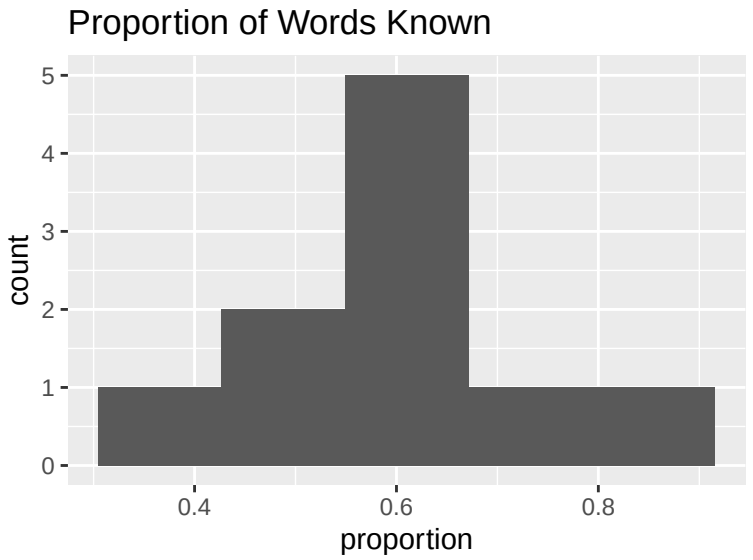
```
#EDA

#Histograms
#Total words
ggplot(data)+geom_histogram(aes(x=totalwords),bins=5) +
  labs(title="Total Words",x='total words')
```




```
#Known Words
ggplot(data)+geom_histogram(aes(x=knownwords),bins=5) +
  labs(title="Known Words",x='known words')
```




```
#Proportion of known words divided by total words
ggplot(data)+geom_histogram(aes(x=knownwords/totalwords),bins=5) +
  labs(title="Proportion of Words Known", x='proportion')
```




```
#Mean and sd of words known
mean(data$knownwords)
```

```
## [1] 31.2
```

```
sd(data$knownwords)
```

```
## [1] 7.509993
```

```
#Mean and sd of proportion
mean(data$knownwords/data$totalwords)
```

```
## [1] 0.6136701
```

```r
sd(data$knownwords/data$totalwords)
```

```
## [1] 0.1419078
```

```r
#Correlation
cor(data$knownwords, data$totalwords)
```

```
## [1] 0.737701
```

```r
# Creating our SRS design
data$n <- nrow(data)
data$N <- 52
data$wts <- data$N/data$n
design_srs <- svydesign(id= ~1, fpc= ~N, weights= ~wts, data= data)

# Estimating the Percentage of words usin ratio estimate
ratio<-svyratio(~knownwords, ~totalwords, design_srs)
ratio
```

```
## Ratio estimator: svyratio.survey.design2(~knownwords, ~totalwords, design_srs)
## Ratios=
##            totalwords
## knownwords  0.5977011
## SEs=
##            totalwords
## knownwords 0.03019337
```

```r
confint(ratio,df=degf(design_srs))
```

```
##                           2.5 %    97.5 %
## knownwords/totalwords 0.529399 0.6660033
```

```r
# Estimating the Total Known words using SRS estimate
totalknown<-svytotal(~knownwords,design_srs)
totalknown
```

```
##             total     SE
## knownwords 1622.4 110.99
```

```r
confint(totalknown,df=degf(design_srs))
```

```
##               2.5 %   97.5 %
## knownwords 1371.334 1873.466
```

```r
# Estimating the Total h words using SRS estimate
totalwords<-svytotal(~totalwords,design_srs)
confint(totalwords,df=degf(design_srs))
```

```
##               2.5 %   97.5 %
## totalwords 2278.855 3149.945
```