The Change in the Number of Four-letter Words in the English Language

Rachel Kahlenberg
Ohio Northern University

Rebekkah Dann
Messiah College

SUMSRI 1999

The English language is constantly changing, but it is almost impossible to detect all of these changes without choosing a specific area of study. Because four-letter words are an integral and sometimes interesting part of the English language, it is worthwhile to contemplate whether their use has changed over the past few decades. However, the task of counting the number of four-letter words would be very time consuming, but through the use of statistical sampling the time this takes is considerably reduced.

The population density method is a statistical sampling strategy that includes applications that estimate the following: the number of diseased trees in a forest, the number of ant hills in a specific area, or the number of defects in a yard of material[1]. The number of four-letter words in the English language, which can be represented by the entries in a dictionary, can also be estimated using the population density method. This method estimates the total number in the population by using quadrat samples. A quadrat is a set piece of the total area where the population is found. The total area for the population would be the total number of pages in the dictionary, and the quadrat for a dictionary sampling could be set at one page. The number of elements from the population found in each quadrat is counted and then used to estimate the total number of elements in the population.

Comparing dictionaries[2] of two different years will yield data that can be used to note the changes in the number of four-letter words in the English language. The total area (A) for the 1950 and 1986 dictionaries is 2,987 and 2,662 pages, respectively. The size (a) of a quadrat is one page.

A random sample was generated using Minitab, a statistical software package, that produced 100 random page numbers from each dictionary. The number of four-letter words[3] on each page was then counted and recorded for later analysis. Table 1 summarizes the frequency of four-letter words per page. For example, in the 1950 dictionary there were 30 pages sampled which did not contain any four-letter words and 34 pages that each contained one four-letter word. Of the 100 random pages sampled for each dictionary, there were no more than 13 four-letter words found on any page.

---

[1] Scheaffer, p. 371-375.

[2] Different editions of the same dictionary were used: *Webster's New International Dictionary* (2nd ed.) 1950 and *Webster's Third New International Dictionary Unabridged* 1986

[3] Words are the bold-faced entries that are defined in a dictionary. The same word with different definitions was counted only once. Four-letter abbreviations and prefixes were included in the count. Hyphens and spaces were disregarded.

Table1

# 4 letter words	0	1	2	3	4	5	6	7	8	9	10	11	12	13
1950	30	34	12	10	5	4	1	3	0	0	1	0	0	0
1986	25	19	19	10	6	5	6	2	4	1	1	0	1	1

In order to estimate the total number of four-letter words in each dictionary ($\hat{M}$), it is necessary to find the estimated number of four-letter words per page ($\overline{m}$), and the estimated density of four-letter words per page ($\hat{\lambda}$).
The following equations were used to calculate these values where n is the number of pages sampled.

	1950	**1986**
$\overline{m} = \frac{1}{n}\Sigma m_i$	$\frac{165}{100} = 1.65$	$\frac{262}{100} = 2.62$
$\hat{\lambda} = \frac{\overline{m}}{a}$	$\frac{1.65}{1} = 1.65$	$\frac{2.62}{1} = 2.62$
$\hat{M} = \hat{\lambda}A$	$(1.65)(2987) = 4928.55$	$(2.62)(2662) = 6974.44$

The estimated number of four-letter words in the 1950 dictionary is approximately 4,929 and 6,974 for the 1986 dictionary. To get a better idea of possible values for the true number of four-letter words in each dictionary, a confidence interval must be constructed to determine the range in which the true number of four-letter words lies. The variance of the number of four-letter words ($\hat{V}(\hat{M})$) must be computed to determine the bound (or range) for the confidence intervals. The bound is two standard deviations from the mean ($\hat{M}$).

	1950	**1986**
$\hat{V}(\hat{M}) = A^2 \frac{\hat{\lambda}}{an}$	$(2987)^2 \frac{1.65}{(1)(100)} = 147215.79$	$(2662)^2 \frac{2.62}{(1)(100)} = 185659.59$
$\hat{M} \pm 2\sqrt{\hat{V}(\hat{M})}$	$4928.55 \pm 2\sqrt{147215.79}$ $(4161.18, \quad 5695.92)$	$6974.44 \pm 2\sqrt{185659.59}$ $(6112.68, \quad 7836.20)$

Thus, the total number of four-letter words in 1950 is within the range of 4,161 to 5,696 and the total number of four-letter words in 1986 is between 6,113 and

7,836. The actual number of four-letter words (M) for each dictionary lies somewhere within these ranges with a 95 percent level of confidence.

There is a visible difference, with no overlapping, in the confidence intervals for 1950 and 1986. A hypothesis test can be performed to detect whether or not this difference is significant. The population of four-letter words in the dictionary has a Poisson distribution, but because the sample size is 100 for each dictionary, the Central Limit Theorem[4] will allow the use of a normal approximation.

A hypothesis test for the difference in population means can be performed to determine whether there is enough statistical evidence at the 0.05 level of significance to conclude that the mean number of four-letter words in the 1950 dictionary (M_1) is less than the mean number of four-letter words in the 1986 dictionary (M_2). This test relies on the assumption that the samples are independent, due to the method used to gather the data. The null hypothesis claims that the means for 1950 and 1986 are equal, and the alternative hypothesis claims that the mean for 1950 is less than the mean for 1986.

$$H_0 : M_1 = M_2$$
$$H_a : M_1 < M_2$$

The test statistic is shown by the following equation:

$$z = \frac{\hat{M}_1 - \hat{M}_2 - 0}{\sqrt{\hat{V}(\hat{M}_1) + \hat{V}(\hat{M}_2)}} = \frac{4928.55 - 6974.44}{\sqrt{147215.79 + 185659.59}} = -3.55$$

The critical region, the region in which the test statistic must fall in order to reject the null hypothesis, is given by the following inequality:

$$z < -z_{0.05} = -1.645$$

The test statistic does fall within the given critical region because –3.55 < -1.645. Another method used by statisticians to validate hypothesis tests is the use of the p-value. The probability (p-value) of having –3.55 as the test statistic given that the null hypothesis is true $P(z < -3.55 \mid H_0) = 0.0002$. Because 0.002 < 0.05, the null hypothesis can be rejected at the 0.05 level of significance. A conclusion can then be made that the mean number of four-letter words in the 1950 dictionary is significantly less than the mean number of four-letter words in the 1986 dictionary.

Through the use of the population density method, the number of four-letter words was estimated. The difference in the estimated 1950 and 1986 means was found to be almost 2,000 words. This statistically significant growth in four-letter words could be attributed to several factors that include an increase in

[4] The Central Limit Theorem states that the distribution of the sample mean can be approximated to the normal distribution as long as the sample size is large.

the following: technological and medical terms, abbreviations, slang, and many other advances in society that require new vocabulary[5].

References

Book Review. World Wide Words. 7 July 1999 http://clever.net/quinion/words

Scheaffer, Richard L., William Mendenhall III and R. Lyman Ott. *Elementary Survey Sampling* (5th ed.). Belmont: Duxbury Press, 1996.

Webster's New International Dictionary (2nd ed.). Chicago: G & C Merriam Co., 1950.

Webster's Third New International Dictionary Unabridged. Chicago: Merriam-Webster Inc., 1986.

[5] Book Review. http://clever.net/quinion/words/