Thailand Statistician
January 2014; 12(1): 25-35
http://statassoc.or.th
Contributed paper

An estimation of the number of words known by undergraduate students: A case study of paragraph level writing analyzed by Capture-Recapture Method

Krisana Lanumteang* [a] and Yutthasak Chuenchaichon [b]

[a] Section of Statistics, Faculty of Science, Maejo University, San Sai, Chiang Mai 50290, Thailand.

[b] Department of English Language, Faculty of Humanities, Naresuan University, Phitsanulok 65000, Thailand.

* Corresponding author; e-mail: k.lanumteang@mju.ac.th

Received: 18 June 2013
Accepted: 26 July 2013

Abstarct

This paper examines the use of capture-recapture method in estimating the number of words known by undergraduate students. The frequency of repeated unique words from paragraph writing is the variable of interest, which can be used to estimate size of words the students have known in a particular writing topic. We generally considered three well-known estimators including maximum likelihood, Chao's and Zelterman's estimators. Considered estimators perform well under the study condition.

Keywords: Capture-recapture method, zero-truncated poisson, maximum likelihood estimator, Chao's estimator, Zelterman's estimator.

1. Introduction

Capture-recapture modelling is a well-known method used to estimate the size of hidden populations. This model has been widely used to estimate not only population size in ecological science, but also the fundamental demographic factors affecting the size of this particular population [1]. The classical capture-recapture model goes back to the Petersen-and Lincoln-methodology [2], which uses the independent information of two identifying sources or lists to construct an estimator of population size. In this model, individuals are captured, marked and then released to mix with the natural population. The number of individuals caught can be observed along with this capture history. This provides the observed frequency of identifying individuals. From the capture history, an estimate of the number of unobserved cases and hence of the population size can be constructed.

In addition, the identifying system can produce a count of how often each unique observed unit has been identified. As a result, the capture-recapture model providing a count of recaptures tends to be widely applied in a variety of fields, such as to estimate the size of human population particularly in public health science and social science [3, 4].

In this paper, we examine an application of capture-recapture method in a linguistics frame work. We look at the use of this method in estimating the size of words known by students from paragraph writing. This is an alternative method using to measure skill of students about the size of vocabulary that they have or use in a particular topic. It should be noted that the size of vocabulary is very important for English language writing learners since it influences how well they express their ideas and intended messages when they write. Therefore, this is worth estimating the number of words known by students. The estimating procedure will be clearly illustrated in the next section.

2. Research Methodology

2.1 Data and Samples

In this study, we used secondary data from recent work of Chuenchaichon [5]: The development of paragraph writing for EFL writers through the use of a reading into writing method. In this study, participants were 54 second year undergraduate English major students in the four years Humanities program at Naresuan University between the ages of 20 and 22. They were Thai students who were lower intermediate English learners enrolled in the paragraph writing course of the academic year 2009. All of them had nearly almost the same academic background. As the aim of his research is to explore the impact that reading can have on written performance, the participating students were divided into two groups. At the beginning of the study, these two groups had a paragraph-writing pretest in order to evaluate their writing ability in terms of grammatical accuracy, grammatical complexity, and coherence and cohesion. Then, students in each group were selected to ensure that their writing abilities were at the same level. Thus, there was no significant difference of writing ability across the students of these two groups. The control group had the traditional way of teaching in a Thai university context, and the experimental group had a newer approach (i.e. receiving extra reading and doing activities with the reading). The number of the students in the control group was 28 and that in the experimental group was 26. After finishing the experimental period, participants were required to write a paragraph in English (at least 10 sentences) with the topic "***What do you like about life at university?***". Here, 10 pieces of student paragraph writing from both control and experimental groups were selected with simple random sampling to be our samples. We can reasonably assume that there are some words known by students in the content of this writing topic, but did not use in the paragraph. If we can estimate this number, we then get the total number of words they have known. This estimating process will be discussed in the next sub-sections *2.2* and *2.3*.

2.2 Formulating Problem

From the capture-recapture experiment, the frequency counts of identified individuals are the variables of interest. The identifying system generally provides a count $Y_i > 0$ of how many times the individual i^{th} has been captured, for $i = 1, 2, \ldots, n$ and $Y_i = 0$ denotes unobserved cases in the system for $i = n + 1, n + 2, \ldots, N$. Hence, if we let p_0 be the capture probability for unobserved individuals, then $N(1 - p_0)$ is the expected number of observed cases which can be estimated by sample size n. This leads to the simple equation to estimate the population size N, $N = Np_0 + N(1- p_0) = Np_0 + n$. Therefore, this equation can be solved for estimating N to provide the Horvitz-Thomson estimator [6] of the form

$$\hat{N}_{\text{HTE}} = \frac{n}{(1 - p_0)} . \tag{1}$$

However, p_0 is typically unknown and the estimator in (1) would require an estimator of p_0. As capture-recapture system provides only count data of observed cases

(Y_i > 0), the zero-truncated Poisson distribution might be a good candidate model for this probability. Some significant formulas under this condition will be seen in sub-section *2.3*.

Similarly, we can identify how many times that each unique word has been repeated in the paragraph level writing. Let $f_1, f_2, f_3, \ldots, f_m$ be the frequencies of unique words written exactly 1,2,3,…,*m* times in the whole paragraph, and $n = \sum_{j=1}^{m} f_j$ is the total number of distinct words. Therefore, $S = \sum_{j=1}^{m} jf_j$ denotes the total word count of the paragraph. In addition, let f_0 be the number of words known by writers, but did not use (hidden words) in the paragraph. Consequently, the total number of words known by writers (*N*) can be now defined as $N = f_0 + f_1 + f_2 + f_3 + \ldots + f_m \; = f_0 + n$. Note that the estimated size of words known by writers covers only the content of this writing topic. This unknown size (*N*) can be obtained if we can estimate the number of hidden words, f_0. Estimating formulas will be shown in the next subsection.

2.3 Considered Estimators

In order to estimate the size of hidden words, we considered three well-known estimators which required to assume a Poisson model as the capture probability. Here, f_j the frequencies of unique words written exactly *j* times in the whole paragraph are the variables of interest and it might be impossible to determine the largest possible count of identifications; consequently, Poisson model might be more suitable for this particular type of data. Considered estimators are not only more appropriate for our study condition but their formulas are also very simple to understand and to use as follows:

2.3.1 Maximum likelihood estimator ($\hat{N}_{MLE}$)

$$\hat{N}_{MLE} = \frac{n}{1 - exp(-\hat{\lambda}_{MLE})}; \tag{2}$$

where $\hat{\lambda}_{MLE}$ is the maximum likelihood estimator of the unknown parameter under the zero truncated Poisson distribution. As $\hat{\lambda}_{MLE} = \bar{y}(1 - exp(\hat{\lambda}_{MLE}))$, $\bar{y} = \frac{1}{n}\sum_{j=1}^{m} jf_j$ is not the closed form, the Expectation–Maximization(EM) algorithm is applied for required iterative method. A variance of (2) can be estimated as:

$$\hat{V}(\hat{N}_{MLE}) = \frac{\hat{N}_{MLE}}{(\exp(\frac{\sum jf_j}{\hat{N}_{MLE}}) - \frac{\sum jf_j}{\hat{N}_{MLE}} - 1)}$$, see [7]. (3)

2.3.2 Chao's estimator ($\hat{N}_{Chao}$)

$$\hat{N}_{Chao} = n + \frac{f_1^2}{2f_2},$$ (4)

and $\hat{V}(\hat{N}_{Chao}) = (\frac{1}{4})^2 \frac{f_1^4}{f_2^3} + \frac{f_1^3}{f_2^2} + \frac{1}{2}\frac{f_1^2}{f_2}$, see [8]. (5)

2.3.3 Zelterman's estimator ($\hat{N}_{Zel}$)

$$\hat{N}_{Zel} = \frac{n}{1 - \exp(-\frac{2f_2}{f_1})},$$ (6)

A simple variance formula for (6) can be obtained as:

$$\hat{V}(\hat{N}_{Zel}) = n(\frac{\exp(-\frac{2f_2}{f_1})}{(1 - \exp(-\frac{2f_2}{f_1}))^2}) \times \delta;$$

where $\delta = 1 + n(\frac{\exp(-\frac{2f_2}{f_1})}{(1 - \exp(-\frac{2f_2}{f_1}))^2})(\frac{2f_2}{f_1})^2(\frac{1}{f_1} + \frac{1}{f_2})$, see [9,10]. (7)

2.4 Testing Hypothesis

In this study, we also would like to compare the average number of words known by students from both control and experimental groups. Moreover, comparing the proportion of hidden words based upon the estimated total size of words known is of our interest. We compute the estimated number of words known by each student from the average of three considered estimators($\bar{\hat{N}}$); $\hat{N}_{MLE}$, $\hat{N}_{Chao}$ and $\hat{N}_{Zel}$. Lanumteang and Böhning [11] showed that the estimated population size($\hat{N}$) has an approximately normal distribution. As a result, independent two-sample *t*-test seems reasonable to be used for testing the difference in means of the two groups.

3. Results

We found that samples of both groups wrote the paragraph on average of 174.35 (±32.39) words, which yielded the mean of unique word count approximately 93.05 (±15.79) words. An average of repeated written words was 1.88 (±0.19) times, of which 62.70 (±2.86) appear only once and 12.50 (±0.61) twice. The maximum counts of repeated written words are 19 and 18 times, which are the word “I” and “you”, respectively. The frequency count of repeated unique words is shown in Figure1.

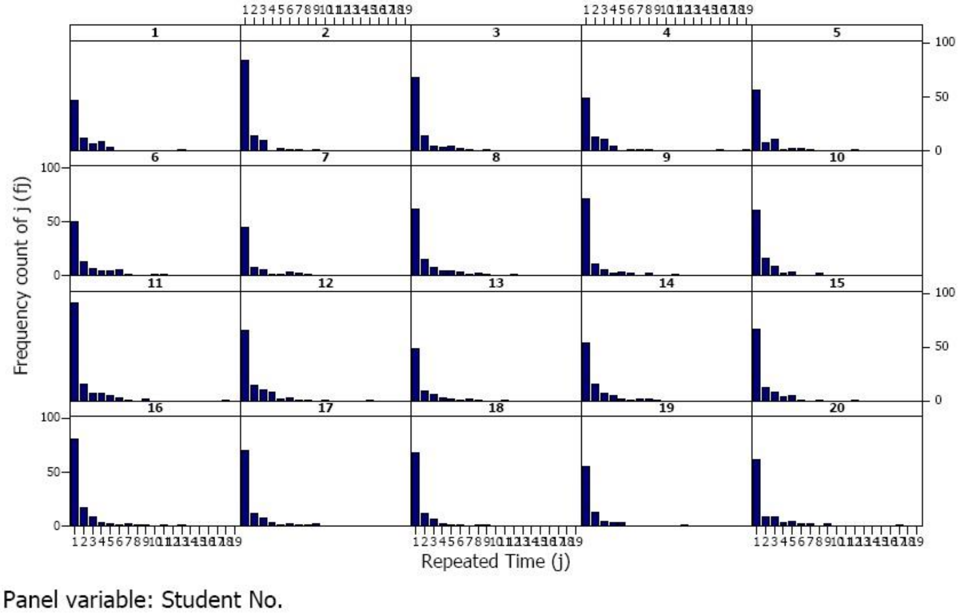

Figure 1. Frequency count of repeated unique words.

The ratio plot$\left(R_j = \frac{(j+1)f_{j+1}}{f_j} \quad vs \quad j \right)$was applied to detect population homogeneity or heterogeneity for approving appropriate uses of considered estimators, see [12] for review. As can be seen from Figure 2, there is a linear trend of the ratio plot of identifying words (R_j) as the counts of repeated words (j) increases. This is evidence of the presence of population heterogeneity in repeated times. As Chao's estimator is developed under the assumption of heterogeneity in capture probability, it might be more appropriate for this application. On the other hand, *MLE* and Zelterman's estimator are more suitable under homogeneity Poisson model.

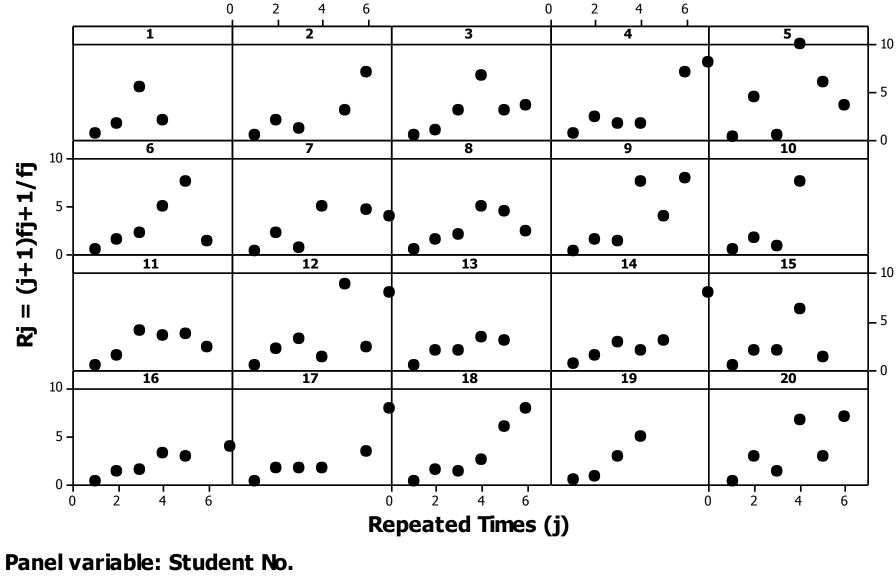

Figure 2. Ratio plot of repeated unique words.

Table 1 shows summary results of the study. *MLE* provided the smallest estimator on average 125.75 (±26.16) words, whereas Zelterman's estimator gave the highest value, about 296.85 (±78.62) words. The largest estimated number of words known by students was 408 (±69.97) words, provided by Chao's estimator. In contrast, the lowest estimated number of words known by students was only 87 (±6.80) words, given by *MLE*. According to an average estimated number of three considered estimators, this yielded the mean of approximately 228.07 (±55.94) words. The highest and lowest estimated values on average of all estimators was 349 (±158.37) and 156 (±52.20) words, respectively. If we denote the proportion of hidden written words as $(\frac{\hat{N}-n}{\hat{N}})\times 100\%$, then we have this proportion on average about 58.10% (±6.44%). Remarkably, the highest proportion of hidden words known by students in this particular writing topic was 69.03% whereas the lowest was 47.23%.

In comparison, the average estimated number of words known by students in control and experimental groups is 217.10 (±57.07) and 238.80 (±55.27) words, respectively. From the testing hypothesis, we do not have strong evidence to conclude that the control group in a mean numbers of words known differs from the mean numbers of words known by treatment group (p > 0.05). Similarly, the proportion of hidden written words the students have known in this particular writing topic of both groups are also not different (p > 0.05). The proportion of control and experimental groups on average are 57.99% (±7.73%) and 58.21% (±5.27%), respectively, see Table 2 and Figure 3.

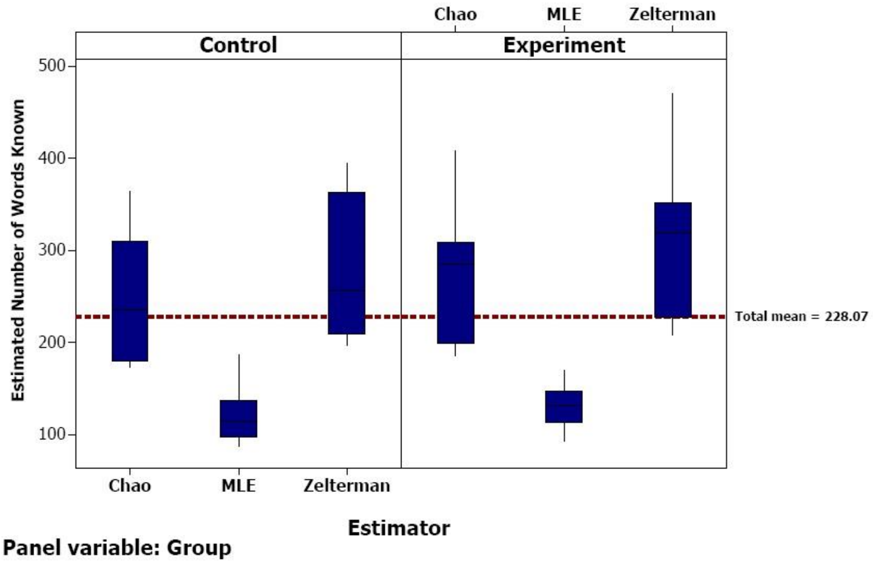

Figure 3. Boxplot of estimated number of words known.

Table1. Estimated number of words known by students.

No.	Word Count	Unique Word Count (n)	Repeated Mean	Estimated number of words known by students				Proportion of hidden words
				$\hat{N}_{MLE}$ **(*SE*)**	$\hat{N}_{Chao}$ **(*SE*)**	$\hat{N}_{Zel}$ **(*SE*)**	$\bar{\hat{N}}$ **(*SE*)**	
1	147	76	1.93	98 (*7.03*)	176 (*34.46*)	203 (*56.50*)	159 (*31.48*)	52.20
2	171	112	1.53	187 (*17.94*)	364 (*66.41*)	395 (*101.16*)	315 (*64.79*)	64.48
3	168	97	1.73	138 (*10.90*)	262 (*47.50*)	287 (*72.19*)	229 (*46.07*)	57.64
4	177	81	2.19	97 (*5.36*)	173 (*30.86*)	197 (*49.45*)	156 (30.14)	47.97
5	145	80	1.81	109 (*8.66*)	304 (*74.83*)	362 (*132.51*)	258 (*76.52*)	69.03
6	191	86	2.22	102 (*5.30*)	182 (*31.84*)	212 (*53.34*)	165 (*32.83*)	47.98
7	128	67	1.91	87 (*6.80*)	194 (*45.29*)	224 (*75.12*)	168 (*41.58*)	60.20
8	165	90	1.83	121 (*8.84*)	301 (*38.22*)	340 (*55.96*)	254 (*67.45*)	64.57
9	173	98	1.77	137 (*10.38*)	327 (*66.18*)	368 (*106.53*)	277 (*71.16*)	64.66
10	159	93	1.71	134 (*11.09*)	209 (*34.85*)	228 (*51.95*)	190 (*28.69*)	51.14
11	256	132	1.94	170 (*9.21*)	408 (*69.97*)	470 (*115.89*)	349 (*91.43*)	62.21
12	223	106	2.10	129 (*6.66*)	257 (*44.26*)	303 (*75.23*)	230 (52.05)	53.85
13	145	73	1.99	93 (*6.51*)	201 (*44.14*)	233 (*73.47*)	176 (42.35)	58.44
14	180	89	2.02	111 (*6.75*)	186 (*30.90*)	209 (*48.56*)	169 (29.59)	47.23
15	182	99	1.84	133 (*9.23*)	286 (*54.81*)	329 (*90.06*)	249 (*59.48*)	60.29
16	222	118	1.88	156 (*9.50*)	311 (*50.79*)	344 (*78.24*)	270 (*57.95*)	56.35
17	177	99	1.79	137 (*10.08*)	303 (*58.78*)	341 (*93.89*)	260 (*62.64*)	61.97
18	146	92	1.59	145 (*14.09*)	285 (*56.12*)	309 (*85.18*)	246 (51.14)	62.65
19	132	79	1.67	117 (*11.03*)	195 (*36.89*)	210 (*53.91*)	174 (*28.83*)	54.60
20	200	94	2.13	114 (*6.14*)	308 (*66.50*)	373 (*119.42*)	265 (*77.80*)	64.53
Mean (*SE*)	**174.35 (*7.24*)**	**93.05 (*3.53*)**	**1.88 (*0.04*)**	**125.75 (*5.85*)**	**261.66 (*15.35*)**	**296.85 (*17.58*)**	**228 .07 (*12.51*)**	**58.10 (*1.44*)**

Note : No. 1- 10 is control group and No. 11-20 is treatment group.

Table 2. An average of estimated number of words known by students of each group.

Group	$\hat{N}$		$' \ \hat{N} - n) \times 100/\hat{N}$	
	Mean	SD	Mean	SD
Control	217.10	57.07	57.99%	7.73%
Treatment	238.80	55.27	58.21%	5.27%
t-Test	$t_{df=18} = -0.86$, $p > 0.05$		$t_{df=18} = -0.08$, $p > 0.05$	

4. Conclusion & Discussion

To sum up, we studied the use of capture-recapture method in estimating the number of words known by undergraduate students in the particular writing topic. The writing task in this study was “***What do you like about life at university?***”. The participants were 20 second year undergraduate English major students in the Humanities program at Naresuan University. Overall, participants wrote the paragraph on average 174.35 (±32.39) words. The mean of distinct word count was 93.05 (±15.29) words and the average of repeated written words was 1.88 (±0.19) times. Considered methods yielded the estimated number of words known by students on average 228.07 (±55.94) words. This led to the proportion of hidden written words the students have known in this particular writing topic approximately 58.10% (±6.44%). If we look at the ratio plot, which shows the heterogeneity of word repeating, Chao’s estimator might be more suitable for this data. In addition, the majority of written words in our studied paragraph appeared only once or twice and Chao’s estimator also requires only the count of words repeating once and twice. This estimator gave the estimated number of words known by participants ranging between 173 (±30.86) and 408 (±69.97) words. It can be said that these might be minimum words known by students as Chao’s estimator generally provides the lower bound. Remarkably, our results (the estimated number of words known by students) cover only this particular of writing topic. The participants in this study wrote/used only words that relate to this writing topic. This might explain why the difference between the means of words known by students from both groups is not statistically significant. As the paragraph writing gave quite small number of distinct words, we should extend the content of the writing topic or look at more pieces of writing for each student in further study. Then, we can more deeply explain the main focused question of this study (the benefit of our work) about how many words students knew but did not use. In addition, the findings of this study shed some light on the teaching of English writing. English teachers should provide students with a wide variety of vocabulary words and encourage their students to practice using them in different contexts. By doing this, students would gain more knowledge about word meanings and word choices. As a result, they can make use of their word banks to improve their writing skills and become better writers.

References

[1] Amstrup, C., McDonald, L., and Manly, J., Handbook of capture-recapture analysis, Princeton Univerity Press, New Jersey, 2005.

[2] Seber, G., The estimation of animal abundance and related parameter 2nd ed., Blackburn Press, Caldwell, NJ, 2002.

[3] Pollock, H., Capture-recapture models, *Journal of the American Statistical Association*, 2000; 95: 293-296.

[4] Böhning, D., and van der Heijden, P., Recent developments in life and social science applications of capture-recapture methods, *AStA Advances in Statistical Analysis*, 2009; 93:1-3.

[5] Chuenchaichon, Y., The development of paragraph writing for EFL writers through the use of a reading into writing method, PH.D. Thesis, University of Reading, UK, 2011.

[6] Horvitz, G., and Thompson, J., A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association.*1992; 47: 663-685.

[7] Chao, A. and Lee, S., Estimating the number of classes via sample coverage, *Journal of the American Statistical Association*, 1992; 87: 210-217.

[8] Chao, A., Estimating the population size for capture-recapture data with unequal catchability, *Biometrics,* 1987; 43: 783-791.

[9] Zelterman, D., Robust Estimation in truncated discrete distributions with application to capture-recapture experiments, *Journal of Statistical Planning and Inference,* 1988; 18: 225-237.

[10] Böhning, D., A simple variance formula for population size estimators by conditioning, *Statistical Methodology*, 2008; 5: 410-423.

[11] Lanumteang, K., and Böhning, D., An Extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics and Data Analysis*, 2011; 55: 2302-2311.

[12] Rocchetti, I., Bunge, J., and Böhning, D., Population size estimation bases upon ratios of recapture probabilities, *Annals of Applied Statistics*, 2010; 5: 1512-1533.