Journal of Statistics Education

How many words do you know? An integrated assessment task for introductory statistics students

David I. Warton
The University of New South Wales, Australia

Journal of Statistics Education Volume 15, Number 3 (2007),
www.amstat.org/publications/jse/v15n3/warton.html

Copyright © 2007 by David I. Warton all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

Key Words: active learning, dictionary, higher education, lexical richness, service courses, vocabulary size

Abstract

A novel assignment exercise is described, in which students use a dictionary to estimate the size of their vocabulary. This task was developed for an introductory statistics service course, although it can be modified for use in survey sampling courses. The exercise can be used to simultaneously assess a range of core statistics skills: sample size estimation, obtaining a simple random sample, estimating a sample proportion, measuring the sample error of this proportion, and similarly for a scalar multiple of a proportion. The outcome of this exercise involves the student discovering something about themself, which serves as a natural motivator and a tool for generating interest in the discipline of statistics.

1. Introduction and Aims

The education literature emphasizes the importance of active learning activities to encourage deep learning experiences (Brookfield, 1991). Many statistics educators emphasize the importance of using teaching tasks that encourage active learning, and design experiments or mini-projects accordingly (Stephens, 1977; Fisch, 1988; Smith, 1998; Zhang, 1998; Paranjpe and Shah, 2000; Hakeem, 2001). Rather than simply taking a dataset and analyzing it, students undertake a small research task, which involves designing a study, collecting the data *then* analyzing it to draw conclusions. There are many advantages to this approach. For example, students will be more likely to see statistics as interesting and relevant to them if they can see it work in a research context. There is often a sense of ownership or responsibility for the data, given that the students collected the data themselves. When different datasets are collected by different (groups of) students, then there is no longer a single right answer, but only a right method, and so students must focus on the desired outcome - learning "the method" rather than trying to get "the answer."

Paranjpe and Shah (2000) proposed a dictionary-based exercise for a class on sampling methods, in which students compare the efficiency of different methods of sampling for estimating the number of words in the *Concise Oxford Dictionary*. I also propose a dictionary-based exercise, but in a different context and with a different intended outcome.

In this paper, I describe an exercise in which the purpose is for students to estimate the size of their vocabulary. I have used this exercise as an out-of-class assessment for statistics service course students, although with a change in emphasis, it could readily be adapted to a survey sampling course. The project involves several steps: sample size estimation; the non-trivial task of obtaining a representative sample of words from a dictionary, for which a range of sampling strategies could be used; estimating a sample proportion and making inferences about it; and inference about a simple transformation of the sample proportion. Hence several different statistics service course topics can all be assessed using this single assessment exercise.

2. The Dictionary Exercise

The basic principle behind the proposed exercise is that if one has access to a dictionary that is sufficiently large that it contains every word that you know, then you can estimate the size of your vocabulary as pN where N is the number of words in the dictionary and p is an estimate of the proportion of words in the dictionary that you know. If, for example, you sample 40 words from a dictionary of 100 thousand words, and you know the meaning of half of them, then the estimated size of your vocabulary is 50 thousand words.

I explain this principle to the class, then tell the class that their task is to use a dictionary to estimate the size of their vocabulary and to construct a 95% confidence interval for the size of their vocabulary. I have implemented this type of assignment task in two different types of course - an introductory statistics course and a survey sampling course. Details of my implementations of the written assignment task are in Appendix A and B. The focus in this paper, except where otherwise stated, is on implementation in introductory statistics service courses.

To complete the task, students will be required to:

- Choose a large dictionary, which they think can reasonably be assumed to contain all the words that they know. The dictionary must state the number of words that are defined in it.
- Determine a random sampling technique for sampling from the dictionary. In the case of introductory statistics students, I specifically ask them to obtain a simple random sample of words from the dictionary.
- Determine a sample size sufficient to estimate the size of their vocabulary, to within a level of precision that they decide upon.
- Decide upon an objective definition for determining whether the student knows a given word or not.
- Obtain a sample and test themselves on the words in this sample.
- Estimate the proportion of words that they know, from their sample, and convert this into a confidence interval for the total number of words that they know.

This is a challenging task and some guidance can be offered at various stages, as appropriate for the group of students and the course in which this task is implemented. The relative importance of learning outcomes related to each of the above tasks should serve as a guide as to the sort of help students should be given. For example, if sample size determination is not taught or is considered too difficult, students could be instructed to sample 100 words. If confidence intervals are not taught by the time that the assignment is given to students, students could be asked to construct an estimate of their vocabulary size, possibly with a standard error, but without interval estimation.

I do not tell students which dictionary to use and suggest to them that they may use a non-English dictionary if they like. Allowing students to choose their own dictionary, but to explain why they chose it, offers a means of assessing their understanding of the necessity of using a large, modern dictionary which can reasonably be assumed to contain all the words that they know. By allowing non-English dictionaries, those for whom English is a second language may do the activity in a language that they are more confident and comfortable with (provided that they have access to a large, modern, non-English dictionary). In addition, some students have an active interest in learning another language and allowing non-English dictionaries takes advantage of this interest.

An important issue in dictionary choice is that the chosen dictionary must state the number of words defined in the dictionary, such that N is known, if one is to use pN to estimate vocabulary size. Many dictionaries do not give the exact number of words defined, but report an approximation (e.g. "over 110,000 words are defined"), usually found in the blurb on the back-cover of the dictionary. I consider such an approximation as being sufficient for the purpose of this assignment, especially considering that sampling error will likely be of a much larger order than error in lack of knowledge of the exact value of N. Not all large dictionaries state the total number of words they contain, but it was not difficult for me to find a few suitable candidates (for example the Macquarie Dictionary, 2001). The requirement that N is known can be lifted for some alternative sampling schemes, discussed later.

Some students suggest the use of an electronic dictionary, particularly one that has a function that generates "random" words. I do not allow the use of electronic dictionaries for this exercise, because many of the intricacies of the project (in particular, the identification of an appropriate sampling scheme) are lost in using an electronic dictionary rather than a physical book. Additionally, before allowing the use of an electronic dictionary's random word generator, one would need to be aware of whether a simple random sample of words is in fact generated, or if there are biases in the sampling approach.

3. Comments on Implementation

The dictionary exercise described above is deliberately open-ended, with students choosing the method of sampling, how to define knowledge of words, and how many words they will sample. By requiring students to make decisions at all of these key stages, rather than following a list of instructions, they are encouraged to think more deeply about the problem and appreciate the difficulties often inherent in research. Specifically:

- By choosing the method of sampling words themselves, students get to explore different sampling strategies and apply principles they have learned in class regarding sampling. They develop an appreciation for how difficult it can be to obtain a simple random sample in some practical situations.
- By having to define knowledge of words themselves, students get hands-on experience with the difficulties often encountered obtaining objective measurements of a variable of interest.
- Students are required to specify a desired level of precision such that they can determine the sample size. Choosing the desired level of precision is a critical stage of sample size determination in any study and it is a stage that many applied researchers have difficulty with. By having to choose a level of precision, rather than being told the level of precision to use, students will develop an appreciation for the issues involved and an awareness of the difficulties identifying a desired level of precision in practice.

In summary, by setting the dictionary task for students with minimal "ground rules" students have a more authentic experience of applied research and will experience a range of challenges key to designing and conducting any observational study.

While it was previously stated that it is necessary that a dictionary is used for which the total number of words N is known, this requirement could be avoided by using an alternative sampling scheme. For example, a simple random sample of pages (instead of a sample of words) could be taken, then knowledge of *all* words on a given page could be tested (or at least, of all words on a randomly chosen column of the page). Then, if there are N_p pages in the dictionary, the total size of your vocabulary could be estimated as mN_p, where m is the average number of words on a page that are known (and similarly if columns were sampled). The advantage of this approach is that it can be applied without knowledge of N, indeed it could even be used to obtain an estimate of N (as in Paranjpe and Shah, 2000). Further, obtaining a simple random sample of pages from the dictionary is a straightforward matter. I have chosen not to use this approach in introductory statistics courses and I instead specifically ask them to obtain a random sample of *words* not pages (as in Appendix A). This way students learn that often it is quite difficult to obtain a simple random sample from a population (as is the case when sampling words, but not pages, from a dictionary).

There are many different methods of sampling in order to estimate the size of your vocabulary from a dictionary and so an exploration of these alternative methods is a useful exercise for a survey sampling course. In sampling words, one can readily obtain a simple random sample (although with some difficulty), a systematic random sample, or a cluster sample (sampling pages, then words). In sampling pages, one can use a simple random sampling or systematic sampling approach, and base inference on the average number of known words per page, or on a ratio estimator (if N is known). Of the random sampling methods studied in a typical survey sampling course, most are applicable for the dictionary problem. I have in fact made use of this and implemented a version of the dictionary exercise for a survey sampling course (Appendix B). In this alternative implementation, students were required to propose several methods for sampling words, use two alternative methods that they think are the most suitable for this problem, and compare their efficiency.

4. Sampling Difficulties

There are many difficulties encountered in attempts to achieve a simple random sample of words from a dictionary and test one's knowledge of those words. It is important for students to obtain feedback on their proposed sampling approaches and for this purpose I dedicate a tutorial class (*i.e.* a weekly one hour class of 20-30 students) on sampling issues to the dictionary exercise. The main task is for students to explore various ideas about how to randomly sample words from a dictionary and get feedback on their ideas from the instructor. The instructor brings to class several dictionaries of different types and sizes, for the students to look at in groups. Students look at the different dictionaries, to get a sense for issues involved in choice of dictionary. Students propose sampling strategies and get feedback from the instructor, who ensures that students are made aware of various difficulties encountered and how to resolve them, as described below.

- Students need to have access to a large dictionary which can be reasonably considered to contain all the words that they know. Most university library collections contain several such dictionaries.
- Languages change over time and from one place to another and so the dictionary chosen should be modern and should contain local jargon.
- It is important to have an objective, replicable definition of whether or not a word is known, such as use of the word in a sentence to illustrate its meaning.
- Lengths of definitions vary considerably across words and the more common words tend to have the longest definitions (due in part to the variety of contexts in which they are used). So if one randomly samples a position on a page, rather than randomly sampling a word on the page, the estimate of the proportion of words they know will be upward biased.
- Different pages contain different numbers of words. So if one uses a two-stage sampling technique to sample words, randomly sampling a page then a word from the page, this will not result in a simple random sample, unless some adjustment is made to the sampling technique.

In class discussions, many of the above issues are identified by students. All the above issues should be addressed in class, so that students have an appreciation for the issues involved in completing the exercise.

There are few successful strategies that deal with all of the above situations, so I tend to be lenient with students who are aware of all of the above issues but have not perfectly resolved all of them in their chosen sampling technique.

An example sampling scheme which does resolve the above issues is the following:

1. Observe the number of pages in the dictionary, N_p.
2. Observe the number of words on several different pages of the dictionary, in order to estimate a number W which will be larger than the maximum number of words on any page of the dictionary.
3. Randomly sample a page $i \in \{1,..., N_p\}$. Go to this page of the dictionary.
4. Randomly sample a number $j \in \{1,..., W\}$. Count down the selected page until you reach the jth word. If the selected page does not contain j words, return to step 3.
5. Test your knowledge of this word, then return to step 3 until you have tested yourself on the required number of words.

This sampling scheme ensures that all words have the same chance of being sampled and that the chance of being sampled is independent of which words have previously been sampled. Note that strictly speaking, the above scheme does not obtain a simple random sample, because it involves sampling with replacement. The chance of repeat words is trivially small, however if a simple random sample is specifically required, this can be achieved by adding an intermediate step after step 4 in which repeat words (which are highly unlikely!) are discarded.

Note also that the above sampling scheme is closely related to Lahiri's method (Lahiri, 1951) for obtaining a random sample with probability proportional to size. Indeed this is in fact Lahiri's method, if the above sampling scheme is used in order to obtain a random sample of *pages*, with probability of page selection proportional to the number of word definitions on each page.

5. Discussion

I have found that this assessment task is well-received by students. A measure of the effectiveness of a task is the depth of the questions one receives in class - and on this evidence it appears that this task is effective at stimulating students to think deeply about sampling issues and the issues involved in sample size determination. The following are some examples of questions I am commonly asked by students: "But how many samples should I take?", "How do I decide how precise I want my estimate to be?" and "Why not use a systematic sample rather than a random sample?". These are all good questions that arise naturally in this exercise, but questions which one would otherwise be unlikely to hear from statistics service course students. This type of deep response from students has been encountered previously in using "research problems" as teaching tools (Smith, 1998), rather than just working through statistical analyses in class without context.

A key attraction of the proposed assessment task is that in estimating the size of their vocabulary, the student is finding out something about themselves. This assignment is expected to stimulate interest due to the personal nature of the subject matter and is intended to give rise to an appreciation for the power of statistical tools through the experience of completing this task.

This task might be expected to be more robust to plagiarism than the typical take-home assignment. In particular, the vocabulary estimation task has no single correct answer and it is has a personal nature. Such assessment tasks tend to be less subject to plagiarism (Carroll, 2002). Of course, as with any data collection assignment, there is the inevitable possibility of students making up their own data, in more pathological cases.

Many students find this task challenging, which in itself is not a problem, given that challenges are often seen by students as the most rewarding and important experiences, empowering students (Brookfield, 1991). Conceptually, some students have difficulty with the original concept - that *pN* can be used as an estimator of the size of their vocabulary, but only if one has a dictionary that contains all of the words that you know. It is important that this idea be carefully worked through with students initially, such that all understand the fundamental theory underlying the exercise.

The very definition of vocabulary size is somewhat nebulous, due in part to difficulties defining knowledge of a word (Laufer et al., 2004) and in part due to the definition of what a word is in the first place (Nation and Waring, 1997). Typical approaches to this exercise result in an estimate of the student's passive recall vocabulary (Laufer et al., 2004) - the number of words that are recognized when given the word and for which the student is able to recall a meaning for the word or construct a use of it. Dictionaries often state their size in terms of the number of "headwords" *i.e.* distinct words (excluding inflected forms) that have a distinct derivation. For example, the Macquarie Dictionary (2001) lists "installer" and "installers" as inflected forms of the headword "install" whereas "installation" and "instalment" are listed as separate headwords. In contrast, language experts often refer to knowledge of "word families" *i.e.* groups of words consisting of a base word, inflected forms, and transparent derivations (Nation and Waring, 1997). Unfortunately, different experts often define word families in slightly different ways. However, most would be expected to classify "install," "installer(s)," "installation" and "instalment" as members of the same word family, given that they share a common derivation. A conservative estimate of the number of word families in the English language is 54,000, and a typical native-speaking university student might be expected to know 20,000 of them (Nation and Waring, 1997). In my experiences with this assignment task, students typically know around 50,000 headwords from a large dictionary.

A potential disadvantage of the proposed task is that marking assignments can be quite time-consuming. Marking is complicated by the fact that there is no single correct answer, so all calculation steps need to be checked, rather than just the final answer. To reduce marking times, I enforce a strict page limit (two A4 or letter pages should be sufficient) and use a simple marking scheme. I have previously marked the project out of five, with one mark each for demonstrating various skills important for completion of the project (*e.g.* recognition of important issues in selecting a dictionary, correct calculation of a confidence interval for *p*...).

For a class of 30 students, it typically takes me approximately one hour to mark responses to this assignment task, which most would consider as a satisfactory time investment, in the sense that it is sufficient time to fairly grade students, while also not being excessively demanding on the instructor's time. In my work environment, most classes are taught and marked by statistics academics who are already under considerable time pressure, so constraining marking loads is an important consideration.

In this paper, I have described implementations of a dictionary-based exercise for both an introductory statistics course and a survey sampling course. Variations on this exercise are possible - for example, a reviewer suggested estimating the relative frequency of words starting with different letters and comparing these frequencies with the point scores available in the game Scrabble. Undoubtedly many other dictionary-based exercises exist and can be used to enliven statistics classes!

Appendix A

Example dictionary exercise question (for an introductory statistics course)

Use a dictionary to estimate the number of words you know. You need not estimate the size of your English vocabulary - you can estimate how many words you know in another language if you like. Note that you there are many large dictionaries available in the reference section of the library.

1. Explain how you did this task, and why did it the way you did. Consider in your answer:
   (a) What dictionary you used, how many words are defined in it.
   (b) How you sampled words. Include sufficient details that someone else could repeat what you did.
   (c) How you decided whether or not you knew a word.
   (d) How many words you included in your sample.
2. Estimate the proportion of words in the dictionary that you know.
3. Calculate a confidence interval for this estimate.
4. Hence estimate the number of words you know, and include a confidence interval for this estimate.
5. What assumptions must be satisfied for your estimate of the total number of words you know to be unbiased?

Appendix B

Example dictionary exercise question (for a survey sampling course)

In this assignment, you will use survey sampling methods to estimate the size of your vocabulary using a dictionary.

There is more than one way to do this - for example, if you know the number of words in the dictionary N, you can use survey sampling techniques to estimate the proportion of words in the dictionary that you know p. Then an estimate of the number of words that you know is pN. Alternatively, you could sample pages, not words - estimating the average number of words you know per page m, then calculating the number of words you know as mN_p where N_p is the total number of pages in the dictionary. Note that for each of the above approaches, there is more than one survey sampling technique that might be suitable (for estimating p or m).

For this assignment, you should use a modern, large dictionary *e.g.* a dictionary containing at least 100,000 words), which states (at least approximately) how many words are in the dictionary. Note that many such dictionaries are available in the library. It is important that you choose a dictionary that is sufficiently large that it is reasonable to assume that (nearly) all the words that you know are in the dictionary. Note that if you speak a second language, you are welcome to test the size of your vocabulary in this other language.

1. First, answer the following preliminary questions.
   (a) Briefly describe the dictionary you will use, how many words are in it, and why you chose to use it.
   (b) Describe how you intend to test yourself on whether or not you know a word. Include sufficient details that someone else would be able to use the same method.
2. Making use of survey sampling methods we have discussed in lectures, construct **four** different sampling strategies that could be used to estimate the size of your vocabulary. In each case, describe:
   (a) The sampling strategy. Include sufficient details that someone else would be able to sample exactly as you intend by following your instructions.
   (b) Some potential advantages and disadvantages of using this approach, as compared to the other three approaches you consider.
3. Choose **two** of the above methods, which you will use to estimate the total number of words that you know. For each of these two methods:
   (a) Briefly explain why you chose these methods rather than the alternatives considered in question 2.
   (b) Estimate the sample size required to achieve a pre-determined level of accuracy in estimating

the number of words that you know.
(c) Sample from your dictionary in order to estimate the total number of words that you know. Use your results for question 3b as a guide for sample size, but also note that you are not expected to spend longer than an hour sampling via either method.
(d) Showing all working, estimate the number of words that you know. Express your answer as a confidence interval.

4. (a) Compare the relative efficiency of the two sampling approaches you used in question 3.
(b) If you were to make recommendations to next year's students concerning a good sampling method to use for this project, what would be your recommendations? Explain.

References

Brookfield, S.D. (1991), *Facilitating adult learning: a transactional process*, Krieger, chap. Grounding teaching in learning, pp. 33-56.

Carroll, J. (2002), *A handbook for deterring plagiarism in higher education*, Oxford: Oxford Brookes University.

Fisch, L. (1988), "Students Become Data - Statistics Comes Alive," *College Teaching*, 36, 153.

Hakeem, S.A. (2001), "Effect of Experiential Learning in Business Statistics," *Journal of Education for Business*, 77, 95-98.

Lahiri, D.B. (1951), "A method of sample selection providing unbiased ratio estimates," *Bulletin of the International Statistical Institute*, 33, 133-140.

Laufer, B., Elder, C., Hill, K., and Congdon, P. (2004), "Size and strength: do we need both to measure vocabulary knowledge?" *Language Testing*, 21, 202-226.

Macquarie Dictionary (2001), *Macquarie Dictionary, revised third edition*, Sydney, Australia: The Macquarie Library.

Nation, P. and Waring, R. (1997), *Vocabulary: description, acquisition and pedagogy*, Cambridge: Cambridge University Press, chap. Vocabulary size, text coverage and word lists, pp. 6-19.

Paranjpe, S. A. and Shah, A. (2000), "How Many Words in a Dictionary? Innovative Laboratory Teaching of Sampling Techniques," *Journal of Statistics Education*, 8.

Smith, G. (1998), "Learning Statistics By Doing Statistics," *Journal of Statistics Education*, 6.

Stephens, L.J. (1977), "Getting the Students Involved in the Elementary Statistics Course, *Two-Year College Mathematics Journal*, 8, 19-21.

Zhang, J. (1998), "Preparing Students To Solve Practical Problems: An Elementary Statistics Course with a Group Final Project Requirement," *Primus*, 8, 67-83.

David I. Warton
School of Mathematics and Statistics
The University of New South Wales NSW 2052
Australia
David.Warton@unsw.edu.au

Volume 15 (2007) | Archive | Index | Data Archive | Information Service | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications