Decoding a Scrambled Text: A Hands-On Project to Illustrate Sampling and Variability

Wlodzimierz Bryc
University of Cincinnati

Journal of Statistics Education v.7, n.2 (1999)

Key Words: Active learning; Statistics education; Substitution code; Teaching statistics; World Wide Web material.

Abstract

Students crack a simple substitution code using character frequencies in texts sampled from web pages. Frequencies are tabulated by a web-based character counter. This quick and simple project reinforces notions of sampling variability and emphasizes the need to complement statistical techniques with intuition.

1. Introduction

1 Instructors want students to think of statistics as a process for gaining information, rather than merely plugging numbers into formulas. They want students to understand variability and how it is related to statistics. Accordingly, educators recommend that introductory statistics courses should emphasize data collection, variation, graphical display of data, design of experiments and surveys, and problem solving (Easton, Roberts, and Tiao 1988; Hogg 1991, 1992; Moore 1992). They also recommend new formats for class activities that rely on writing, projects, and experimentation (Fillebrown 1994, Giraud 1997, Roberts 1992).

2 A number of interesting class activities can benefit from the Internet. Perhaps the best known Internet resources are searchable datasets and online case studies (http://lib.stat.cmu.edu/DASL/DataArchive.html, http://www.stat.ucla.edu/cases/), online statistical software (http://fonsg3.let.uva.nl:8001/Service/Statistics.html, http://www.stat.ucla.edu/calculators/cdf/), and online textbooks (Dear 1995, Lane 1993, Stockburger 1996). A sampling activity described in Schwarz (1997) uses the Internet in a novel way that would be difficult to implement without a computer.

3 This note presents a short hands-on project that relies on interactive capabilities of the Internet. The World Wide Web format assures access to the project from all operating systems. It also allows the instructor to assign different versions of the project among the students, or groups of students. With Web-based computer support this is actually a short project, not much more than a homework assignment.

4 This project introduces students to a simple decoding problem. Students are asked to decipher an encrypted text provided by the instructor. The task is straightforward to explain, yet at first seems inapproachable. The decoding of the encrypted text is based on comparisons between the frequencies of characters in the encrypted text and the frequencies of characters in a plain text of the student's choice, or in several such texts. The

statistical tools used in the project are summary statistics and sampling. To complete the project, students must also incorporate critical thinking to make use of all available information.

5 The project forces the students to face the fact that sample data vary and different texts can give different percentages for the same character. Students realize that by analyzing multiple texts they can estimate the range of values attained by the sample proportion of a character. They learn to use approximate percentages of characters. They also learn that while statistics provides a good starting point, intuition, critical thinking, and even lucky guesses are important.

2. Decoding a Substitution Code

6 The encoded text consists of characters: letters, spaces, punctuation marks, numbers, and other printable keyboard symbols, which are encoded by a *simple substitution code*. This means that every character in the original text is replaced by some other character, its substitution code. The term "simple" refers to the fact that a single substitution rule is used throughout the encoded text. Capital letters might be coded by a different character than the corresponding lower case letters, and all punctuation characters and spaces are encoded too.

7 The basic method for decoding a substitution code is to compare the frequencies of characters in the encoded text and the frequencies of characters in a text written in plain English. The characters that occur most frequently in the coded text correspond to the most frequent characters in plain English. Similarly, the frequencies of the combinations of characters (pairs, triplets) within the coded text should match the frequencies of corresponding combinations of characters in any sufficiently long text.

8 The main statistical step is to create tables of character frequencies that will be used to compare the coded text to a plain English text. To facilitate statistical comparison, the encrypted texts should be quite long, although the objective of the project is to decode only the first sentence. The cumbersome task of counting the characters is automated by the online character counter. The character counter works best with texts in electronic formats that can be copied and pasted into the WWW-based input form. The output of the program is a printable table of the frequencies of characters, including punctuation marks and other printable characters in the text.

9 Comparing the proportions of characters is only a good starting point. Students quickly discover the meaning of a couple of the characters in the cipher. The first character to crack is the space that separates the words; the proportion of spaces in English is substantially higher than the proportions of other characters. Next, the letter "e" is easily identified in most texts. Typically the direct comparison method breaks down at this point and should be stopped because other characters cannot be reliably matched solely on the basis of their frequencies.

10 The next step is to decipher more characters by trying to decode short and frequent words, like "a," "the," "and," and "is." Statistics helps to choose among the alternatives when a word can be decoded in more than one way. For example, we might encounter an encoded three-letter word that we decode as "ha_", with just one letter missing. This word could be completed as "has," "had," or "ham" ["hat" can be discarded assuming we already know the code for "t" from deciphering the code for "the"]. Since the proportions of letters "m," "s," and "d" are spaced far apart, we can make a reasonably certain choice by inspecting the table of frequencies of characters in the coded text. Our choice can also be confirmed by comparing the frequencies of pairs of letters "am," "ad," and "as" with the frequency of the coded combination in question. Similarly, three-character combinations can be inspected. Every newly deciphered word increases the number of deciphered characters in the encoded text and serves as an extra verification, or falsification, of the previously established matches.

3. Experiences With the Decoding Project

11 This project was used in Elementary Probability and Statistics at the University of Cincinnati. This is a service course taken by students to fulfill a college math requirement. Initially, the project was assigned late in the course to a small section of 23 students; by this time these students had already been exposed to statistical

reasoning, including confidence intervals and hypothesis testing. Since the project does not rely on formal statistical procedures, it was also assigned in another section of 45 students early in the course.

12 Before the project was assigned, students were introduced to sample variability and the use of online software. As a joint lab activity, students used the character counter to collect data on the percentages of the letters "e" and "t" in various texts. Collected data were recorded directly in the class announcement, and then used to make a histogram and a stem-and-leaf plot, to compute the five-number summary of the data, and to discuss outliers. In my opinion, a similar exercise should precede the project.

13 Prior to the actual decoding project, students should be exposed to sampling variability by analyzing several texts from the Web. This is also an appropriate moment to discuss the issue of random sampling. Students easily concur that selecting unusual texts, like specialized technical papers, is not the best strategy if the text to decode is suspected to come from a textbook or a novel.

14 The encoded texts for the project were assigned to students via Web links from the class list. Sample project instructions are available online. The task of preparing encrypted texts was simplified by the encryption function which is available as part of the character counter. Its output is a ready-to-save Web page. Here are samples of the encoded texts produced by the program: text1 text2 text3 text4.

15 To make tables of comparative frequencies, students were asked to select their own samples of text. The project description and class announcement provided links to online textbooks and the Gallup poll site as suggested sources. Most students followed the suggestion and picked samples of moderate size (several thousand characters). Some students chose very small sample texts from other sources, like the actual text of the project instructions. Some students analyzed huge pieces of text, hoping to estimate the proportions more accurately.

16 Confronted with an unmanageable (at first sight) jungle of symbols, students learned to appreciate the very first insight that came from comparing proportions of characters: they could immediately divide the continuous string of encrypted text into separate words. They could often recognize a couple more characters in a similar fashion, sometimes because of numerical coincidences in the frequencies. Some students commented on the fact that often such initial matches had to be discarded because they wouldn't form sensible words in the encrypted text. Then the students were confronted with the fact that most of the frequencies in the encoded text did not have any clear match, and it became harder to recognize which code should match which character. Inexact matching and trying to choose letters to make sensible short words was a trial and error process that used both common sense and occasional insights.

17 An effective strategy, noticed in students' solutions in the first, more experienced class, became part of the project instructions for the second, less experienced class. The strategy was to use the output of the character counter to produce a set of potential matches for each character. This device helps the students to avoid fixing the matches too early and to visualize their options.

18 Most students had no difficulty deciphering the first sentence, and many went on to decode more than was required. A few good students failed to complete the project because of lack of flexibility -- their attempts to match the characters rigidly by the order of frequencies led them into a blind alley. Some mathematically weaker students found the project very easy and asked for more such projects so that they might improve their grades.

19 None of the students used the frequencies of pairs or triplets. Two -- perhaps complementary -- explanations are that students may have found this information too complex to take into account, and that they were able to complete the project using only single-character frequencies and their ability to fill in the missing letters in the text.

4. Character Counter Program

20 The character counter program used in this project is a short (about 300 lines of code) Common Gateway Interface (CGI) script written in Perl 5.0. The script is invoked from any Web browser that supports forms. The script is executed on the server machine and thus is entirely independent of the computer system used by the students. The script can be installed on any system that supports Perl 5.0 and the CGI.pm library. At the time of this writing, the script can be moved to any Win32 or Unix platform with the minimum version of Perl and corresponding library. The script requires no compilation and can be modified to support other projects by editing its file and reformatting the output. Instructors do not have to install their own copy of the script. They can point their students to the common access page, like the one linked to in the sample project and sample encoded texts (text1 text2 text3 text4). Access pages to the same script at a single location can also be personalized by instructors and placed anywhere on the Web.

21 Some aspects of the character counter script are not entirely intuitive and might need clarification.

- The script counts the lower case characters separately from the upper case characters. For example, lower case "t" might be reported as about 6% of the text, while upper case "T" as only .8%. This is consistent with the encryption scheme used in the project where the upper case letters are encrypted by different characters than the lower case letters.

- The character counter treats white space (blanks) differently than the rest of the characters. Invisible white space is actually made of several different computer characters: tab, space, new line (sometimes called end-of-line), and "enter." These are all lumped together, except for the "new line" which is not counted at all. Consecutive occurrences of white space, like two consecutive spaces, or a tab and a space, are counted as a single occurrence.

- The access page provides an option to count single characters, pairs, or triplets of characters. If more than one option is selected, then groups of various sizes (singles, pairs, triplets) are all counted together. Separate tables of frequencies for pairs and triplets are easily produced by activating one option at a time or by using a single-option access page.

- To avoid long meaningless printouts, the character counter prints the counts for the first 30 characters. It will print additional counts for up to 100 character combinations that might occur in a similar text. A normal approximation is used to approximate the probability of occurrence and to calculate the "margin of error" for each percentage reported.

Acknowledgments

The sample encoded texts linked to in this note come from introductory texts in statistics (Dear 1995, Lane 1993, Stockburger 1996). I would like to thank the anonymous referees of this paper who provided excellent suggestions which significantly improved the decoding project and helped to clarify its description.

References

Dear, K. (1995), SurfStat, Online textbook. (http://surfstat.newcastle.edu.au/surfstat/main/surfstat.html) See also *Maths&Stats; newsletter*, November 1997.

Easton, G., Roberts, H. V., and Tiao, G. C. (1988), "Making Statistics More Effective in Schools of Business," *Journal of Business and Economic Statistics*, 6, 247-260.

Fillebrown, S. (1994), "Using Projects in an Elementary Statistics Course for Non-Science Majors," *Journal of Statistics Education*, [Online], 2(2). (http://www.amstat.org/publications/jse/v2n2/fillebrown.html)

Giraud, G. (1997), "Cooperative Learning and Statistics Instruction," *Journal of Statistics Education*, [Online], 5(3). (http://www.amstat.org/publications/jse/v5n3/giraud.html)

Hogg, R. V. (1991), "Statistical Education: Improvements Are Badly Needed," *The American Statistician*, 45, 342-343.

----- (1992), "Report of Workshop on Statistical Education," *Heeding the Call for Change*, ed. Lynn Steen, MAA Notes No. 22, Washington: Mathematical Association of American, pp. 34-43.

Lane, D. (1993), HyperStat, Online textbook. (http://www.ruf.rice.edu/~lane/hyperstat/contents.html)

Moore, D. S. (1992), "Teaching Statistics as a Respectable Subject," in *Statistics for the Twenty-First Century*, eds. Florence Gordon and Sheldon Gordon, MAA Notes No. 26, Washington: Mathematical Association of America, pp. 14-25.

Roberts, H. V. (1992), "Student-Conducted Projects in Introductory Statistics Courses," in *Statistics for the Twenty-First Century*, eds. Florence Gordon and Sheldon Gordon, MAA Notes No. 26, Washington, DC: Mathematical Association of America, pp. 109-121.

Schwarz, C. J. (1997), "StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling," *Journal of Statistics Education*", [Online], 5(2). (http://www.amstat.org/publications/jse/v5n2/schwarz.html)

Stockburger, D. (1996), Introductory Statistics, Online textbook. (http://www.psychstat.smsu.edu/sbk00.htm)

Wlodzimierz Bryc
Department of Mathematical Sciences
University of Cincinnati
PO Box 210025
Cincinnati, OH 45221-0025

brycw@math.uc.edu
