Using Large-Scale Datasets to Teach Abstract Statistical Concepts: Sampling Distribution

Gibbs Y. Kanyongo, PhD

Associate Professor of Educational Statistics and Research

Duquesne University, Pittsburgh, PA 15282, USA

Abstract

With the advancement in computer technology, more and more statistics instructors now rely on simulations, web-based statistical applets and other artificially-generated datasets to teach abstract statistical concepts. While this approach may be useful, this paper shows how instructors can use large-scale datasets to make these concepts more real for students thereby facilitating their understanding of the concepts. The paper uses the Education Longitudinal Study of 2002 (ELS: 2002), a large-scale real dataset to demonstrate the concept of sampling distribution and standard errors. The paper focuses on the distribution of sample means and sample standard deviations.

Introduction

Statistics is used in almost our everyday lives and it has applications in a wide variety of settings, for example in weather reports, in business, in crime reports, in finance, and in education. This wide application of statistics makes it essential for students' to have a good understanding of statistical concepts and become statistically literate. However, students often do not grasp certain statistical concepts due to the abstract nature of the subject matter. According to the Guidelines for Assessment and Instruction in Statistics Education (GAISE), an introductory statistics course should: (1) promote statistical literacy and statistical thinking; (2) use real data; (3) promote conceptual understanding; (4) foster active learning; (5) use technology for developing conceptual understanding and analyzing data; (6) use assessments to improve and evaluate student learning. This paper will focus on the second guideline, the use of real data in teaching statistics.

One of the ways to enhance the understanding of statistical concepts, one needs to run real experiments that generate reliable data (Akram, Siddiqui, & Yasmeen, 2004). In most cases, the main limitation for running real experiments is the lack of reliable real data. Even if in cases where real data may be available, the challenge then becomes how to effectively integrate those data in a statistics curriculum. It is difficult to run real experiments during the teaching period in the university. Because of the challenges in using real datasets in teaching statistics, one option that is commonly used by statistics instructors is the use of artificial data in conducting statistical experiments. Statisticians developed simple and very economical experiments, which can be performed in the class by the students (Akram, Siddiqui, & Yasmeen, 2004). A number of studies have been conducted over the years that demonstrate improved students' learning through the use of statistical experiments and simulations. For example, Larwin & Larwin (2010) conducted an experimental study in which they examined the effect of teaching undergraduate business statistics students using random distribution and bootstrapping simulations. Their results indicate that students in the experimental group-where random distribution and bootstrapping simulations were used to reinforce learning demonstrated significantly greater

gains in learning as indicated by both gain scores on the Assessment of Statistical Inference and Reasoning Ability and final course grade point averages, relative to students in the control group.

In another study, Raffle & Brooks (2005) reported results of the effectiveness of a Monte Carlo simulation method in teaching a graduate-level statistics course. In their paper, they reported how computer software (MC4G, Brooks 2004) can be used to teach the concepts of violations of assumptions, inflated Type I error rates, and robustness in an introductory graduate course statistics course.

While simulations and simulated data are valuable tools in teaching statistical concepts, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) calls for the use of real datasets to teach these concepts. In line with this guideline, there are several studies that report results for using real-life situations or examples in teaching statistical concepts. In one such study, Leech (2008) used the game of poker and small group activities to teach basic statistical concepts. He noted that this approach helped to reinforce the understanding of basic statistical concepts and helped students who had high anxiety and it made learning these concepts more interesting and fun. In another study, Connor (2003) illustrated statistical concepts using students' bodies and the physical space in the classroom. The concepts covered included: central tendency, variability, correlation, and regression are among those illustrated. In this study, these exercises encouraged both active learning and the visual-spatial representation of data and quantitative relations. Connor (2003) reported that students evaluated the exercises as both interesting and useful. Makar & Rubin (2009) presented a framework for students to think about informal statistical inferential reasoning covering three key principles of informal inference--generalizations "beyond the data," probabilistic language, and data as evidence. In their study, they used primary school classroom episodes and excerpts of interviews with the teachers to illustrate the framework and reiterate the importance of embedding statistical learning within the context of statistical inquiry.

In their paper, Schumm, Webb, Castelo, Akagi, Jensen, Ditto, Spencer Carver, & Brown (2002) use of historical events as *examples* for *teaching* college level *statistics* courses. They focused on *examples* of the space shuttle Challenger, Pearl Harbor (Hawaii), and the RMS Titanic. Finds *real life examples* can bridge a link to short term experiential learning and provide a means for long term understanding of *statistics*. Stork (2003) describes a pedagogy project that keeps students interest. He used student survey to gather data on students' university experience and demographics. The class used the Statistical Package for the Social Sciences (SPSS) data set for demonstration of a range of statistical techniques. The student survey instrument and data set provided examples and problems for each of the major topics of the course. He concluded that student comments and performance demonstrated the project's positive results.

In most cases when talking about sampling distribution, most students think this only applies to the distribution of sample means. The main reason for this is that most textbooks explain this concept using the distribution of sample means. Yet, in actual fact sampling distribution applies to any sample statistic that we calculate to estimate population parameters. It is important for students to understand that when talking about sampling distribution, this applies to standard deviation, variance, correlation coefficient, and many other statistics that we can calculate. What is also important for students to understand is that the sampling distribution is a model of a distribution of scores, just like the population distribution, except that the scores are not raw scores but statistics. The resulting distribution of statistics is called the sampling distribution of that statistic. For example, if standard deviation is the statistic of interest, the

distribution is known as the sampling distribution of standard deviation. In this paper, sampling distribution is illustrated for sample means and standard deviations. While this may not be new, what's unique is that a large-scale real dataset is used do illustrate these concepts.

Method

Data used in this paper came from the Education Longitudinal Study of 2002 (ELS: 2002). This survey was designed to monitor the transition of a national sample of young people as they progress from tenth grade through high school and on to postsecondary education and/or the world of work (National Center for Education Statistics 2008). During the 2002 base year, students were measured on achievement, and information was also obtained about their attitudes and experiences. These same students were surveyed and tested again, two years later in 2004 to measure their achievement gains in mathematics, as well as changes in their status, such as transfer to another high school, early completion of high school, or leaving high school before graduation, and also in 2006. This cohort will be interviewed again in 2012 to measure their transition into the job market (National Center for Education Statistics 2008). For the purpose of this paper, only data collected during the base year (2002) was be used.

Procedure

One of the several variables contained in ELS: 2002 is a measure of mathematics achievement of the cohort of 2002 10^{th} graders using a standardized mathematics score (MathScore). This variable was selected for use in this paper. The entire ELS: 2002 dataset contains 10, 094 cases, and this represents the population (N=10,094) for the purpose of this paper. Initially, the population distribution of MathScore was obtained, and this is displayed using a histogram (μ= 51.40, σ = 10.094) in Figure 1. Next, using an SPSS syntax, 100 random samples each of size 100 were drawn from the population and an SPSS dataset with n=100 was created with mean and standard deviation for each sample produced, resulting in 100 means and 100 standard deviations. This was repeated for 500 samples, and 1000 samples, and in each case, the size of each sample drawn was 100 generating means and standard deviations.

Results

The population distribution is shown in Figure 1 while the sampling distribution of the means and standard deviations for the samples on n=100, n=500, and n=1000 are presented in Figures 2 to 4. Population parameters, expected values and standard errors are presented in Tables 1 below. The mean of the sample means in these distributions is the expected value of the sampling distribution of the mean, and the mean of the sampling distribution of the standard deviation is the expected value of the sampling distribution of the standard deviation. Both of these parameters are estimators of the corresponding parameters. When the expected value of a statistic equals a population parameter, the statistic is called an unbiased estimator of that parameter. Students should note that the standard deviation of the sampling distribution of the mean is called the *standard error of the mean*. Similarly, the standard deviation of the sampling distribution of the standard deviation is called the *standard error of the standard deviation*. Students will have the opportunity to learn that sample statistics are most likely to be different from the population parameters due to random sampling errors. However, as sample size

increases the sample statistics approach the population parameters, thereby reducing the discrepancy between the statistic and the parameter. This discrepancy is known as the sampling error, and students will be able to see that large samples yield small sampling error. In addition, the histograms show that sampling distribution of the mean will approach a normal distribution as sample size increases.

Figure 1. Population Distribution

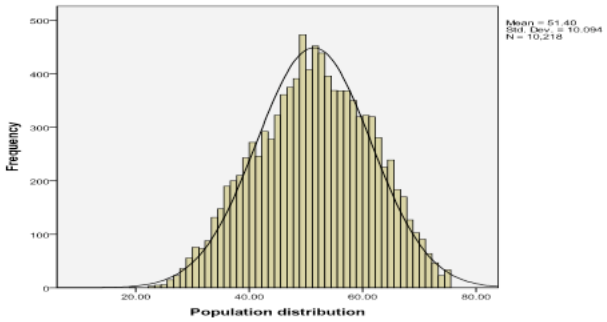


Figure 2a. Distribution of Sample Means (n=100)

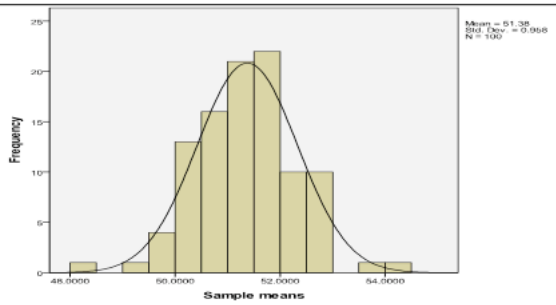


Figure 2b. Distribution of Sample Standard Deviations (n=100)

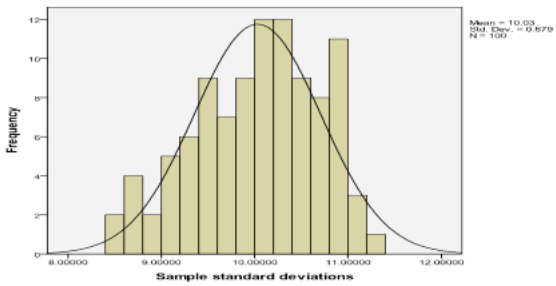


Figure 3a. Distribution of Sample Means (n=500)

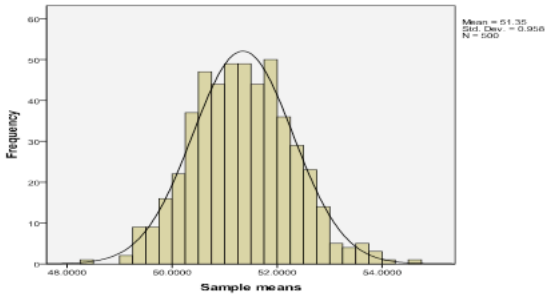


Figure 3b. Distribution of Sample Standard Deviations (n=500)

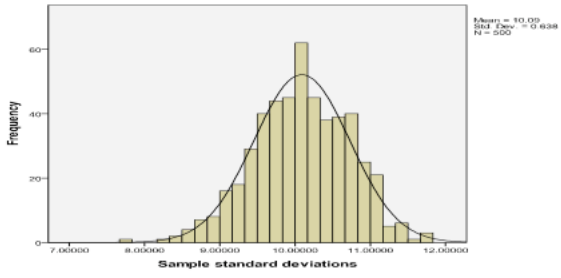
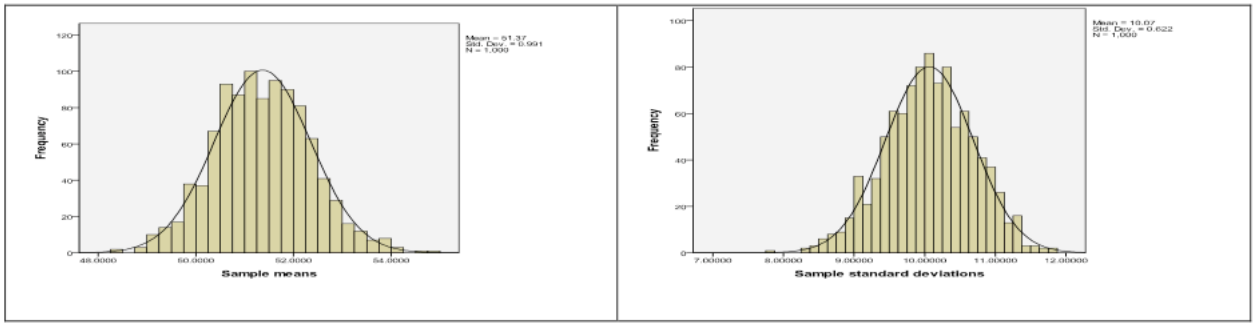


Figure 4a. Distribution of Sample Means (n=1000)

Figure 4b. Distribution of Sample Standard Deviations (n=1000)



Conclusion

In this paper, I have demonstrated the use of a real dataset to teach sampling distribution. While artificial data is usually used to demonstrate these concepts, this paper showed that it's possible to use real data to achieve the same objective with the added advantages that real data bring to students' learning experiences. The dataset ELS: 2002 contains thousands of variables, and students will be able to select variable(s) of their choice, that make sense to them. The fact that ELS: 2002 and other large-scale datasets are publicly available should be an incentive for statistics instructors to use them as teaching tools.

References

Akram, M., Siddiqui, A. J., and Yasmeen, F. (2004). Learning statistical concepts. *International Journal of Mathematical Education in Science and Technology, 35*(1), 65-72.

Brooks, G. P. (2004). MC4G: Monte Carlo Analyses for up to 4 Groups [Computer software and manuals].

Connor, J. M. (2003). Making statistics come alive: Using space and students' bodies to illustrate statistical concepts. *Teaching of Psychology, 30*(2), 141-143.

Larwin, K. H. and Larwin, D. A. (2010). Evaluating the use of random distribution theory to introduce statistical inference concepts to business students. *Journal of Education forBusiness, 86*(1), 1-9.

Leech, N. L. (2008). Statistics poker: Reinforcing basic statistical concepts. *Teaching Statistics:An International Journal for Teachers, 30*(1), 26-28.

Makar, K. and Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105.

National Center for Education Statistics [cited November 11, 2010]. Education Longitudinal Study of 2002[Online]. Available at http://nces.ed.gov/surveys/els2002/

Raffle, H. and Brooks, G. P. (2005). Using Monte Carlo software to teach abstract statistical concepts: A case study. *Teaching of Psychology, 32*(3), 193-195.

Schumm, W. R., Webb,F. J., Castelo,C. S., Akagi, C. G, Jensen, E. J., Ditto, R. M.,Spencer, C.E., and Brown, B. F. (2002). Enhancing learning in statistics through the use of concrete historical examples: The Space Shuttle Challenger, Pearl Harbor, and the RMS Titanic. *Teaching Sociology, 30*(3), 361-375.

Stork, D. (2003). Teaching statistics with student survey data: A pedagogical innovation in support of student learning. *Journal of Education for Business, 78*(6), 335-