Journal of Mathematics and Statistics: 4(3): 168-173, 2008
ISSN: 1549-3644

Using the Spreadsheet to Understand Random Sampling Procedures in Relation to the Central Limit Theorem

Ramzi Naim Nasser
Center of Educational Development and Research, College of Education,
Qatar University, POBOX 2713 Doha, Qatar

Abstract: In many statistics courses in colleges and universities, random sampling has not been presented to students through a hands-on and practical approach. This research described a real-life method to random sampling was used as an exercise in the statistics class. The method allowed students to recursively select 5, 10, 20, 30, 40, 50, 60 and 70 samples and list these samples in a matrix-spreadsheet. Students used the spreadsheet software-EXCEL, calculated the statistics and performed graphical computations. Students were engaged to reflect and develop a sense of meaning of randomness in its relation to the law of large numbers. Specifically, the method explored an increase in the number of samples, how it changed the means, mean of means or standard deviations- This is an instructional study which devices a working approach to random sampling.

Key words: Central limit theorem, law of large numbers, random sampling, resampling, mean, mean of means, shrinking means

INTRODUCTION

In various science and engineering academic degree programs, students take probability and statistics course(s) as a requirement. This course(s) introduces students to probability and normal distributions with substantial problem solving activities. The course(s) is usually crammed with probability, descriptive and inferential statistics. Teaching statistics is not simple and straight forward, it is very difficult for instructors to demonstrate in real-time, concepts that have abstract and extensive computational requirements. When such concepts are treated theoretically in statistics, it may not be all that meaningful for the student. As stated by[1], students in problem-solving situations can plug-and-chug but cannot make a meaningful interpretation of the calculations. They may follow a description or an example and then try to study out a general formula. Once they make the calculation, generalization about the meaning of the process is not interpretable. However, in non-procedural problems, heuristic approaches, have specific steps that cannot be identified by a set of rules or algorithms. Statistical heuristics are not easily adapted to spreadsheet applications. Often heuristics are complex and interconnected with layers of abstract concepts; when applying a heuristic, the spreadsheet utility works well to provide the computational requirements needed to solve abstract non-exercise based problems.

The spreadsheet has been used extensively in engineering education as in course grading of numerical solutions, differential equations and network analysis[2,3]. Rarely has the spreadsheet been used to demonstrate statistical applications for science and engineering students. The spreadsheet is an excellent tool for problem solving and records keeping (listing of problem-solving steps). In this study, the use of random sampling heuristic demonstrates the central limit theorem. The exercise presented could allow statistics teachers to show the effects of random sampling on sampling distribution statistics using the spreadsheet.

There are spreadsheets heuristics run through applets, which students interact with, by changing variables, inputting unknowns and then observing the behavior of the data, results, or graphical output. Several business statistical texts have supplementary material that consider and explore random number generators[4,5]. Statistical applets used the spreadsheet applications and run on EXCEL, they include random number generators and a number of other statistical applications. These applets are interactive such that by increasing one of the relational variables, changes are seen dynamically. Specifically, the central limit applet compares a normal curve with an exponential distribution. By changing n (sample size) of the exponential distribution, the user can graphically observe the distribution and convergence to the normal. If these supplementary applets are integrated within the heuristic scheme for problem-based learning, they become excellent educational material. However, these applets are non-enabled exercises and do not engage the student in the learning process.

The spreadsheet can be used to simulate problems, by setting up equations, changing variables, co-variations, extensions, recursions, iterations, modeling, adding data with results on one worksheet or template[2]. The spreadsheet is very useful when faced with the tedium of study and pencil classroom approaches.

Specifically, the use of the spreadsheet to model random sampling techniques can not be done with functions or macros, especially that heuristics can not be easily translated into generic functions. Change within model parameters, data selection, or interconnections of layers of consecutive generalizations, produce possibilities for higher levels of mathematical thinking[6]. Simulation models also contrive lists of nominal events; a case in point would be the entry of random numbers from tables rather than numbers generated from spreadsheet functions or applets.

The advantages in the use of automated spreadsheet programs in college statistics are numerous. The primary advantage is that, it enables the student to focus on the statistical concepts by implementing computational models without the recourse to program writing that exploit macros or Visual Basic[7,8]. In addition, the spreadsheet program is constructivist; from a model, it allows the user to build miniature--conceptual structures, meaningful matrices for complex calculations[9]. The spreadsheet software, as earlier mentioned, is also user friendly allowing the student to experience the how, the result and its presentation through various graphical and tabular outputs[10,11].

Law of large numbers and sampling concepts with the spreadsheet: The law of large numbers says that if a random number is drawn from a population the larger the sample, the closer the mean of the sample to the mean of the population[12]. The larger the sample is, the increased resemblance of the sample to the population. This particular concept is often hard to expose to students in statistic classes. But when students experience the process of random sampling and then observe the behavior of these numbers through reiterative sampling procedures, it helps in developing a meaning to random sampling theory. Sampling of data helps the student experience the impact of random number selection through observing the behavior of selected numbers through its distribution. Software tools as the graphical calculator or even EXCEL can generate random data; however, a concrete approach to random sampling helps students to make connection between the activity and theoretical precepts.

Sampling inferences are rarely connected to hands-on, practical and real applications in the statistics classroom. Particularly, the generalization and development of the random selection concept in connection to sample size is seldom made into an exercise but a passing mention in university-level statistics courses. The connection between sampling, i.e., variability, law of large numbers and the central limit theorem are generally difficult to teach because of settings in the classroom (time and tools)[13,14]. Exploratory ways of teaching problem solving requires time and computational power; the spreadsheet as a stand-for-all schema has the capacity for calculating multi-layered datasets found in tedious statistical exercises or even every complicated problem.

With the assumption that descriptive and theoretical discussion in the classroom has little practical and authentic value in understanding the random sampling concept, however, a hands-on approach through the use of the spreadsheet would seem to have educational advantages. Thus, this study presents a teaching module where students discover the actual process of random sampling.

Traditionally, random instances in probability theory are illustrated through coin flip examples or the selection of objects from an urn. The occurrence of any possible outcome signifies the cognition about the nature of the events. The representative heuristic[15], perhaps, more easily expressed by the example of the birth of Boys (B) or Girls (G). The outcomes of event; BBBGGG versus the GBBGBG, concludes that the latter has a higher probability of occurrence because of its representative feature-appearances as a random type. A process approach to random sampling helps channel the concept of the law of large numbers in a different sense than what appears as a representative feature in sampling[15].

The resampling technique allows the learner to vary sample size and number of samples to experience the random selection procedure. A change in one or more dimension in the resampling procedure consequently would alternate the behavior of the distribution. So by altering one or more variables the spreadsheet allows for changes in preserving the primary or selected datasets.

Simulation modeling has then important and appropriate learning conditions for exploration and discovery[6]. Thus, this classroom exercise study will stress on the role of modeling the operational schema. The model contrives the generation of sample means, ranges and standard deviations with the sense that there

are two distributions one being, a random variable distribution and the other being the sampling distribution.

The purpose of the exercise: The purpose of this exercise was to draw students to observe and explore a re-sampling procedure as a framework for understanding random sampling concept in its relation to the law of large numbers. The author of this study in a naturalistic approach observes and records student-learning experience with the spreadsheet for problem-based learning. The exercise allows students to simulate a number of samples and re-samples to show how numbers behave under random effects.

The next section lays a formal description of the proposed procedure, followed by the exercise section. The exercise includes a description of a selection procedure and the calculation of statistical values (mean, range and standard deviation followed by a graphical output).

MATERIALS AND METHODS

Use of spreadsheets in statistics instruction: The spreadsheet applications and software has important functions in the teaching and learning of statistics[1]. The spreadsheet software use of matrixes, crossed attribute tables, repetitive calculation and layers of repetitive functions can handle complex calculations[2], large datasets, carry elaborate calculations, produce dynamic representation, statistical functions and have the results with the input data all appearing in one structure[1,2].

The most popular and available spreadsheet software is the EXCEL program. The software is inexpensive, run on machines with modest specification and widely used by the academic community. EXCEL has a variety of data objects, including texts, numbers, graphical utility, logical and computational formulas[7,8].

A statistics class at a private university was opportunistically selected to for the study. The class was carried in an "intelligent" classroom where the students had networked computers in which they communicated with the instructor through the Blackboard platform. Students worked individually and were give prior instruction in the use of the Blackboard and EXCEL. Each student was given the random number Table and handed self-explanatory guidelines to select numbers from a random numbers table.

Exercise: A sample of 22 students participated in this study taking an introductory probability and statistics course. Numbers selected from the random numbers table were considered random such that the distributions will be as unbiased as to being very close to being homogeneously random. For each of the exercises assigned, students were asked to submit their study through Blackboard.

Re-sampling procedure: In the first exercise, each participant is given instruction to choose a sample of n = 5 observations and k = 5 samples. The selection of numbers comes from the random number table. The numbers listed in the random number table are six digit numbers. Participants are instructed to select the first 3 digit of the 6 digit number, read from the right-hand side of the random numbers table, such that the number could start with a 0 and range to 999. The procedure and exercise goes as follows:

- From the Table of Random Numbers, participants draw k = 5 random samples of size n = 5. Participants obtain k = 5 samples each of size n = 5 resulting in n x k = 25 numbers which is then placed in a 5×5 spreadsheet matrix. Participants obtain a sample mean and standard deviation for each sample n, resulting in 5 means and 5 standard deviations for the k = 5 samples. In addition, the range for each sample and range for sample means is calculated. Each new sample is combined in a single column in the spreadsheet structure. Thus, there will be samples of size n and the number of samples of size k. Each participant will be asked to produce an nxk randomly selected numbers. Participants are asked to submit the 25 randomly selected numbers of the 5 samples to the data manager
- The spreadsheet is used to calculate the statistics for each random sample (mean and standard deviation). The resampling procedure allows for the variation of k and n. Each observation in sample k is chosen at random. The data manager (investigator) chooses n = 5 elements for each of the k samples. The selection of an element in a sample n may occur more than one time because of the random selection
- Once participants finish the selection of numbers, they send the data to the investigator (data manager). The investigator ends up with 22×25 = 550 numbers from all the participants. In the second class exercise the investigator as a data manager reorganizes the 25 numbers from each of the 22 participants. The data manager resubmits 22×25 numbers back to participants. Thus, each student will have 22×25 numbers. Participants are then asked to compare the grand mean of the 550 numbers to the mean of the k = 5 of their data.

Participants are also asked to compare the standard deviation, range between sample means and the grand mean. The data manager organizes the numbers in a random fashion onto one worksheet and sends the class data back to participants

- Participants obtain means and standard deviations by increasing number of samples to k = 5, 10, 20, 30, 40, 50, 60 and 70 samples of size n = 5, for example, the k = 20 samples would have 20×5 elements, then participants graph on the x-axis, the sample size and the means on the y-axis. In this approach participants discover the act of random selection in relation to the law of large numbers and central limit theorem

RESULTS

Observations from the classroom: Figure 1, shows the 25 random numbers selected by one of the students in the study, in the eighth row, a mean of the five numbers and a mean of the means is calculated in the H-Column (Fig. 1). The re-sampling procedure is run five times, with calculations made for each sample. In the next three rows, a minimum and maximum function is used and then the absolute value difference between them is calculated. The range for each sample and range between the means of the samples is calculated. A second sheet was created for 550 random numbers organized by the data manager and send to each participant in the class. Participants are asked to calculate the grand mean, range and standard deviation and then compare to values of the sample that each student obtained. They are then asked to write down their observations.

In the third exercise participants repeat exercise 1 with n = 5 elements and different sized samples of k = 5, 10, 20, 30, 40, 50, 60 and 70. The mean, range and standard deviations are calculated for each of the k = 5, 10, 20, 30, 40, 50, 60 and 70 samples. Participants then draw the histogram for the means crossed with the different sized samples (Fig. 2).

Observation 1: Participants recognize that the range between each sample of n = 5 was greater than the range of sample means. Participants are then asked to explore why the range in the sample was less than the range of the means. Several participants recognized that the means of the five samples were close and the value of the means was shrinking to the middle of 0-999.

Observation II: Participants compared sample means to the mean of means and to the grand mean respectively. Invariably, all participants recognized that when n was increasing, the means were shrinking and getting very close to the middle point. Thus, as k increased, the central measures got very close to the population mean. Participants realized that the ranges where shrinking and means where getting close to each other by increasing n. In addition, participants observed, that the values of numbers were getting closer to the middle (between 0-999) for the three digit numbers. They observed that the distribution of sample means had a lower range. Participants were very surprised by the results and where instigated to explore these issues individually without any direction. It was emphasized that sampling distribution would eventually approximate the normal distribution.

Participants obtained the statistics of means, standard deviations and ranges. To determine whether the standard deviation decreased as n increased, 40% of the participants who were probed, realized that the behavior of the mean values for each k=5, 10, 20, 30, 40, 50, 60 and 70 sample did not lend itself to the theoretical understanding of the law of large numbers.

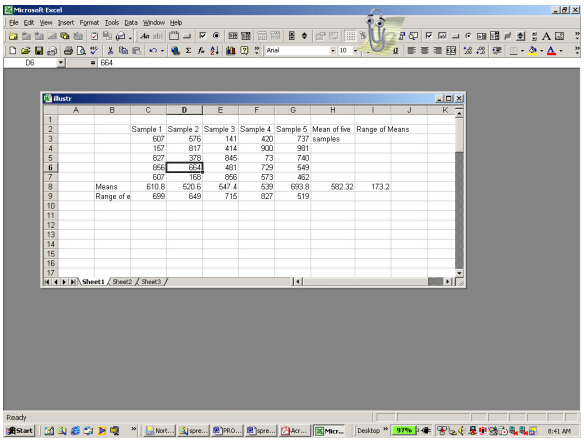

Fig. 1: Sample of the 5×5 matrix and calculations of the means

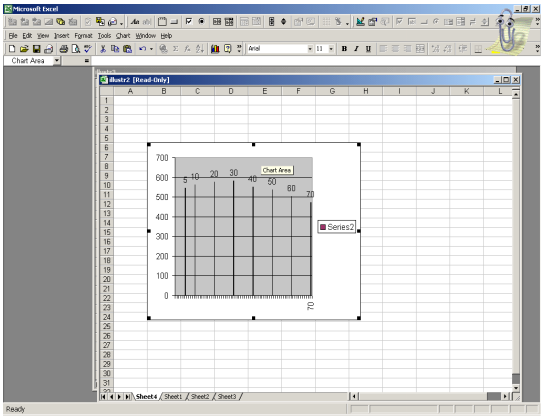

Fig. 2: Sampling distribution graph showing

When the participants later graphed the values of the means for n=5, 10, 20, 30, 40, 50, 60 and 70, they realized that the means were getting very close to the middle value (between 0 and 999). The purpose of this exercise was to have participants go through active sampling so they understand constructively the concept of random sampling procedure and thus be able to generalize from action to theory and make the association between the empirical evidence resulting from the statistics. Particularly, participants were successful at making the appropriate generalization by deducing the shrinking mean for larger n and mean of means.

DISCUSSION

Participants selected a 3-digit number from a random numbers table, they then simulated the resampling procedure using different sized samples. A comparison made between the central tendency statistics with the hypothetical measures of the population helped students observe larger sample means convergence onto middle values. The method devised in this exercise showed the relation between sample size to the population and law of large numbers through a selected set of random numbers to central limit theorem.

The graphing utility of the spreadsheet was helpful in providing real-time presentation for students to observe the behavior of numbers as well as sample means. The most notable observation made by participants was the shrinking of the mean and the mean of means, as well as the shrinking range of means. A large number of the participants realized that the mean approached the middle point between 0 and 999. Some students remarked about the behavior of the standard deviation. Those who did not see that the standard deviation shrinking, observed it was getting larger and where instructed to take larger samples to observe the changing behavior.

Perhaps the most interesting finding was participant's inability to determine the behavior of the standard deviation through sampling, re-sampling, statistical calculations and recalculation. When asked whether the standard deviation decreased or increased as n increased, students where prone to refer to the analysis of data and the results. Some students where asked to take 20 samples by increasing n at each turn, this approach allowed them to merge the theoretical precepts with the results. Some students discovered that through the procedure they became aware of random data characteristics.

University students were constructively involved in the process and interpretation, they appeared to understand that the averages of large samples was more likely to resemble the population mean (middle value between 0 to 999) as the means were getting closer to samples of larger n. The challenge in the heuristics steps was to simulate students to think about the questions and the theory and to interact with one another in the lecture.

As one participant said:

- I feel that the pattern recognition was so interesting, I never thought statistics is actually about the behavior of numbers. When a problem had a definite answer, once it is found the problem is closed. I feel that in the group study, we all shared in generalizing about these concepts. It would have been very hard if we considered these concepts on individual basis or tried to work them out individually

CONCLUSION

Overall, the study-helped students make distinctions between distributions of averages for large and small samples and distribution of scores of random numbers. In the model students were asked to vary the sample size and number of samples. Students then observed the differences between sampling averages and sample scores. The use of the spreadsheet came in handy by changing sample sizes and number of samples and to work through the calculations and compare to statistics as the mean and the mean of means. Students could make the appropriate generalization and development of the central limit theorem by relating it to the random sampling selection of numbers.

Most participants appreciated the complex issues associated with statistical sampling. They realized the need for larger samples in understanding sample means being similar to that of the population, thus knowing that central tendency measures as the mean, median and mode, fall in the middle between 0 and 999. Participants understood the importance of random sampling and sample size, subsequently the effects of sample size on variability.

The application of the spreadsheet is overly stressed in this research. The heuristic presented in this study was simulated through the spreadsheet and thus the model stressed on variability number of sample means, changing of means and other central tendency measures. A final point concerns pupils statistical problem-solving heuristics in statistics. Often these heuristics are overcome by statistical intuitions that are inappropriate[16]. In fact, statistical thinking involves more than a procedural approach to solving problems.

This demonstrative study explored the central limit theorem through an experimental base activities lending practical methods to a better pedagogical treatment of a difficult concept taught through theoretical methods. Thus, participants gain authentic concrete problem-solving experience and engage in the process of discovery.

There are several limitations to this study. On one hand, the samples selected were all of sample size 5. Thus, varying the sample size would have substantially altered the distributions and a fuller meaning of the random sampling procedure. In addition, the purpose was to see the random effects of random samples, contrasting and comparing non-random sample helps students observe the behavior of two types of sampling distributions.

REFERENCES

1. Boye, J., R.J. Soukup and P.F. Williams, 1993. Using spreadsheets to teach problem solving in a first year class. IEEE Trans. Educ., 36: 68-71. DOI: 10.1109/13.204819
2. Coskey, C., 1988. Spreadsheet illustration of engineering economics. IEEE Trans. Educ., 31: 270-275. DOI: 10.1109/13.9753
3. Huelsman, L.P., 1984. Electrical engineering applications of microcomputer spreadsheet analysis programs. IEEE Trans. Educ., 27: 86-92. DOI: 10.1109/TE.1984.4321669
4. Moore, D., 2003. Basic Practice of Statistics. W.H. Freeman Company. 3rd Edn. New York. ISBN-10: 0716758814, pp: 674.
5. Levin, R., D. Stephan, T. Krehbiel and M. Berenson, 2001. Statistics for Managers Using Microsoft Excel. 3rd Edn. Prentice Hall, New Jersey. ISBN-10: 0130290904, pp: 913.
6. Abramovich, S., 2003. Spreadsheet-enhanced problem solving in context as modeling. Elect. J. Spreadsheets Educ., 1: 1-17. http://epublications.bond.edu.au/cgi/viewcontent.cgi?article=1000&context=ejsie
7. Macho, S., 2002. Cognitive modeling with spreadsheets. Behav. Res. Methods Instrument. Comput., 34: 19-36. http://www.ingentaconnect.com/content/psocpubs/brm/2002/00000034/00000001/art00003
8. Humberto, B., 2001.Teaching comparative statistics with Microsoft Excel. J. Econ. Educ., 32: 397. http://www3.wabash.edu/econexcel/compstatics/
9. Drier, H., 2001. Teaching and learning mathematics with interactive spreadsheets. School Sci. Math., 101: 170-179. http://ssmj.tamu.edu/abstract/abs_apr01.html#Teaching
10. Brophy, J. and J. Alleman, 1991. Activities as instructional tools: A framework for analysis and evaluation. Educ. Resr., 20: 9-23. DOI: 10.3102/0013189X020004009
11. Mitchell, M., 1997. The use of spreadsheets for constructing statistical understanding. J. Comput. Math. Sci. Teach., 16: 201-222. http://www.editlib.org/index.cfm?fuseaction=Reader.NoAccess&paper_id=20945
12 Groth, R., 2006. An exploration of students' statistical thinking. Teach. Stat., 28: 17-21. DOI: 10.1111/j.1467-9639.2006.00003.x
13. Shaughnessy, J., 1992. Research in Probability and Statistics: Reflections and Directions. In: Handbook of Research on Mathematics Teaching and Learning, Grouws, D.A. (Ed.). Macmillan, New York, pp: 465-493. ISBN-10: 0029223814
14. Shaughnessy, J.M., J. Garfield and B. Greer, 1997. Data Handling. In: International handbook of Mathematics Education, Bishop, J., M.A.K. Clements, C. Keitel and J. Kilpatrick (Eds.). Springer, Dordrecht, Netherlands: Kluwer. ISBN-10: 0792335333, pp: 1364.
15. Kahneman, D. and A. Tversky, 1972. Subjective probability: A judgment of representativeness. Cognitive Psychol., 3: 430-454. DOI: 10.1016/0010-0285(72)90016-3
16. Pollatsek, A., C. Konold, A. Well, and A. Lima, 1984. Beliefs underlying random sampling. Memory and Cognition, 12: 395-401. http://www.psychonomic.org/search/view.cgi?id=9964