SOME ISSUES OF DATA PRODUCTION IN TEACHING STATISTICS

Carl Lee
Department of Mathematics, Central Michigan University, United States of America
carl.lee@cmich.edu

Statistics educators, for a long time, have stressed the importance of using real data and focusing on variability of data in teaching statistics. Real data are messy. Due to the difficulty of creating a 'real' process of data production in a classroom setting, the issues of data production seem to be ignored in teaching statistics. Some of the issues may include (a) choice of measurement units, (b) robustness of measuring techniques, (c) the importance of operational definition, (d) subjective sampling vs. random sampling (e) observational vs. experimental studies (f) outliers Vs. errors, and (g) underlying target population. The use of real-time hands-on activities to teach students how to handle the issues of data production in a classroom setting is demonstrated.

INTRODUCTION

Data production is a process in the cycle of a statistical investigation. Due to the large amount and high dimension of data collected in many statistical problems, such as data mining and business intelligence applications, the process of data production often takes over 75% of the time for a project (Berry & Linoff, 2004). With the huge amount of time required for data production in such applications, research on data quality has become an important topic in these applications (e.g., Mallik, 2008; Batini & Scannapieca, 2006). Statistics educators and professional organizations have also stressed the importance of discussing the issues associated with producing data in a statistics curriculum.

The National Council of Teachers of Mathematics (NCTM), Mathematical Association of America (MAA) and the American Statistical Association (ASA) have developed recommendations for guiding the teaching and learning of statistics. The recommendations given in the 2005 GAISE report include (1) Emphasize statistical literacy and develop statistical thinking; (2) Use real data; (3) Stress conceptual understanding rather than mere knowledge of procedures; (4) Foster active learning in the classroom; (5) Use technology for developing conceptual understanding and analyzing data; (6) Use assessments to improve and evaluate student learning. Many statistics educators and researchers have pointed out the importance of actively involving students in the entire process of statistics investigation. Learning statistical concepts by using real data has become common recommendation among statistics educators. Cobb (2007, p. 338) summarized it well: "Ultimately statisticians are studying data, for which context is essential. What patterns mean, and indeed, whether they mean anything at all, depends on context. … Every data set has its own unique context, and its own special features."

In this article, the issues related to the process of data production will be addressed and the experience on how to engage students in identifying some important issues for proper data production will be shared. Through engaging students in the process of data production, students are led to learn seven important issues related to data production. These issues are identified using real-time hands-on activities developed by Lee and Famoye (2006). Section Two discusses the types of data that are available for teaching statistics. Section Three presents a list of issues related to data production and the use of real-time hands-on activities for students to investigate these issues, followed by some discussions and conclusions in Section Four.

TYPES OF DATA AVAILABLE FOR TEACHING AND LEARNING OF STATISTICS

Real data are usually messy. Variability can occur in every step during the process of data production. Data quality is clearly an important issue in every statistical application. However, most of statistics textbooks still provide 'artificial', 'cleaned' or 'nearly cleaned' data and emphasize mainly analysis and interpretations. This leaves a huge gap that is difficult for instructors to fill in their

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

instruction. Wild and Pfannkuch (1999) suggested that students need to be given numerous situations where they can address data collection issues and the impact choices may have on conclusions. Various pedagogies have been proposed to integrate data production as part of the teaching and learning of statistics. Among them are small-group activities to solve problems or analyze a set of data (e.g., Garfield, 1993; Garfield & Ben-Zvi, 2007) and using hands-on activities that collect students' own data. The Workshop Statistics (Rossman & Chance, 2003) and Activity-based approach (Gnanadesikan, *et al*., 1997) are among two very popular and successful approaches.

Cobb (2007) classified the sources of data into three categories: (1) archives (2) simulation and (3) hands-on data. He further discussed the advantages and disadvantages of each type of data for instruction, and emphasized the importance of using all of these different data sources in the teaching and learning of statistical concepts. He pointed out (Cobb, 2007, p. 341) that "the main advantage of activity-based data is that active participation gives our students the sense of ownership. A second important advantage is that some activities offer opportunities to think about issues of design."

Finzer and colleagues (2007) indicated that instruction in the introductory statistics course has relied on carefully honed collections of data sets well-suited to the illustration and investigation of each topic addressed and wrote (Finzer, *et al.,* 2007, pp.1) that "this judicious selection deprives students of the experience of data discovery." They discussed several approaches for data gathering: (1) Census microdata – samples from a huge population, (2) real time data, and (3) collecting classroom data using an online survey. They further elaborated the advantages of using the online survey for data collection: (a) it can collect many kinds of data, (b) classroom discussion about data collection can be beneficial for discussing the data measurement and issues related to data collection, (c) students can enter the data at anytime in different locations based on the students' convenience, (d) it does not take additional time for data entry by the instructor, (d) there is no waste of time managing data, and (e) instructors can focus on the concepts and questions to be discussed in class without interrupting the class time.

Lee and Famoye (2006) developed a real-time online database for hosting data collected from hands-on activities. The real-time online database site is at http://stat.cst.cmich.edu/statact/. These real-time online hands-on activities provide the opportunity for students to collect their own data, enter data online, and access the data immediately after the data entry. The following section discusses how these real-time hands-on activities are used to engage students in the process of discovering seven important issues related to data production.

TEACHING THE ISSUES RELATED TO DATA PRODUCTION USING REAL-TIME HANDS-ON ACITIVITIES

Statistics educators recognize that the data production process is an important aspect of statistical thinking for students and that it is crucial to integrate the data production process in the teaching of statistics. Chance (2002) identified two mental habits that students need to develop to think statistically: (a) keeping alert on validity about data and (b) considering how to best obtain meaningful and relevant data to answer the question at hand. The following demonstrates the use of real-time hands-on activities to introduce the issues that occur during the process of data production in an introductory statistics course.

Hands-on activities have been considered useful and successful activities that create active learning environments and are capable of engaging students in the process of statistical investigation. However, due to the long process needed for collecting and managing data using hands-on activities, only limited such activities can be facilitated in a classroom setting. The real-time online database developed by Lee and Famoye (2006) attempts to minimize the time needed for data production using hands-on activities and focus on the teaching and learning of statistical thinking by involving students in the entire process of statistical investigation. Data collected from each student have been stored and cumulated in the database. Some of the activities have thousands of cases. Stored data can be downloaded immediately by '*download the most recent n cases*', '*download your class data*' or '*randomly select n cases*'. The data in the online activity database are collected by students, about students from different classes in different institutions. The data sets are 'real', 'messy', 'large', and dynamically changed in 'real time'. These characteristics mimic real world projects well and provide unlimited scenarios for addressing the issues related to the process of data production. Many statistics instructors have used some of the real-time activities in their instructions. There are ten real-time hands-on activities currently available at http://stat.cst.cmich.edu/statact/.

A six-step framework to apply these activities in their introductory statistics instructions (Lee & Famoye, 2006) includes the following tasks related to the process of data production:

- choosing an activity and presenting a project scenario,
- discussing the variables of interest and ways of measuring the data,
- exploring the data to look for problematic cases,
- discussing the possible causes and taking action to handle these cases.

These tasks are often ignored in textbooks and as a consequence, are ignored in typical statistics instruction. By using the real-time hands-on activities, students are able to investigate and learn to handle the issues related to data production. Table 1 summaries the issues related to data production investigated by students and the corresponding real-time hands-on activities used to address these issues. The next section demonstrates how students are engaged in the investigation of these issues using some of these activities.

Table 1. Issues Related to Data Production and Real-Time Hands-on Activities Used

	Issues Related to Data Production	Activities
1	Choice of measurement units	Distance, Hand_size, Exercise
2	Robustness of measuring techniques	Distance, Hand_size, Exercise
3	The operational definition of variable	Distance, Hand_size, Exercise
4	Subjective sampling or random sampling	Sampling, Vote, Random_selection, Raisins
5	Outliers Vs. errors	Distance, Hand_size, Exercise
6	Observational Vs. experimental study	Exercise, Vote, Colleg_life, Random_selection
7	Underline target population	Raisins, Vote, Distance

CASE STUDIES FOR DEMONSTRATING THE USE OF REAL-TIME HANDS-ON ACTIVITIES TO INVESTIGATE THE ISSUES RELATED TO DATA PRODUCTION

In the past few years, the activities listed in Table 1 have been used to engage students in investigating the issues related to data production. The follow are case examples on how the issues of data production are addressed in the classroom setting.

- Choice of measurement units: The first step in the framework proposed by Lee and Famoye (2006) is "defining the project scenario and choosing the hands-on activity for the project." Once a problem scenario is introduced, the next step is to discuss the variables of interest and how the variables should be measured. Prior to collecting data, choice of measurement units must be determined. For example, the problem scenario of investigating 'How far are you away from home?' involves how to measure distance. Here are some common responses from students:
  'I am 100 miles away from home.' 'I am 2 hours away from home.'
  'We use kilometers in my country.' ' I am 800 kilometers from home.'

This distance activity brings lively discussions about the importance of proper choice of measurement. The cultural aspect can play a role in the choice of measurement unit. For the 'distance' activity, the choice of miles or kilometers clearly reflects the cultural difference.

- Robustness of measuring techniques: Which measurement unit is more robust is often ignored in most statistical instruction. By using the hands-on activity such as 'How far are you away from home?', students discuss whether' miles' or 'hours' , is more robust in terms of how the data are measured and who measures the data. Here are some common responses from students:

  'I think mile is a better measure for distance, since it is actually measures the distance'.

  'I think hour is not a good measure since people may drive in different speeds.'

  For this distance activity, using 'time' to measure distance results in more variability of distances because individuals drive in different speeds. Besides, it is not a direct measurement of distance. One may argue that if miles are used, then different routes will give very different distances. This comes down to another interesting discussion on the issue of 'operational definition'.

- The importance of operational definition of variable: Once the discussion of choosing measurement units has begun, the following questions are commonly brought up among students:

  'How about using kilometers? If this activity is conducted in my country, people will certainly use kilometers. How can we be sure that students enter data in miles?'

  'But, if I drive on county roads, the distance would be very different from driving on a highway'

  'There are several highways; which should I use to estimate the distance?'

  Statisticians are aware of the importance of operational definitions of the variables to be measured in order to reduce unexpected data errors. The 'Distance' activity quickly leads students to bring up this issue. The instructor can then provide a few operational definitions needed for measuring the distance, such as:

  'We will assume you drive on the major freeway'. 'The measuring unit will be miles'.

  'This is an estimate; you do not need to be very precise. Think about the exit number from where you get in and where you get out to help you'.

- Subjective sampling vs. random sampling: Random sampling is an important concept for statistical investigation. A clean data set given in textbooks or existing real data do not provide an opportunity for students to investigate the importance of 'randomness'. In the 'subjective sampling vs. random sampling' activity students investigate this issue. Students are given a population consisting of 100 blocks; each a different length, and are asked to subjectively select a sample of size ten that best represent the population. At the same time, a computer randomly selects a sample of size ten. The data are recorded; sample mean and standard deviation are computed. Some common comments from the discussions by students are:

  'I thought the sample I choose really is better than the random samples. Because I see the entire population, and I should be able to choose the best ten representing the population.'

  'It is surprising to see how wide the differences in averages are among individual's subjective selections, and how close the averages are from different random samples.'

  'It seems, from the histogram, the center of the averages computed from the subjective selection is smaller than the actual population mean, but the center of the averages computed from the random samples is much closer to the population mean.'

  These comments often generate some curiosity regarding 'why this happens, and does this happen in other classes'. At the time this activity is conducted, students have not yet learned the sampling distribution of a sample statistic, nor have they learned about the estimation concept. However, by learning about the potential problem of subjective sampling during the data production process, they have a first experience about sampling distributions of sample mean and sample standard deviation.

- Outliers vs. errors: When a 'clean' data set is presented in textbook, unusual observations can artificially be introduced as outliers. However, it is not possible to discuss the difference between 'outlier' and 'data error' without a real context. Using hands-on activity, this is a natural topic for discussion. 'Is the data value an outlier or an error?' and 'What are the possible causes that may result in this data value?' are among commonly raised questions during the class discussion.

The activity 'Does one minute exercise dramatically increase your pulse rate?' often generates a discussion about outlier vs. data error. In this activity, students are asked to measure their pulse rates prior to the exercise, then do one minute of aerobic exercise with music provided. Pulse rates are taken immediately after the exercise. Here are some common observations from students:

> 'Look at this case, pulse rate after exercise is smaller than that before exercise. Something must be wrong.'; 'How about this case, the change of pulse rate is from 60 to 135. Is it possible?'

Curiosity from students usually leads to discussion about possible causes for such results in the data. Some possible sources of errors commonly pointed out by students are:

> 'It may be that students did not really do the exercise.' 'It may be a data entry error.'
>
> 'It may be a counting error.' 'I can see there is a mistake; it is not possible to have a smaller pulse rate after the exercise. But, it is possible that the pulse increased from 60 to 135. It may not be an error. It is just rare.'

The instructor can take this opportunity to address the difference between possible outliers and possible data errors, and talk about how to handle the situation.

- Observational vs. Experimental Study: The hands-on activities in the real-time online database are observational studies. The instructor needs to provide different examples to illustrate principles of designing an experiment, such as case control studies to examine the effect of a new drug, and ask students to design an experiment by giving them a problem scenario. One activity is to ask students to design an experiment to investigate if one minute of exercise dramatically increases pulse rates as a homework problem. A surprise is that many students indicated 'we also need to collect weight, age and gender as they are potential factors affecting the pulse rates.' Many students suggest 'using a treadmill as the instrument, separating male and female groups, and considering using electronic devices to measure the pulse rates.'

  Without introducing different types of experimental designs, students can build various insights about designing experiments through this activity.
- Underlying target population: Many real world data come from more than one population, especially observational studies. The hands-on activity 'How many raisins in a ½ ounce box of raisins?' provides the opportunity to investigate this issue. Due to the fact that the data are collected from different classes in different institutions, instructors may use different brands of raisins or different weights of raisin boxes. As a result, data may be from different populations (different brands or different weights of raisin boxes). The instructor purposely downloads a set of raisin data that show a bimodal histogram. Here are some common responses from students:

  > 'Is this because of counting error?' 'It looks like they are from two different classes. Observations 1 to 40 are similar, and observations 40 to 60 seem to have more raisins.' 'I think one of the classes may have a larger box of raisins.' 'It may be different brands'.

  In the discussion the instructor can point out the issue that data are from more than one population and ask students to discuss how to handle the problem.

DISCUSSION AND CONCLUSION

Although statistics educational reform and recommendations have been stressing the need to involve students with the process of data production, changing the curriculum to include data production seems to be slow. Most textbooks have included the use of real data to provide the context behind the data. However, most textbooks have not yet included activities that allow students to experience the process of data production. As a consequence, there is a lack of opportunity for students to consider the issues related to data production and the importance of data quality. Garfield (1995, pp. 30) wrote, "Regardless of how clearly a teacher or book tells them something, students will understand the material only after they have constructed their own meaning for what they are learning." Melton (2004, p. 11) pointed out "When students recognize data collection as a process and begin to apply concepts of statistical thinking to this process, they begin to look at statistics in a different way, rather than seeing statistics as an isolated quantitative course. … Helping students recognize some of the reasons that variation can exist in the data collection process is the first step toward helping them evaluate the usefulness of data for a given situation."

The investigation of the issues related to data production requires students to construct their own meaning behind the issues. This paper presented seven issues related to the process of data production that are discovered and discussed among students in an introductory statistics course. Real-time hands-on activities can engage students in the investigation of these seven issues. The goal is to help students not only to be able to analyze data and interpret the results but also to be able to diagnose the potential problems in the production of data and take action to ensure the high quality of data prior to analysis. The seven issues are common issues that occur in real world data and are checked by data analysts. The data quality issues are usually addressed in various graduate statistics course, but rarely are discussed in an introductory statistics course. These real-time hands-on activities have been successfully used in introductory statistics courses for students to learn these important data production issues.

REFERENCES

Batini, C., & Scannapieca, M. (2006). *Data quality: Concepts, methodologies and techniques*. Springer, Inc.

Berry, M., & Linoff, G. (2004). *Data Mining techniques* (2nd Ed.). Wiley Publishing.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). Online: www.amstat.org/publications/jse/v10n3/chance.html.

Cobb, G. W. (2007). One possible frame for thinking about experiential learning. *International Statistical Review*, *75*(3), 336–347.

Finzer, W., Erickson, T., Swenson, K., & Litwin, M. (2007). On getting more and better data into the classroom. *Technology Innovations in Statistics Education*, 1.

GAISE Report (2005). Online: it.stlawu.edu/~rlock/gaise/.

Garfield, J (1993). Teaching Statistics Using Small-Group Cooperative Learning. *Journal of Statistics Education, 1*(1). Online: www.amstat.org/publications/jse/v1n1/garfield.html.

Garfield, J. (1995). How students learn statistics. *International Statistical Review, 63*(1), 25-34.

Garfield, J. & Ben-Zvi , D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review,* 75(3), 372–396.

Gnanadesikan, M., Scheaffer, R., Watkins, A., & Witmer, J. (1997). An activity-based statistics course. *Journal of Statistics Education*, *5*(2). Online: www.amstat.org/publications/jse/v5n2/gnanadesiakn.html.

Lee, C., & Famoye, F. (2006). Teaching statistics using a real time online database created by students. *Proceedings of the Seventh International Conference on Teaching Statistics*. Salvador, Brazil: International Statistical Institute and International Association for Statistical Education. Online: www.stat.auckland.ac.nz/~iase/publications.

Mallik, Sanjib (2008). Data Quality Issues for the 21st Century. Online: www.datagym.com/company/DQ_Seminar/Sanjib/index.htm.

Melton, K. (2004). Statistical thinking activities: Some simple exercises with powerful lesson. *Journal of Statistics Education*, *12*(2). Online: www.amstat.org/publications/jse/v12n2/melton.html.

Rossman, A.J., & Chance, B.L. (2003). *Workshop Statistics*. Key College Publishing.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*, 223-265.