INFLUENCE OF CONSULTING IN THE SELECTION OF TOPICS WHEN TEACHING STATISTICS

Edith Seier
East Tennessee State University, United States
seier@etsu.edu

Statistical consulting not only provides real life examples to mention in class; it provides a reality check that influences the way we teach and the choice of topics to teach or emphasize. The focus of this paper is on topics that are frequently omitted or not emphasized but which we consider important as a direct result of consulting experiences. The first four belong to introductory courses and among them are data issues, the test of hypothesis about proportions for small samples using the binomial distribution and some topics on categorical data analysis. The last two topics, one on statistical models and the other in time series, belong to upper division courses.

INTRODUCTION

Intramural consulting in walk-in services inside universities, external consultation, private practice and collaborations with faculty from other departments all provide statisticians with practical experience that can enrich their teaching. Students usually become very attentive in class when we mention a survey or any other example we have worked in as professionals. There are excellent books about case studies that can be used to assign readings at different levels such as Lange *et al*. (1994), Peck *et al*. (2005) and previous editions of the same collection. Cases of their own consulting provide, for many statisticians, examples for topics already included in courses. Villagarcia (1998) shows how a consulting case provides real life motivation to teach engineering students about the sum of random variables and the Central Limit Theorem. Truran and Arnold (2002) introduced consulting in an introductory statistics course. According to Esfandiari and Lew (2004), the use of case studies allows us to introduce statistics as a mean of answering real world questions instead of a series of procedures.

This paper focuses on certain topics that are frequently omitted or that could be treated to a greater extent without increasing the difficulty of the course but increasing its usefulness. These topics have been identified as appropriate to include after several years of working in walk-in consulting services inside universities, developing long time collaborations with faculty members of other departments, and at the same time teaching a variety of statistics courses; in particular, the introductory algebra-based statistics course for non-majors.

What to include in a first statistics course for non-majors is a frequently addressed issue. A good reference about what an educated citizen should know about statistics is found in Utts (2003). Reviewing several introductory statistics books**,** one can sense that most authors agree on the list of topics that should be taught in a first algebra-based s**t**atistics course. Nevertheless, the instructor's own experiences can be used to enrich the content of the course. Of course, the experience of each consultant is different and it is not the intention to overcrowd a course with too many or too specialized topics that are seldom used. The point is that consulting can give us a perception about the relative importance of topics in the environment in which we work and can help us to detect some deficiencies. Consulting can also help us realize how some non-statisticians, who need statistics, perceive certain fundamental statistical ideas and in that way identify the need to make those concepts and ideas more clear in our courses. It is to be noted that the influence of consulting in teaching goes beyond what we discuss with our students in statistical consulting courses for statistics majors.

There are some topics that in our opinion should receive more attention in statistics courses. Some belong to introductory statistics, one to statistical methods and the last to time series analysis.

TOPICS IN INTRODUCTORY STATISTICS COURSES

Population/Sample/Data File/Tables/Graphs: A Smooth Presentation Through a Single Story

Many of the cases in walk-in-consulting are about simple issues related to data that people have collected from individuals and have put in a data file to be analyzed. However simple the task, there are several issues, questions, anxieties about quantitative analysis, and small mistakes that surface in the moment of consulting, which could be avoided if they were routinely addressed in courses. Introductory textbooks all include material about producing data, frequency tables and two way tables. These tasks are part of a sequence in real life but they sometimes appear fragmented by occurring in different sections of the textbook, skipping some details to which we need to pay attention between the production of the data and the analysis. Among the forgotten details are: to number the questionnaires so that we can go back to them in case we see some non-reasonable value during the analysis or to do spot checking for the quality of the data, how to design a data file, simple ways to deal with questions in which the respondent can mark more than one answer; and to use a simple frequency table to check that all the values are reasonable before proceeding with the analysis. It is not uncommon that people jump into the analysis without first checking the data, and later realize that some of the answers included codes for missing values which are not really values of the quantitative variable, or that a serious typo appeared.

It is not the intention to convert the introductory course into a methods course, but if all the parts are already there, why not put the puzzle together, and make a smooth transparent transition from data collection to data analysis. It is nice to work with real data without overwhelming the students with a too complicated situation. Usually during the first class I ask the students 'Why do we do surveys?' That gives the opportunity to talk about populations, samples, parameters and statistics. I tell them about some survey I have worked on in the past; public health or opinion surveys are usually examples that can get them interested. The example can be used to stress the importance of randomness in the selection of the sample. We comment that each questionnaire has a number to identify it, each individual becomes a row in the data file and each question is a column, but if a question has several possible answers we can create a column for each one of them. They are then given a data file with a few variables to explore. A data file with data about several variables including age, gender, smoking status, marijuana use and alcohol dependence for 500 individuals selected at random in a certain region can be found in http://www.etsu.edu/math/seier/1530/drugsurv.mtw. This gives us the opportunity to discuss the nature of the variables (quantitative or categorical, discrete or continuous, nominal or ordinal) and to check if the values of the variables are reasonable by doing a simple tally which brings us to the idea of tables. We talk about describing the sample by its demographic variables, if any, and then I start asking them questions about the answers given by the respondents. Almost without noticing, the topics of frequency tables, two-way tables and the usual graphs flow as needed to answer the questions. Time is saved by introducing several topics with a single case, and a more connected view of data collection and data display is offered to the student. Then we move to other data sets to get a wider variety of shapes of distributions in the case of quantitative variables.

Test of Hypotheses for Proportions in the Small Sample Case

Most introductory algebra-based books do a good job discussing the test of hypotheses for proportions using large samples and the assumptions the normal approximation requires. However, most of them do not treat the case of small samples using the binomial distribution, an exception being Agresti and Franklin (2007) that includes an explanation (p. 385) and two optional exercises about this case. The exact test for proportions using the binomial distribution is also absent from several calculus-based introductory books, an exception being Rossman and Chance (2005). The test of hypotheses about a proportion using the binomial distribution is actually easier to explain than the large sample case using the normal approximation, as shown in Seier and Robe (2002), because it requires no formulas and the only requisite is that the student should be familiar with the binomial distribution. The small sample case could be used to introduce the topic before the large sample method is explained.

The need for a tool to test hypotheses with small samples usually arises not in large scale surveys but in the academic environment when students do small research experiences as part of their course work, or in experiments when large samples are not possible. The same tool in the hands of school teachers could be very useful if they advise students in science fair projects.

Independence is Not the Only Question

Most modern introductory books distinguish between the test of homogeneity and the test of independence. However 'independence' and 'Chi-square' are the two words that commonly come to the mind of non-statisticians when they see a two-way table. After being accustomed only to the idea of independence, it takes a while for them to later accept a more general framework and become aware that independence might not always be what we are curious about. The matched-pairs situation, in which either agreement or symmetry are of interest happens in real life. Some second courses in statistics for non-majors include the 'agreement of raters' situation but this topic could be easily introduced in the introductory course. In a consulting case, the hips (left and right) of a group of patients are classified either as having or not having osteoporosis based on the measurement of the bone mineral density (BMD). Osteoporosis might be present in just one, both or neither hip. A current debate in medicine is whether it is necessary to measure both hips or only one. The bone mineral density of the left and right hip are highly correlated; however the diagnosis of osteoporosis in each hip can be different because in one hip the BMD could be above -2.5 and in the other it could be below that value. Knowing that the BMD in both hips are correlated we might no longer be curious about testing for independence but we could be interested in discussing if osteoporosis is more often present in either the right or the left hip, and how much coincidence is in the diagnosis done with each one of the hips. In another case a graduate student in computer science is developing a new method to detect if a hard drive is damaged or not and wants to compare it to the commercial software that is traditionally used with this purpose. In both cases, testing for symmetry and the measurement of agreement are important.

More on Association; R is Not Alone

Many of the consulting cases refer to the analysis of categorical variables; however the overall content of introductory textbooks tends to put more emphasis on quantitative variables. Most college educated people, non-statisticians, remember the Pearson correlation coefficient but they are not aware that association between categorical variables can be quantified. It is true that there is no association statistic in the categorical world with the universality of the ***r*** in the quantitative world but at least the odds ratio and the relative risk could be introduced when talking about two-way tables. These two statistics are easy to calculate and useful not only for people applying statistics but for the general public that receive information in the media about public health studies. Several introductory algebra-based statistics textbooks published in the U.S., with at least one edition after the year 2000, do not include odds ratio or/and relative risk. However, Utts (2005), Utts and Heckard (2004), Aliaga and Gunderson (2006), and Agresti and Franklin (2007) do include either odds ratio, relative risk or both. Rossman and Chance (2005) does include odds ratio and relative risk.

TOPICS IN UPPER DIVISION COURSES

Logistic Regression- More Useful than it is Generally Advertised

Multiple Regression and Analysis of Variance have been traditionally the main topics in Statistical Models or Statistical Methods courses. Logistic regression has been gradually introduced in the last two decades but sometimes it is not given all the importance it deserves, particularly when the statistical methods courses for non-majors are taught not in the statistics or Mathematics Departments but in other departments, and not by statisticians but by specialists in the application fields.

Cases in which Logistic Regression is an appropriate tool happen frequently in many fields of applications. We have had cases in consulting from such different fields as experiments with chemicals and asphalt, surveys in criminal justice or tobacco use, experiments with the learning process in bees, and applications of graph theory to genetics.

Introducing Spectral Analysis in the Undergraduate Time Series Course

It is not uncommon for undergraduate Time Series Analysis courses to cover exploratory analysis, the decomposition approach and ARIMA models, but to overlook spectral analysis. We think it is possible to do a gentle short introduction to spectral analysis starting with harmonic functions, going to the periodogram and from there to the spectrum. With that purpose we have developed class notes that at present are evolving into a tutorial and will be available on line. The tutorial includes programs in *MATLAB* and *R*. The inclusion of the basics of spectral analysis gives a more panoramic view to students in the Time Series course. We have included the topic for many years and found it useful in exploring seasonality among other things. The tutorial and programs have been used not only for the students in the course but to explain the basic concepts to professionals from other fields who have not had training in the topic and need to apply it. The new version of the tutorial is influenced by long time collaboration with biologists doing research on circadian rhythm.

CONCLUSION

Consulting provides not only examples to share with our students in classes at different levels, but also helpful information for the design of courses and the identification of concepts that need to be emphasized. Based on consulting experiences, we have identified the value of the inclusion of, or more emphasis on, certain topics in introductory statistics and some upper division courses. Consulting should be encouraged among statistics faculty and their experiences formally discussed in the context of improving course content and the teaching of statistical concepts.

REFERENCES

Agresti, A. and Franklin, C. (2007). *Statistics: The Art and Science of Learning from Data*. Upper Saddle River, NJ: Pearson Prentice Hall.

Aliaga, M. and Gunderson, B. (2006). *Interactive Statistics* (3rd Edition). Upper Saddle River, NJ: Pearson Prentice Hall.

Esfandiari, M. and Lew, V. (2004). Bridging the gap between consulting and teaching. Center for Teaching Statsitics, University of California, Los Angeles. Paper 2004010101.

Lange, N., Ryan, L., Billard, L., Brillinger, D., Conquest, L., and Greenhouse, J. (Eds.) (1994). *Case Studies in Biometry*. New York: Wiley.

Peck, R., Casella, G., Cobb, G. Hoerl, R. Nolan, D., Starbuck, R., Stern, H. (Eds.) (2005). *Statistics a Guide to the Unknown* (4th Edition), Belmont, CA: Duxbury.

Rossman, A. and Chance, B. (2005). *Investigating Statistical Concepts, Applications, and Methods* (Preliminary Edition). Belmont, CA: Brooks Cole.

Seier, E. and Robe, C. (2002). Ducks and green: An introduction to the ideas of hypothesis testing. *Teaching Statistics*, 24(3), 82-86.

Truran, J. and Arnold, A. (2002). Using consulting for teaching elementary statistics. *Teaching Statistics*, 24(2), 46-50.

Villagarcia, T. (1998). The use of consulting work to teach statistics to engineering students. *Journal of Statistics Education*, 6(2).

Utts, J. (2003). What educated citizens should know about Statistics and Probability. *The American Statistician*, 57(2), 74-79.

Utts, J. and Heckard, R. (2004). *Mind on Statistics* (2nd edition). Belmont, CA: Duxbury, Brooks Cole-Thompson.

Utts, J. (2005). *Seeing Through Statistics* (3rd edition). Belmont, CA: Duxbury, Brooks Cole-Thompson.