Fundamental Statistical Analysis for a Future with Big Data

Session Leaders and Moderators
Robert L. Andrews,
Virginia Commonwealth University,
Department of Supply Chain Management and Business Analytics,
Richmond, VA. 23284-4000, 804-828-7101, randrews@vcu.edu

Wilma M. Andrews,
Virginia Commonwealth University, Department of Information Systems,
Richmond, VA. 23284-4000, 804-827-0956, wandrews@vcu.edu

ABSTRACT

Session examines issues that the advent of Big Data present relative to introductory business statistics instruction. It is followed by an interactive discussion of what changes, if any, should be made in introductory business statistics instruction to prepare students for a world with big data. The discussion piece addresses whether the typical statistics course prepares business students with the skills they will need for a future where big data sets and decision making based on data will be more prevalent. The session will include a demonstration of capabilities for analyzing data in Excel 2013 focusing on recently introduced capabilities that are available through Excel and able to handle bigger data sets.

INTRODUCTION

It is a well-known fact that the amount of data being collected and available for analysis is rapidly increasing. As a result the size of data sets is also increasing and there are new types of data available for analysis. Big Data is the name often used to refer to such data. However, big data, like many terms we use, does not have a universal definition of what constitutes big data. Emerson and Kane [4] define big data as the quantity of data and use of the percent of the computer's random access memory as a measure of big since the computer is what is being used to analyze the data. An Internet search for Big Data yields the April 28, 2013 definition by Wikipedia to be "a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications." This definition focuses on size relative to processing capability and as such is not really germane to a foundational business statistics class other than to the selection of computational software that should be used in the class. One might note that using a handheld calculator is not providing any real value for analyzing big data. One part of this session will include an overview of analysis procedures available through Excel that are appropriate for a fundamental business statistics class and these will be presented later.

In contrast to the definitions of big data that focused on the size of data, Michael Horrigan [5] from the Bureau of Labor Statistics has a slightly different view and sees "Big Data as nonsampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference." This definition focuses on the purpose of the data rather than the size of the data set and purpose of data analysis should be

germane to a foundational class in business statistics. Most business statistics textbooks and one would also infer that most business statistics courses focus on descriptive statistics that summarize the characteristics of the observed data and on inferential statistics that strive to reach conclusions about characteristics of a larger population or process that was the source for the observed data. According to Horrigan's definition of Big Data, standard inferential statistical instruction is not properly preparing students for the analysis of big data.

An IBM Institute for Business Value study of Chief Financial Officers [7] found that the CFOs felt that a significant gap existed between skills required for today's business environment and the skills currently available in the workforce. This, along with other similar pronouncements, indicates that our present educational process is deficient for the perceived future needs of businesses. The business statistics class is one that should have a clear focus on analysis of data. Historically this has included a set of analysis procedures/techniques listed in the course description with a primary focus on descriptive statistics, understanding probability distributions, confidence interval estimation and testing for the statistical significance of hypotheses. Testing of statistical hypotheses may match well with the analysis many faculty members are doing to get their research published but it does not match well with the analysis needed for big data using Horrigan's "nonsampled data" as the definition for big data.

Brown and Kass [2] in their critical look at statistics state, "Degrees in Statistics have emphasized a large suite of techniques, and introductory courses too often remain unappetizing. The net result is that at every level of study, gaining statistical expertise has required extensive coursework, much of which appears to be extraneous to the compelling scientific problems students are interested in solving." As their comment indicates, their orientation is on the physical and health sciences and neuroscience in particular, but we feel their comments are equally applicable to business in the sense that much of the coursework in a business statistics class may be considered as extraneous to the business problems our graduates will need to solve using big data. They continue saying, "Computer scientists have been especially influential in the past decade or so. … As others leap daringly into the fray, attempting to tackle the most difficult problems, might statistics as we know it become obsolete?" Faculty members teaching business statistics are often heard expressing concern about the diminished role of statistics in the business curriculum at their school. Its role in the business curriculum should be well established if the business statistics class is providing students with skills they will need for a future that is forecast to be reliant on data-based decision making using big data.

Based on their findings Brown and Kass conclude, "We are worried. While we expect that in many institutions–perhaps most–there may exist specific courses and programs that are exemplary in certain aspects, in the aggregate, we are frustrated with the current state of affairs. The concerns that we have articulated here are not minor matters to be addressed by incremental improvement; rather, they represent deep deficiencies requiring immediate attention." They suggest two overarching principles for curricular revision:

1. A focus on statistical thinking and
2. Flexible cross-disciplinarity.

Specifically they say "the primary goal of statistical training at all levels should be to help students develop statistical thinking." But they point out "According to syllabi and lists of requirements, statistics courses and degree programs tend to emphasize mastery of technique." It

is easier to have a narrow focus and teach mastery of technique rather having a broader focus that includes learning technique as well as statistical thinking and cross-discipline applications.

Hal Varian [8] presented a brighter picture and is cited as saying that he "sees statisticians as part of the reconfiguring of the business industry's future" and that "Information is easily accessible, and statisticians can help organizations analyze the information to improve productivity." A place for statisticians to do this is to better prepare our business students for their future in our business statistics class. The cited statements from Brown and Kass expressed the short-comings of current statistical instruction while those from Varian focused on the potential for those properly trained in statistics to play an important role in the future of business.

A window of opportunity exists for an improvement to be made in business statistics instruction. This window has been opened through the deluge of data presented by big data and the potential strategic advantage that can be obtained by proper analysis being touted by Davenport and Harris [3] in their book *Competing on Analytics*. Even though Brown and Kass criticize an almost exclusive focus on techniques that does not mean we should stop teaching analysis techniques.

Today's hot quantitative topic is analytics. Peter Bell [1] says the best way to explain analytics is to use the SAS 8-levels of analytics framework with "(1 = standard reports, 2 = Ad hoc repots, 3 = query drilldown, 4 = alerts) which are familiar to most, and this framework also highlights the big step from 5 (the statistical analysis of historic data) to 6 (forecasting) and 7 (predictive modeling)." The business statistics class should be introducing students to foundational procedures in levels 5, 6 & 7. He points out that their level 8 (optimization) is not the "pinnacle of analytics" from everyone's perspective because "many firms have more difficulty with risk analysis and coping with uncertainty." Learning the fundamentals of risk analysis and coping with uncertainty should be a primary learning objective of the business statistics class to truly prepare students for making decisions that are guided by knowledge obtained from big data.

An important aspect of big data is that almost all of what people refer to as Big Data are data gathered over time. The Wikipedia information about Big Data includes this statement, "Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks." All of these data sources are obtaining data over time and not just taking measurements at regular intervals. The classic business statistics addresses inference for situations where the data are a sample from a fixed population. Most people teaching business statistics would agree that Horrigan's "nonsampled data" definition for big data mean that the data are not a representative sample from a population. Some statistics classes address the situation that may exist when the sampling frame does not match the target population of interest and the implications of this mismatch. However, most introductory business statistics textbooks do not address how to deal with data gathered over time, except in a time series or forecasting chapter. The techniques in these chapters are for data gathered on a single variable of interest and at regular intervals. The introductory class should not attempt to cover everything but it should make students aware of important things to consider relative to using data for the SAS levels 5 (the statistical analysis of historic data), 6 (forecasting) and 7 (predictive modeling). Also the introductory class should provide some applicable data analysis skills using software.

APPROPRIATE SOFTWARE FOR THE BUSINESS STATISTICS CLASS

Analysis of big data, however you define it, must be done with computational software. Which software is most appropriate for the introductory business statistics class? The answer certainly depends on the criteria used. We believe that the primary criterion is what would be best to prepare the students for their future and not what would be best to help me teach the things I want to teach in the course. They need to know tools that will allow them to analyze data when they get a business job. Cheryl McKinnon [5] states, "Big data gets all the hype today, but enterprises around the globe continue to be run by big spreadsheets." She also cites a survey by Deloitte in 2009 reporting that 99.7% of businesses surveyed said that they used spreadsheets and 70 percent of respondents had 'heavy' reliance on spreadsheets to support critical portions of their businesses. She reaches these conclusions, "Spreadsheet data is enormous in size and impact" and as a result "do not skimp on analytic tools that can be used and understood by a typical business worker."

The 2013 KDnuggets Software (14th annual) Poll had 1880 voters answering the question, "What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project?" Excel finished third with 28% reporting that they had used it (29.8% had reported using it in the 2012 survey). The winner for 2013 was Rapid-I RapidMiner/RapidAnalytics free edition with 39.2% (26.7% had reported using it in the 2012 survey placing it behind Excel). The open source statistical software R was second with 37.4% (30.7% had reported using it in the 2012 survey placing it slightly above Excel). It should be noted that this "RapidMiner has been very successful in motivating their users, and got the most votes."

In addition to the above support for Excel, Peter Bell [1] recommends the use of Excel. The bottom line is that Excel is our recommended software because a graduate hired by business is almost sure to have Excel available for use at the business and a large percentage of analysts doing analytics, big data, data mining, data science analysis use it. Plus Excel has added several new capabilities in Excel 2013 that increase its big data capabilities and the statistical packages JMP and SAS can be accessed as an add-in with a tab on Excel.

EXCEL CAPABILITIES FOR THE BUSINESS STATISTICS CLASS

The session will end with a presentation of some Excel based analysis methods/techniques that we believe will prove to be of future value to students.
The overview of Excel topics will include:

- New Charting Capabilities
- Conditional Formatting
- Flashfill
- GeoFlow
- PowerPivot 2013
- PowerView
- Quick Analysis Tools Slicers
- Sparklines
- Standard Pivot Tables

SUMMARY

The main focus of the session will be on what is being taught in the introductory business statistics class and what should be taught to prepare students for a future that will be more reliant on decision making using data. Big Data is a hot topic now and as a hot topic it presents an opportunity for the introductory business statistics to have a more important place in the business curriculum if the course is properly structured to provide real value to the students. If statistics faculty do not take advantage of this opportunity then faculty from other disciplines will fill the void and the value of the introductory business statistics class will decline. The reasons cited support the authors' conclusion that Excel should be the software of choice for the introductory statistics class so that the students are best equipped for their future. Hence the session also covers Excel capabilities and illustrates how the knowledge of using these effectively will be valuable for students in analyzing big spreadsheet data.

REFERENCES

[1] Bell, Peter C., "Innovative Education: What every business student needs to know about analytics," *OR/MS Today*, Vol. 40, No. 4 (August, 2013), pp 24-27.

[2] Brown, Emery N. and Kass, Robert E., "What is Statistics," *American Statistician*, Vol. 63, No. 2 (May 2009), pp. 105-110.

[3] Davenport, Thomas H. and Harris, Jeanne G., *Competing on Analytics: The Science of Winning*, Harvard Business School Press, Boston, MA, 2007

[4] Emerson, John W. and Kane, Michael J., "Don't Drown in the Data," *Significance*, Vol. 9 Issue 4 (August 2012), pp. 38-39.

[5] Horrigan, Michael W., "Big Data: A Perspective from the BLS," *Amstatnews*, Issue#427 (January 2013), pp. 25-27.

[6] McKinnon, Cheryl, "Never mind 'big data' We're still coping with the era of 'big spreadsheets," http://www.fiercecontentmanagement.com/story/never-mind-big-data/2012-11-05#ixzz2cqobwYF2, (November 2012)

[7] IBM Institute for Business Value, *The New Value Integrator: Insights from the Global Chief Financial Officer Study*, IBM Global Business Services, Somers, NY, March 2010

[8] Varian, Hal, "Statistician: A Sexy Job," *Amstatnews*, Issue#381 (March 2009), pg. 17.