120th ASEE Annual Conference & Exposition
FRANKLY, WE DO GIVE A D*MN
June 23-26, 2013

Paper ID #6365

Experimental Assessment of Higher-Level Data Analysis Skills

Capt. Julie Ann Layton, Rensselaer Polytechnic Institute

CPT Layton is a master's degree candidate in the RPI Dept. of Industrial and Systems Engineering.

Prof. Thomas Reed Willemain, Rensselaer Polytechnic Institute

Professor of Industrial and Systems Engineering, Rensselaer Polytechnic Institute

Experimental Assessment of Higher-Level Data Analysis Skills

Introduction

Data analysis is a process learned over time that requires comprehension of all facets of statistics (summary statistics, statistical graphics, estimation, and inference). Data analysis is a unique skill set required in nearly every field of study, from business sectors such as supply chain management and quality control to science-based occupations including physics, materials science, and medicine. One must be able to identify problems or trends within the data to make progress in research and avoid recurring mistakes. These additional skills allow burgeoning analysts to evolve from passively following along with someone else's data analysis to effectively conducting their own analytical experiments.

The goal of this research is to understand how better to help novices become expert data analysts. The first step to solving this problem is description (i.e., find out how novices analyze data by presenting them with a properly designed exercise in data analysis). The second step is to observe the behavior each subject exhibits and try to find behavior correlations based on factors such as the number of statistics courses taken, final exam grades, year group, gender, or course level. The next goal is to identify distinct types of behaviors. This paper reports on an experiment. The methods used to analyze this data are analysis of variance, Markov chains, and chi-squared analysis to compare two Markov chains.

Need for data analysts

"Companies are increasingly turning to analytics to gain a competitive edge. As they do, they must resolve unique demands on their information technology, their structure, their processes, and their culture. Most critical, however, is the challenge posed by analytical talent, the people at all levels who help turn data into better decisions and better business results." [1]

The above quote is from a publication by Accenture, a global consulting firm, about the importance of recruiting analytic talent for businesses. In response to this need several colleges have developed graduate programs in analytics, one of the first being North Carolina State University's Institute of Advanced Analytics, which boasts job placement rate of 100% [2]. Several other articles echo the need for trained data analysts in the information age where large amounts of data are being collected with a growing need for those that can make "data-driven decisions" [3]. McKinsey Global Institute, a business and economic research firm, claims that with the growth of digital data, the United States is going to need an additional 140,000 to 190,000 analysts and more than 1.5 million managers capable of performing data analysis [4]. Additional calls have been made for more statisticians in the federal system, working in places such as the Bureau of Labor Statistics or the United States Census Bureau [5].

These pleas are not new, however; even in the early 1980's authors were writing about the need to make the field of statistics as a separate discipline [6] and recognizing the growing need to develop undergraduate and graduate programs to help develop professional statisticians who are ready to contribute to research and industry [7]. Professional statisticians are especially important today with the emergence of big data and the development and analysis of experiments.

There are several opinions about what really makes a good analyst, but one common trait is mastery of the methodologies of statistics. Thomas Willemain goes beyond just the technical skills, and emphasizes the importance of analytical creativity and analytical decision-making.[8] These decision-making skills are echoed by employers who do not want an analyst who "…blindly believes the output of a piece of software. A true analyst understands how the output was calculated, assumptions behind the output, and the inferences that can be made from it." [3] Creative thinking is what improves the quality and effectiveness of decision-making and the entire analysis [9]. Evans defines creativity, a key step in the decision-making and management sciences, as finding new relationships that were previously unknown. This creativity also fosters the different phases of data analysis especially when facing open-ended problems that are ill-structured. Employers desire analysts who do not strictly think statistically, but also have a business sense that will allow the analyst to use creativity to find relationships in solving a business problem, instead of just blindly doing analysis with no application for the client.[3] Once an analysis is complete, the information must be presented either in writing or in person to a client. Therefore, it is important for analysts to have the ability to communicate the results of an analysis in a way that can influence decisions. However, this is the portion of training that some analysts, like LaBarr[3] and Starbuck [10], think needs additional attention and training for young statisticians. Nonetheless, the best communication in the world will not help if these novices do not understand the basic theories and methodology behind the analysis; therefore all skills should be developed and practiced simultaneously throughout the entire learning process.

Further research and experimentation is needed to find what work styles and behaviors contribute to successful analysis. The skills necessary to be a successful analyst, which are technical, creativity, decision-making, and communication skills, will drive the analysts through the different phases of data analysis.

Experiment Background

In a pilot study conducted in spring 2012, volunteers from two statistics-based classes were given thirty minutes to analyze an open-ended problem and to report as many interesting ideas as possible. Participants submitted both a Word document with their findings, and a project document that recorded their menu choices when using the Minitab software. Initial concerns about the experiment are its small sample size and the similarity of the two courses in terms of the level of expertise required in each. However, this experiment has enabled the authors to start generating and testing hypotheses and working out ways to analyze the students' behaviors.

This pilot study tested the abilities of twenty-seven students in two undergraduate statistics classes at Rensselaer Polytechnic Institute to analyze an open-ended problem wherein students received a dataset and were asked to draw useful conclusions in thirty minutes. Every student in both classes was asked to fill out a form with some basic information, such as gender, year group (freshman, sophomore, etc.), major, and number of statistics courses taken. The experiment took place on the morning of Tuesday, May 8, 2012. The significance of the date is that this was the final class of the semester, listed in the syllabus as a review day. As a result, many students in each class chose not to attend class. Furthermore, since participation in the experiments was voluntary and end-of-term pressures were building, substantial numbers of students who came to class elected to skip the experiment and leave class early. Since subjects cannot be coerced into participation in an experiment, a possible result of these circumstances was a self-selection bias that created a subject pool different from the overall population of engineering undergraduates. However an analysis of participants versus nonparticipants showed that there was not a bias.

The two courses that were given the opportunity to participate were Modeling and Analysis in Uncertainty (MAU) and a Quality Control (QC) course. MAU is a required course for all engineering majors except Electrical and Computer Systems Engineering. For most enrolled students it is their first (and last) course in probability and statistics. It places a heavy emphasis on the ability to execute data analysis using the MINITAB® 14 Student Edition Statistical Computing Software.[1] The academic experience of students taking MAU is diverse because it includes students from all four classes, from freshmen to seniors. The potential participants came from nine different majors that include Aerospace Engineering, Biomedical Engineering, Chemical Engineering, Environmental Engineering, Mechanical Engineering, Material Engineering, Management Engineering, and Nuclear Engineering. The goal to for the students is to gain an appreciation and understanding of uncertainties and the conditions under which they occur within the context of the engineering problem-solving pedagogy of measurements, models, validation, and analysis. MAU will be called the Level 1 course for the remainder of this paper.

The second class involved in the experiment, Quality Control (QC), is an upper-class elective course for Industrial and Systems Engineers, with occasional enrollees from other engineering departments. For this particular class, thirty-eight out of thirty-nine possible participants had a major within the Industrial and Systems Engineering Department, one being a Mechanical Engineer. Most students in QC are juniors and seniors with more than one prior course in statistics. By the end of this course students should have the ability to identify, formulate and solve engineering problems, and model the stochastic nature of management systems and engineering relationships to the planning, organization, evaluation and control of human centered systems. The course places a heavy emphasis on control charting using Minitab 16. QC will be called the Level 2 course for the remainder of this paper.

At the start of the experiment, students provided various items of demographic information (e.g., gender, number of previous statistics courses). Later, final exam grades were added to the dataset. Each record was de-identified and given a random identification number based on the student's current course (e.g., MAU04 or QC12). Since the experiments were embedded within a normal course format, student subjects are unlikely to have perceived an extraordinary stress, which in any case should be less than that of a conventional course requirement (e.g., class assignments), particularly since performance on these exercises was not used in a calculation of the course grade. The experimental stimulus selected was the Web Visitors exercise (See Appendices A and B). It was chosen because of its relative simplicity, open-endedness, and compatibility with the data size limitation in the Minitab 14 Student Edition software used in the Level 1 Course. A sample of the data the participants analyzed is shown in Table 1.

[1] MINITAB® and all other trademarks and logos for the Company's products and services are the exclusive property of Minitab Inc. All other marks referenced remain the property of their respective owners. See minitab.com for more information.

Table 1.Sample of the data from the Web Visitors Exercise

DayOfWeek	Date	TotalVisits	TotalPageViews	AvgTime	BounceRate
Friday	January 14,2011	40	79	91	0.7
Saturday	January 15,2011	22	44	35	0.591
Sunday	January 16,2011	20	32	49	0.65
Monday	January 17,2011	58	194	213	0.466
Tuesday	January 18,2011	52	118	62	0.558
Wednesday	January 19,2011	47	120	117	0.553

To test for volunteer bias, tests were performed to analyze if there was any difference among students who chose to participate, those who chose not to participate, and those who were absent from the class that day. Of the 89 possible participants, only 29 chose to participate in the data analysis study, of which 10 were Level 2 students and 19 were Level 1 students. This framework served as the basis for comparing participants and potential participants in terms of gender, class year, number of previous statistics courses, and final exam grade. After an initial analysis, there were no differences between those who participated and those who did not participate in the experiment. This suggests that there was no bias in the volunteers and that generalization about the data could apply to the entire population of students in both the Level 1 and Level 2 course.

Data generated by experiment

The record of each menu choice made by the participants was extracted from the Minitab Project history file for further analysis. Two participants from the Level 1 class had corrupt data files, so their responses were eliminated from the remainder of the analysis. There were 45 different Minitab commands used, which were further organized into eight groups. The eight groups are listed in Table 2.

Table 2. Commands used during the Web Visitors Experiment, split into eight groups

Group	**Minitab Commands**
Descriptive Statistics	Describe, Mean, Statistics, St. Dev, Results
Quality Control Method	Cchart, Dcapa, Pareto, IMR Chart, MargPlot, XBarChart, XRChart
Visual Depiction	Boxplot, Chart, Dotplot, Histogram, Indplot, Lplot, Matrixplot, Plot,
Statistical Inference	ANOVA, Correlation, Onet, Two Sample, TwoWay, Pplot, Gsummary
Regression	Fitline, GReg, Ologistic Regress
Time Series Analysis	ACF, Trend, TSPlot
Codes/Calculations	Code, Let, Name, Notem, Numeric, Random
Subsetting	Sort, Split, Stack, Subset;, Unstack

To calculate the variety of steps used by each participant we used a method called normalized entropy. This metric quantifies the diversity of the choice of commands; which has been used in

several fields, to include a study about the genetics causing rheumatoid arthritis.[11] The equation to calculate Normalized entropy is:
:

$$(\text{Normalized})\text{Entropy} = \frac{-\sum \text{P}_\text{i} \times \log(\text{P}_\text{i})}{\log \text{N}}$$

where P_i is the probability of each event (or command) and N is the total number of categories, in this case it is eight, for each one of the command groups. The entropy is transformed with a logit transformation to linearize and normalize the variable:

$$\text{Logit8} = \log\left(\frac{\text{Entropy}}{1 - \text{Entropy}}\right)$$

The second response variable collected was the count of commands each participant used while analyzing the data. The new variable, Volume, is calculated using the Freeman and Tukey transformation for Poisson counts [12]:

$$Volume = \sqrt{Count} + \sqrt{Count + 1}$$

The factors of interest collected during this study include the following:

1) Class level – Categorical variable that is MAU (Level 1 course) or QC (Level 2 course)
2) Year Group – Categorical variable that includes freshman, sophomores, junior, seniors and graduate students. This group is further broken down so that Freshmen and Sophomores are one group, Juniors are one group, and Seniors and Grad Students for a third group
3) Final Exam Grade– This is a continuous variable that is changed into a categorical variable by classifying each score as High (above the mean score) or Low (below the mean score)
4) Gender – Categorical variable of Male or Female
5) Number of previous statistics courses taken (Courses) – This is a count that varied from a minimum of one previous statistics course to a maximum of six courses. This factor is coded into categories by dividing into groups of those with only one course, those with two courses, and those with three or more courses.

Analysis of overall style

The initial hypothesis was that there would be a distinct difference between the students in the Level 1 course versus students in the Level 2 course. Using a one-way analysis of variance (ANOVA) resulted in no significant main effects; however, a two-way ANOVA had significant interaction effects for both Logit8 and Volume when Class was the blocking factor. For the variety (Logit8), there was a marginally significant interaction between the class level and the final exam grade (p-value = 0.054). This translates into a higher value for the Logit8, or more variety of steps taken, when the exam grade was below the mean and the student was in the Level 2 course. The Volume, transformed by a logarithm for linearization, had a significant interaction between gender and class (p-value =0.084): the volume of steps decreased for males in the Level 2 course. After several iterations of testing different combinations of factors there were still no significant main effects.

Although the initial hypothesis for this research was that we were going to see a clear difference between the QC students and the MAU students, a far more complicated interaction of several effects are affecting variety since main effects are not significant, but interaction effects are significant. So far, the data is suggesting that the class may not be a key factor. Instead, there may be high-level thinkers and low-level thinkers in each class that cannot clearly be defined by a certain attribute that was collected during the study.

The problem with both of these ANOVAs is that there is multicollinearity. Multicollinearity occurs when factors that are correlated. One of the main problems caused by multicollinearity is that the least squares estimates of the factors will have inflated variances. Most software packages are able to check for a Variance Inflation Factor (VIF) that will detect multicollinearity. Anytime there is a VIF greater than 1.0, there is multicollinearity [12]. In the above ANOVAs for both volume and variety, nearly all the factors had a VIF ranging from 3.0 to infinity. For example, someone may be able to approximate how many statistics classes a student may have taken based on that individual's year group. Additionally, there is a correlation between the year groups and the level class (Level 1 or Level 2), since a student in the Level 2 class has prerequisites that include taking the Level 1 course. A.C. Harvey wrote that, "Multicollinearity is a problem of degree rather than kind" meaning, there will usually be multicollinearity, however, the degree to which it exists will vary. Harvey further specified that multicollinearity does not mean there is evidence of model misspecification [13]. Michael Crawley also emphasizes that if a factor is statistically significant or close to being statistically significant, then it is worth including in a larger scale experiment with repetitions with proper blocking [14]. A well designed experiment based on this pilot study will have balanced, orthogonal data, and factors will be fixed at two or three levels. Despite the multicollinearity from the specified models for volume and variety, there is predictive power in the models created. The two plots in Figure 1 show the fitted values for each model used for the above ANOVA models against the actual values for both the Volume and Variety metrics. There seems to be some confounded, or lurking, variable that was not collected during the study that may better identify experience level other than year group or class level. The authors' hypothesis is that experience from either jobs or internships may have had an impact on how well the subjects performed on the tests and behavior that resulted.

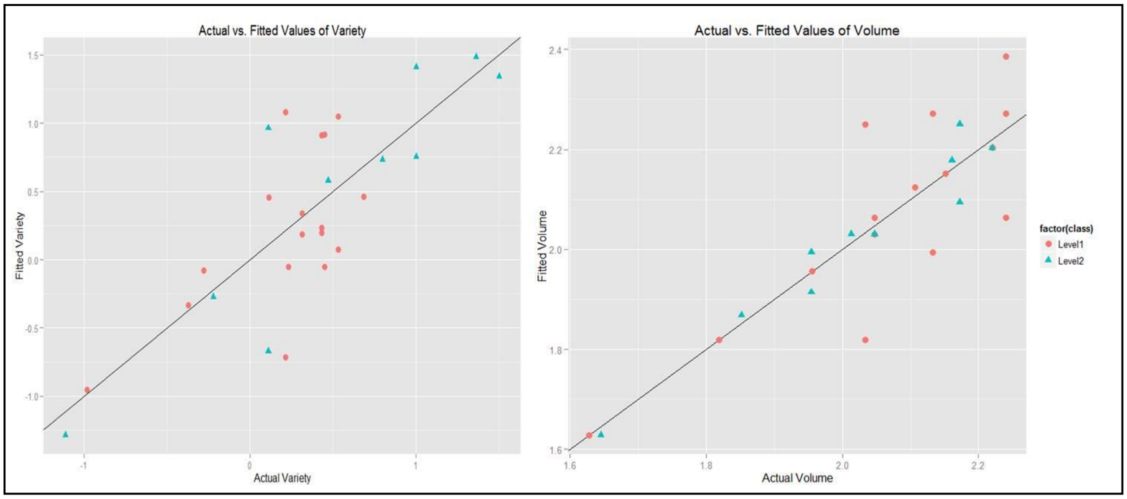
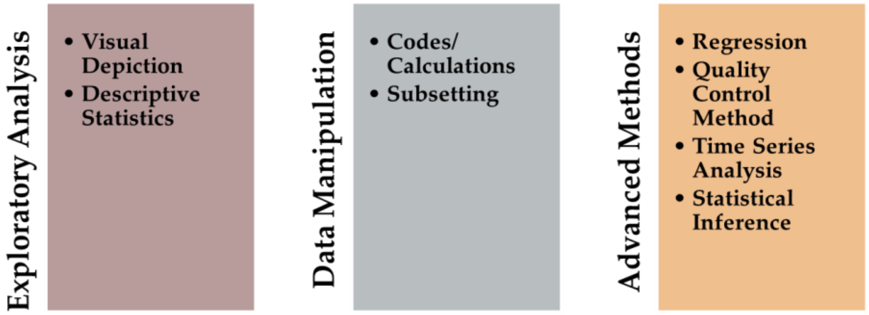


Figure 1. Scatter plots of the actual values versus the fitted values of the ANOVA models show that there is predictive power in the models despite the multicollinearity

Analysis of group behavior by sequential analysis

The purpose of the sequential analysis is to describe the sequence of data analysis decisions made by each subject. This method uses the Markov Model Paradigm as a framework for describing the behavior of subjects, or groups of subjects. The sequential analysis takes each group outlined in Table 2 , which lists all the commands used in each participant's analysis, and further breaks them down in three categories of analysis that include exploratory analysis, advanced methods, and data manipulation. Figure 2 gives a breakdown of which commands fall within each type of analysis.



Exploratory Analysis

- Visual Depiction
- Descriptive Statistics

Data Manipulation

- Codes/ Calculations
- Subsetting

Advanced Methods

- Regression
- Quality Control Method
- Time Series Analysis
- Statistical Inference

Figure 2. Three types of analysis steps that a subject may take when analyzing data

Each participant also has a certain probability of moving from one-step to the next. For example, someone can start the exercise and go to an exploratory step, move to data manipulation by creating a new variable, then try an advanced method like regression analysis. After thirty minutes of going through these steps, they will end the exercise. The participants can also stay on the same type of step for two or three commands and can move in any order among the three types of analysis steps. Figure 3 is a flow chart of the different steps a subject could take during the thirty-minute exercise.

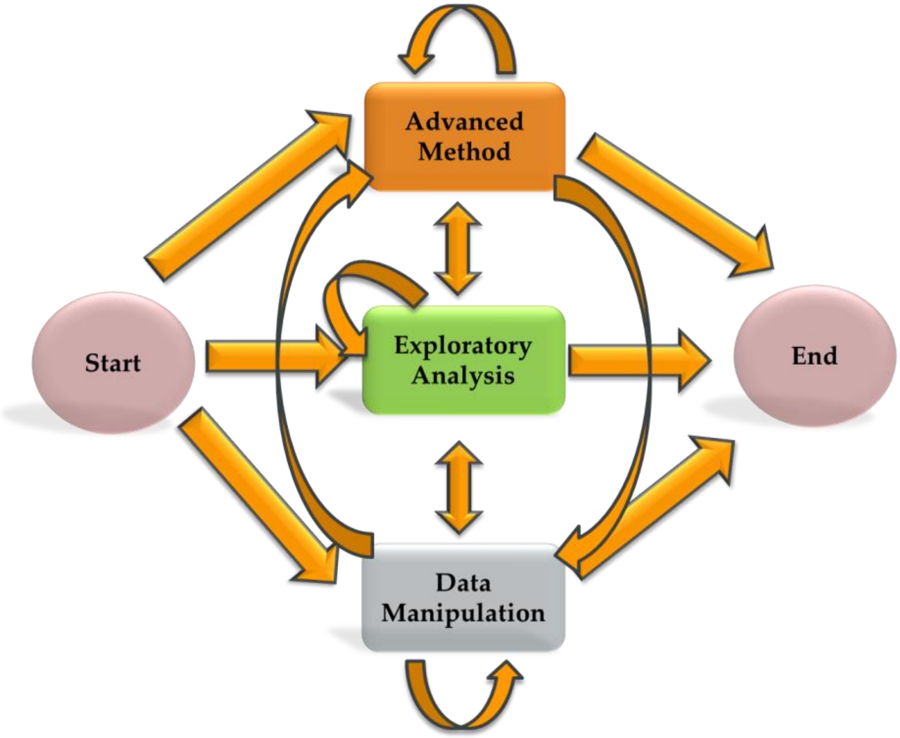


Figure 3. Flow chart showing the possible sequence of steps a participant could take during the exercise

Using this Markov process, the subjects were divided into groups based on descriptions. The counts were then transformed into transition probabilities. However, looking at one Markov Chain may not be good enough. The next goal after getting the counts of each transition is to divide the students into groups – for example, comparing the Markov chains of the Level 1 students' to those of the Level 2 students. A chi-squared test for independence can determine if there is a statistically significant difference between two transition probability matrices. The following formula implements a statistical test to compare two Markov chains of an *s*-by-*s* matrix, where s is the number of different steps (there are five in this analysis) [15].

$$\chi^2 = \sum_{ij} \frac{\left[f_{ij} - f_i \frac{f_{ij}+g_{ij}}{f_i+g_i}\right]^2}{f_i \frac{f_{ij}+g_{ij}}{f_i+g_i}} + \sum_{ij} \frac{\left[g_{ij} - g_i \frac{f_{ij}+g_{ij}}{f_i+g_i}\right]^2}{g_i \frac{f_{ij}+g_{ij}}{f_i+g_i}} = \sum_{ij} \frac{f_i g_i}{f_{ij}+g_{ij}} \left(\frac{f_{ij}}{f_i} - \frac{g_{ij}}{g_i}\right)^2$$

Let f_{ij} and g_{ij} be the transition counts of two samples, independent of each other, from Markov chains with transition matrices p_{ij} and q_{ij}. The null hypothesis is that the probability matrices are equal, with the alternative hypothesis that the matrices are not equal. The degrees of freedom are equal to $s(s-1)$. For is experiment, there are 20 degrees of freedom. Degrees of freedom are "The number of independent comparisons that can be made among the elements of a sample." [16] This comparison was done for the following groups

1) Class Level – Level 1 versus Level 2
2) Gender – Male versus Female
3) Final Exam Grade – Above the Mean versus Below the Mean
4) Number of Statistics Courses Taken – Less than or equal to two versus Greater than two
5) Year Group – Freshmen/Sophomores/Juniors versus Seniors/Graduate Students

Of all these groups tested, the only comparison that yielded statistically significant results was the comparison by year group. First, visualization of the transition probabilities can be seen in a butterfly chart, where there is a side-by-side comparison of the probabilities of each step. Figure 4 shows a butterfly plot of the probabilities (x100) for groups of freshmen, sophomores, and juniors versus seniors and graduate students. Since the plot is asymmetrical, it appears that there may be a difference between the two groups. The chi-squared value for this comparison is 40.64, which is larger than the critical value of 31.41. The p-value for this comparison is 0.004, so the null hypothesis is rejected.

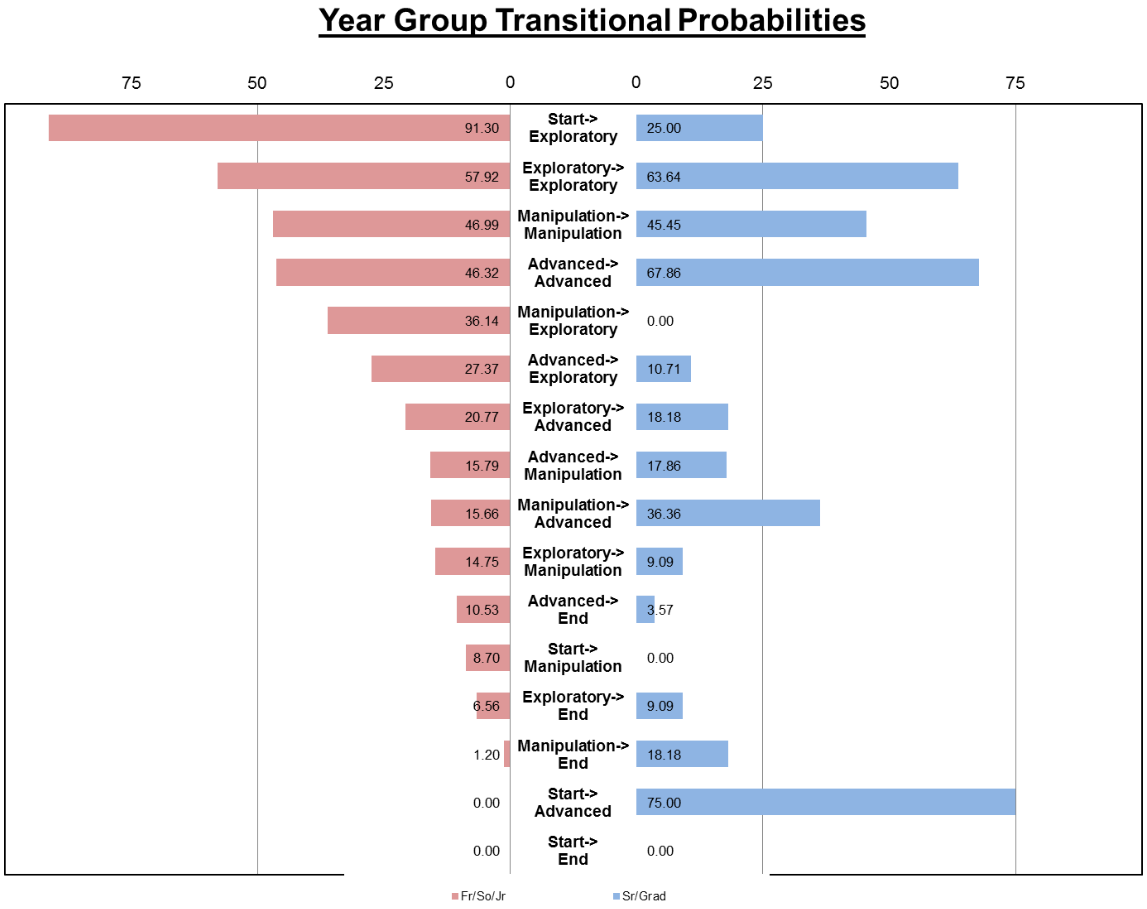


Figure 4. The butterfly plot of younger versus older students

The problem with this particular chi-squared analysis is that it does not meet one of the assumptions, which requires every cell to have an expected count of five. This is what makes creating a full-scale experiment so crucial to this research. This will ensure that the analysis that is used for these types of problems will meet all the proper assumptions for the results to be accurate.

Conclusion

Preliminary conclusions from this pilot study show that there are correlations between the experience level and the behavior of each participant and that there is a certain amount of prediction that can be obtained from the models used in this article. There are also suggestions that students that are more senior tend to skip the initial exploratory steps and go right to advanced methods. This may be because the older students know more advanced techniques and fail to use the basics to determine the general information that the data visualizations may provide. The data visualizations allow an individual to see what types of analysis could be performed with the data and allow for even more creativity when doing a confirmatory analysis. John Tukey, well known for both exploratory and confirmatory analysis, argued that although students need to learn both, exploratory is more important and should be taught first and foremost, which will lead individuals to perform a more complete and sufficient confirmatory analysis.[17]

Besides recommendations for a larger and balanced experiment with a clearly defined experience level, there are several other lessons that can be taken from this pilot study and applied in the classroom. During the analysis, there were no clusters that were distinguishable by descriptive attributes such as gender, year group, course, previous courses taken, or even by the final exam grade. The conclusion was that there are high-level and low-level thinkers among all these different groups. Some students struggle with all aspects of being an analyst. There are also some students who master technical skills and can use these skills on the data provided, but have no idea what the results mean, or what steps they should take next based on the previous results. The high-level thinkers are able to see the results of an analysis, assess information, and make new connections that can help make a decision. High-level thinkers do not always have the advanced and elaborate skills, but understand what statistical tests mean and can provide valid feedback. Besides being able to recognize those high-level thinkers, it is important that teachers recognize that whatever the level of the class, there will be high-level and low-level thinkers.

The second recommendation is practice, practice, practice! But not just practice, perfect practice! This practice at handling open-ended complex problems will only benefit these novice analysts in the end, preparing them for the complexities the students may see in their future careers. Giving the students more opportunities to practice their skills on these large, complex open-ended problems will allow them to refine their technical skills and learn from their mistakes. An after action review (AAR) will help advance any student willing to participate. These exercises will challenge the students to “think outside the box” when solving problems and allow them to apply these methods across multiple fields. More importantly, those students who have experience in analytics with industry should share their experiences. This gives the students the “so what” factor, where they can hear first-hand accounts of the methods learned in the classroom being applied in operational environments by their peers.

There is still more work to be done to find out what behaviors in novices lead to success or failure when given an open ended problem. The next step is to explore ways to qualitatively analyze the subjects' responses to describe the subjects' analytic judgments and creativity to better draw evidence-based recommendations for instructional reform. Additionally, an experiment that analyzes expert analyst behaviors would provide a reference for novice performance. Completing the next phase will allow a rating of each individual response and comparison of the response to each individual's behavioral patterns which were explored in this article using Markov chains. The lessons learned from this experience will translate into a larger scale experiment with repetitions to create a balanced design. This research could possibly hold the key to educational reform and finding out what qualities, behaviors, and techniques will make the most successful analysts for the future.

Bibliography

[1] J. Harris, E. Craig and H. Egan, "Counting on Analytical Talent," Accenture, 2010.

[2] M. Rappa, Master of Science in Analytics: Goals, Learning, Outcomes, 2013.

[3] A. LaBarr, "The Emergence of Analytics in the World of Business Decisionmaking," AMSTATNEWS, pp. 15-16, September 2012.

[4] J. Manyika, M. Chui, B. Brown and J. Bughin, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011.

[5] R. Little and T. Wright, "Wanted! Statsticians in the Federal Statistical System," AMSTATNEWS, pp. 10-11, November 2012.

[6] P. D. Minton, "The Visibility of Statistics as a Discipline," The American Statistician, vol. 37, no. 4, pp. 284-289, 1983.

[7] R. Bradley, "The Future of Statistics as a Discipline," Journal of the American Statistical Association, vol. 77, no. 377, pp. 1-10, 1982.

[8] T. R. Willemain, "Insights on Modeling from a Dozen Experts," Operations Research, vol. 42, no. 2, pp. 213-222, 1994.

[9] J. R. Evans, Creative Thinking in the Decision and Management Science, Cincinnati, Ohio: South-Western Publishing Co., 1991.

[10] R. Starbuck, "Communication, Influence Keys to Success in Statistics," AMSTATNEWS, pp. 10-12, Novermber 2012.

[11] P. F. Velleman and D. C. Hoaglin, Applications, Basics, and Computing of Exploratory Data Analysis, Boston: Duxbury Press, 1981.

[12] E. R. Mansfield and B. P. Helms, "Detecting Multicollinearity," The American Statistician, vol. 36, no. 3, Part 1, pp. 158-160, 1982.

[13] A. C. Harvey, "Some Comments on Multicollinearity in Regression," Journal of the Royal Statisical Society, Series C (Applied Statistics), vol. 26, no. 2, pp. 188-191, 1977.

[14] M. J. Crawley, The R Book, West Sussex, England: John Wiley & Sons, Ltd., 2007.

[15] P. Billingsley, "Statistical Methods in Markov Chains," The Annals of Mathematical Statistics, vol. 32, no. 1, pp. 12-40, 1961.

[16] D. C. Montgomery and G. C. Runger, Applied Statistics and Probability for Engineers, New York City: John Wiley & Sons, Inc., 2003.

[17] J. W. Tukey, "We Need Both Exploratory and Confirmatory," The American Statistician, vol. 34, no. 1, pp. 23-25, 1980.

[18] P. F. Velleman and D. C. Hoaglin, Applications, Basics, and Computing of Exploratory Data Analysis, Boston: Duxbury Press, 1981.

[19] M. J. Crawley, The R Book, West Sussex, England: John Wiley & Sons, Ltd, 2007.

[20] I. W. Witten, E. Frank and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Burlington, MA: Morgan Kaufmann, 2011.

[21] T. R. Willemain, "Model Formulation: What Experts Think About and When," Operations Research, vol. 43, no. 6, pp. 916-932, 1995.

[22] T. R. Willemain and S. P. Powell, "How Novices Formulate Models. Part II: A Quantitative Description of Behavior," Journal of the Operations Research Society, vol. 58, no. 10, pp. 1271-1283, 2007.
[23] L. B. Waisel, "The Cognitive Role of Visualization in Modeling," Rensselaer Polytechic Institute, Troy, NY, 1998.
[24] W. N. Venables and B. D. Ripley, Modern Applied Statistics with S, 4th ed., New York: Springer, 2002.
[25] J. W. Tukey, "We Need Both Exploratory and Confirmatory," The American Statistician, vol. 34, no. 1, pp. 23-25, 1980.
[26] M. A. Shelly, "Exploratory Data Analysis: Data Visualization or Torture?," Infection Control and Hospital Epidemiology, vol. 17, no. 9, pp. 605-612, 1996.
[27] G. Runger, "Introduction to Data Mining," Phoenix, 2012.
[28] S. G. Powell and T. R. Willemain, "How Novices Formulate Models. Part I: Qualitative Insights and Implications for Teaching," Journal of the Operational Research Society, vol. 58, no. 8, pp. 983-995, 2006.
[29] D. C. Montgomery and G. C. Runger, Applied Statistics and Probability for Engineers, 3rd ed., New York City, New York: John Wiley & Sons, Inc., 2003.
[30] D. Jawaheer, W. Li, R. R. Graham, W. Chen and A. Damle, "Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis.," American Journal of Human Genetics, vol. 71, no. 2, pp. 585-594, 2002.
[31] D. C. Hoaglin, F. Mosteller and J. W. Tukey, Understanding Robust and Exploratory Data Analysis, New York: John Wiley & Sons, Inc., 1983.
[32] D. C. Hoaglin, "John W. Tukey and Data Analysis," Statistical Science, vol. 18, no. 3, pp. 311-318, 2003.
[33] F. S. Hillier and G. J. Lieberman, "Markov Chains," in Introduction to Operations Research, 9th ed., D. B. Hash, Ed., New York, McGraw-Hill, 2010, pp. 723-758.
[34] D. Hand, H. Mannila and P. Smyth, Principles of Data Mining, Boston: MIT Press, 2001.
[35] D. Gao, Y.-X. Zhang and Y.-H. Zhao, "Random Forest Algorithm for Classification of Multiwavelength Data," Research in Astronomy and Astrophysics, vol. 9, no. 2, pp. 220-226, 2009.
[36] J. Friedman, T. Hastie and R. Tibishirani, The Elements of Statistical Learning, 2nd ed., New York: Springer Science + Business Media, LLC, 2009.
[37] T. D. Cook and D. T. Campbell, Quasi-Experimentation: Design and Analysis Issues for Field Settings, Chicago: Rand McNally Publishing Company, 1979.
[38] C. Chatfield, "The Initial Examination of Data," Journal of the Royal Statistical Society, Series A (General), vol. 148, no. 3, pp. 214-253, 1985.
[39] D. T. Campbell and J. C. Stanley, Experimental and Quasi-experimental Designs for Research, Boston: Houghton Mifflin Company, 1963.
[40] J. R. Buxton, "Some Comments on the Use of Response Variable Transformations in Empirical Modeling," Journal of the Royal Statistical Society, Series C (Applied Statistics), vol. 40, no. 3, pp. 391-400, 1991.
[41] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[42] P. Billingsley, "Statistical Methods in Markov Chains," The Annals of Mathematical Statistics, vol. 32, no. 1, pp. 12-40, 1961.
[43] Minitab, Inc., Minitab Statistical Software, State College, PA: Minitab, Inc., 2013.
[44] A. A. Schoenfeld, Mathematical Problem Solving, New York: Academic Press, 1985.

Appendix

A. Web Visitors Exercise Instructions

Instructions for voluntary experiment in data analysis

Step 1. Read and, if you choose, sign the attached consent form. Note that you need not participate in this experiment. If you do choose to begin the experiment, you can stop and walk away without penalty at any time.

Step 2. Provide the following background information:

Your name (note that all analyses will be anonymized): ___________________________________

Circle your course: ENGR2600 MAU ISYE4230 Quality Control

Circle your year: Freshman Sophomore Junior Senior

Indicate the number of statistics courses you have ever taken in your life: _______________

Step 3. Start Minitab and read in the data file *WebVisitors.MTW* from the course web site.

Step 4. Open the problem statement *WebVisitors.doc* from the same course web site. Read the problem statement, explore the data, then type your responses directly into the Word file. You will have up to 30 minutes to complete the exercise, but you may stop earlier if you wish. Please do not refer to any course notes, textbooks, or other references, and do not collaborate.

Step 5. Save your Minitab file as a project file (not a worksheet!) with your name on it, e.g., *TomWillemain.MPJ*.

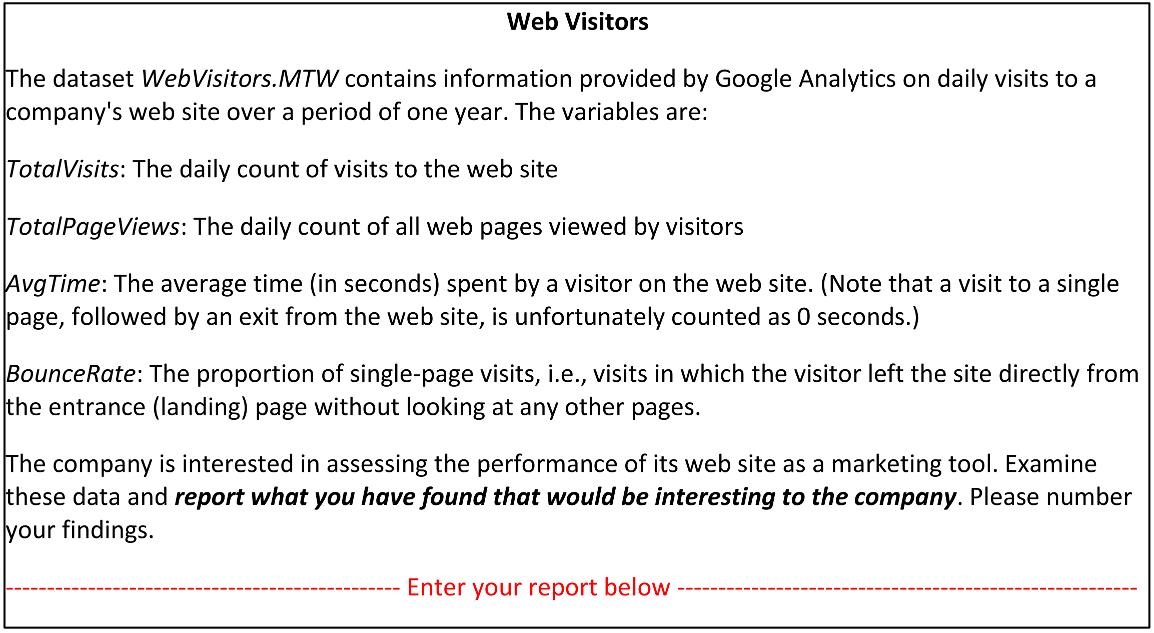
Step 6. Save your answers as a Word file with your name on it, e.g., *TomWillemain.doc*.

Step 7. Email both your MPJ project file and Word file to willet@rpi.edu.

Step 8. Turn this paper with the signed consent form back to Prof. Willemain.

Thank you for your participation in this research project. We hope your data will help evolve our courses to better educate engineers in the art of data analysis.

B. Prompt for Web Visitors exercise



Web Visitors

The dataset *WebVisitors.MTW* contains information provided by Google Analytics on daily visits to a company's web site over a period of one year. The variables are:

TotalVisits: The daily count of visits to the web site

TotalPageViews: The daily count of all web pages viewed by visitors

AvgTime: The average time (in seconds) spent by a visitor on the web site. (Note that a visit to a single page, followed by an exit from the web site, is unfortunately counted as 0 seconds.)

BounceRate: The proportion of single-page visits, i.e., visits in which the visitor left the site directly from the entrance (landing) page without looking at any other pages.

The company is interested in assessing the performance of its web site as a marketing tool. Examine these data and ***report what you have found that would be interesting to the company***. Please number your findings.

-- Enter your report below --

C. List of Minitab commands and definitions

The following is a list of the Minitab commands used by participants in this pilot study (Minitab, Inc., 2013).[3]

ACF - Autocorrelation Function; calculates the correlation between observations of a time series separated by k time units.

ANOVA - Analysis of Variance; Tests the hypothesis that the means of two or more populations are equal. ANOVAs evaluate the importance of one or more factors by comparing the response variable means at the different factor levels. The null hypothesis states that all population means (factor level means) are equal while the alternative hypothesis states that at least one is different.

Boxplot - Boxplot; A graphical summary of the distribution of a sample that shows its shape, central tendency, and variability

CChart - C-Chart; Tracks the number of defects and detects the presence of special causes. Each entry in the specified column contains the number of defects for one subgroup, assumed to have come from a Poisson distribution with parameter µ.

Chart - Bar Chart; Used to visually compare bar heights of category measures. Bar charts can be made of category tallies, of different statistics by categories, or of summary values. The height of the bars signifies the magnitude of the values.

[3] Portions of information contained in this publication/book are printed with permission of Minitab Inc. All such material remains the exclusive property and copyright of Minitab Inc. All rights reserved.

Code - Coding or Calculation Done; change a value or set of values to new values

Correlation - Correlation Coefficient; Calculates the Pearson product moment correlation coefficient between each pair of variables you list. The Pearson product moment correlation coefficient can be used to measure the degree of linear relationship between two variables.

Dcapa - Individual Distribution Identification; Use to evaluate the optimal distribution for your data based on the probability plots and goodness-of-fit tests prior to conducting a capability analysis study.

Describe - Descriptive Statistics; produce statistics for each column or for subsets within a column.

Dotplot - Dot Plot; plots each observation as a dot along a number line (x-axis)

Fitline - Regression; performs regression with linear and polynomial (second or third order) terms, if requested, of a single predictor variable and plots a regression line through the data, on the actual or log10 scale

GReg - General Regression, performs least squares regression when you have continuous and categorical predictors or a polynomial model.

Gsummary - Graphical Summary; The graphical summary includes four graphs: histogram of data with an overlaid normal curve, boxplot, 95% confidence intervals for mean, and 95% confidence intervals for the median. The graphical summary also displays the Anderson-Darling Normality Test statistics, Descriptive statistics, confidence intervals for mean, standard deviation, and the median

Histogram - Histogram; A graph used to assess the shape and spread of continuous sample data.

IMRChart - Individuals- Moving Range Chart; Plots individual observations (I chart) and moving ranges (MR chart) over time for variables data. Use this combination chart to monitor process center and variation when it is difficult or impossible to group measurements into subgroups. This occurs when measurements are expensive, production volume is low, or products have a long cycle time.

Indplot - Individuals Plot; Use to assess and compare sample distributions through individual data values, with optional grouping by categorical variables.

Let - Calculation done using a formula in the calculator function; allows arithmetic operations, comparison operations, logical operations, functions, and column operations.

Lplot - Line Plot; Use to compare response patterns for two or more groups.

Margplot - Marginal Plot; Use to assess the relationship between two variables and examine their distributions. A marginal plot is a scatterplot with histograms, boxplots, or dotplots of the x-

and y-variables in the margins. This two-in-one graph allows you to compare individual variables and their distributions at the same time.

Matrixplot - Matrix Scatter Plot; Use to assess the relationship among several pairs of variables at once. A matrix plot is an array of individual scatterplots. There are two types: matrix of plots and each Y versus each X.

Mean - arithmetic average, is the sum of all the observations divided by the number of observations.

Name - Names a new column

NOTE - Data Table to changed or updated

Numeric - Numeric Coding; allows you to change your data from text or date/time to numeric.

Ologistic - Ordinal Logistical Regression; perform logistic regression on an ordinal response variable. Ordinal variables are categorical variables that have three or more possible levels with a natural ordering, such as strongly disagree, disagree, neutral, agree, and strongly agree.

Onet - Student's t-test; Use 1-Sample t to compute a confidence interval and perform a hypothesis test of the mean when the population standard deviation, is unknown

Pareto - Pareto Chart; type of bar chart in which the horizontal axis represents attributes of interest, rather than a continuous scale. These attributes are often "defects." By ordering the bars from largest to smallest, a Pareto chart can help you determine which of the defects comprise the "vital few" and which are the "trivial many." A cumulative percentage line helps you judge the added contribution of each category.

Plot - Scatterplot; Use to explore the potential relationship between a pair of continuous variables. When you create a Scatterplot, you usually display the response variable on the y-axis and the predictor variable on the x-axis for each observation.

PPlot - Probability Plot; Use to evaluate the fit of a distribution to your data, estimate percentiles, and compare different sample distributions. Plots each value vs. the percentage of values in the sample that are less than or equal to it, along a fitted distribution line.

Random - Random Variable assigned; commands used to obtain random samples, generate random data, and calculate probabilities for different distributions.

Regress - Regression; Use this procedure for fitting general least squares models, storing regression statistics, examining residual diagnostics, generating point estimates, generating prediction and confidence intervals, and performing lack-of-fit tests.

Results - displays the text output of your analysis, such as statistical test results and related notes or error messages

Sort - Order of number/character sorted by ascending or descending order

Split; - splits, or unstacks, the active worksheet into two or more new worksheets based on one or more "By" variables.

Stack - moves data from two or more columns to one longer column within your current worksheet or to a new worksheet.

Statistics - Stores descriptive statistics

StDev - Standard Deviation; estimates the "average" distance of the individual observations from the mean

Subset; - Used to copy specified rows from the active worksheet to a new worksheet

Trend - Trend Analysis of a Time Series Plot; Trend analysis fits a general trend model to time series data and provides forecasts. Choose among the linear, quadratic, exponential growth or decay, and S-curve models.

TSPlot - Time Series Plot; A plot of a sequence of observations over regularly spaced intervals of time.

TwoSample - Two Sample t-test; Use 2-Sample t to perform a hypothesis test and compute a confidence interval of the difference between two population means when the population standard deviations, are unknown.

Twoway – Two-way ANOVA; A two-way analysis of variance tests the equality of population means when classification of treatments is by two variables or factors

Unstack - splits the contents of a stacked column or block of columns into two or more shorter columns within your current worksheet or copy the split columns to a new worksheet.

XbarChart - A control chart of subgroup means. You can use X charts to track the process level and detect the presence of special causes.

XRChart - Displays a control chart for subgroup means (an X chart) and a control chart for subgroup ranges (an R chart) in the same graph window.