BENEFITS OF USING TECHNOLOGY RESOURCES WHILE TEACHING THE CHI-SQUARE TEST FOR INDEPENDENCE/DEPENDENCE

Ramon Gomez
Florida International University
813 NW 133rd Court
Miami, FL 33182
gomezra@fiu.edu

1. Introduction

Chi-square is a statistical test commonly used to compare observed frequencies against expected frequencies according to a specified hypothesis. It is taught in a variety of both undergraduate and graduate Statistics courses at university level. Chi-square actually designates a family of tests based on a common sampling distribution for their test statistics. Pearson's chi-square test is the best-known of this family and one important of its applications can be found when the dependence of two categorical variables is hypothesized. In this particular case the null hypothesis states that two random variables are independent (not associated) against the research or alternative hypothesis that they are dependent (associated). Sample data for the two categorical variables is collected and observed frequencies are organized in a matrix format, known as a contingency table. Students' understanding of this statistical test is usually conditioned by a proper interpretation of the statistical concepts of independence and dependence of random variables. Moreover, the test may also involve long and tedious computations, depending on the size of the contingency table, creating a distraction for students' comprehension.

It is widely accepted that the appropriate use of technology contributes to students' motivation and understanding of statistics. Using technology can make college teaching of statistics more effective as it improves the quality of instruction, encourages students' active learning, and provides them with psychological incentives (Garfield, 1995; Higazi, 2002). A committee created by the American Statistical Association produced in 2005 the Guidelines for Assessment and Instruction in Statistics Education (GAISE), which recommended the use of technology resources for statistics courses at university level.

In this regard, PowerPoint has permeated all aspects of college teaching as a presentation technology resource. This success has been associated with the appropriate use of text, images, and graphics. Furthermore, the use of statistical software provides students with a tool that enhances their learning experience, by allowing them to engage the contents actively and analytically. The Statistical Package for Social Sciences (SPSS) has been identified as one of the most commonly used packages at college level (Hulsizer & Woolf, 2009). The use of PowerPoint and statistical software in undergraduate courses has been previously described in the literature as a facilitator of learning statistics (Lock, 2005; Cryer, 2005; Gomez, 2010; Fernando, 2012; Robinson & Kimmel, 2013).

This paper summarizes the present author's experience with technology resources while teaching the chi-square test for independence/dependence of two categorical variables to undergraduate students at Florida International University during recent years. The benefits of using Power Point and the SPSS software are discussed.

2. Method

2.1 Context

Florida International University is a public institution located in Miami (USA) with current enrollment near 48,000 students. The Introduction to Statistics I and II courses (STA-2122/3123) are requirements for psychology majors and prerequisites for the 'Research Methods' class. The Statistics I and II courses (STA3111/3112) are intended for science students. The STA3193/3194 sequence is especially designed for selected biology scholars, enrolled in a special program, that use computers in the classroom. The STA-3112, STA-3123, and STA-3194 are three credit-hours classes covering a range of topics: hypothesis testing based on one and two samples, analysis of variance models, regression analysis, categorical data analysis, and non-parametric statistics. They are second statistics courses having as a prerequisite a preceding class that includes descriptive statistics, probability and hypothesis testing based on a single sample.

The textbook for both STA-3112 and STA-3123 is "Statistics" by McClave and Sincich (2013) that emphasizes inference methods and stresses the development of statistical thinking. It includes many proposed exercises for which real data is utilized to illustrate statistical applications. These courses encompassed contents from chapter 8 to 14, where chapter 13 is devoted to the Categorical Data Analysis. In particular, the Chi-square test for Independence-Dependence of two categorical variables is presented in section 13.3. The textbook for STA-3194 is "Biostatistics" by Daniel and Cross (2013) which requires mid-level mathematical prerequisites. It includes exercises focused on applications to the health sciences and introduces the Chi-square tests in chapter 12.

The present author has taught the STA-3112 and STA-3194 courses each spring term between the 2010 and 2013 years integrating PowerPoint and the SPSS software. Typical enrollment for STA-3112 is 55 students seating in a classroom with a computer projection system. The STA-3194 course is taught in a computer lab with classes not exceeding 25 scholars.

2.2 Design and organization

The traditional approach to teaching Statistics consists of using a board during lectures, a textbook as a reference, a calculator for computations and, more recently, supplementary material posted on a website. Two technology additions were integrated in our courses between 2010 and 2013: the daily use of PowerPoint for lectures as well as statistical software (SPSS) for data computations and analyses. This integration allowed for more class time to discuss statistical concepts and applications. Thus, a broader conceptual

understanding of the material was promoted as well as active learning in the classroom. Consequently, an interactive learning environment was generated where students had the opportunity to develop an increased rank of statistical literacy and reasoning.

The PowerPoint presentations, developed by the present author for this course were structured with the goal of increasing student participation during lectures in addition to satisfying the needs of scholars with a more visual oriented learning style. A course pack comprising the PowerPoint slides for all lectures was made available to the students at the beginning of the course, eliminating the hassle of frantic note-taking in class. The incorporation of SPSS involved the use of computer output to illustrate various topics allowing a more effective discussion of the statistical concepts and applications. The Instructor's style for lectures consisted of projecting slides from a presenter device and discussing their content with the students while moving around the classroom. This approach allowed a more direct interaction with learners.

2.3 Teaching approach for the chi-square test of independence/dependence

While teaching the Chi-square test of independence/dependence for two categorical variables, the present author organized the discussion in several steps: a) Introduction of contingency tables b) Interpretation of the independence/dependence concepts, c) Hypothesis testing procedure, d) Test statistic computation, and e) Analysis of the results.

To illustrate this organization, a typical exercise used in class, taken from Biostatistics by Daniel and Cross (2013), is presented here. The problem describes a study whose objective is to determine the association or dependence between two categorical variables, Ethnicity and Hemoglobin level, for a population of children under 15 years in the inner-city area of a large city. Three Ethnic groups were considered (A, B, C) and Hemoglobin levels were classified in three categories (Low, Medium, High). A sample of 695 children was selected from the given population for this study. The contingency table summarizing the observed frequency data is presented in Table 1.

Table 1: Contingency table with observed frequencies

	Hemoglobin Level			
Group	High	Medium	Low	Subtotal
A	80	100	20	200
B	99	190	96	385
C	70	30	10	110
Overall	249	320	126	**n = 695**

Source: Biostatistics by Daniel & Cross (2013)

A discussion involving the information provided by this contingency table is conducted. After this preamble the research hypothesis is introduced as follows: Ethnicity and Hemoglobin level are related or associated. This statement can be translated as "The categorical variables Ethnicity and Hemoglobin level are not independent" (they are dependent). A more tangible interpretation will consist of the statement that the distribution of Hemoglobin levels is not the same for every Ethnic group. The observed frequency data may not be sufficiently clear to support this statement; however, a clearer picture is obtained when the percent frequency for each cell relative to the Ethnic group subtotal is shown (Table 2).

Table 2: Contingency table with observed % frequencies

	Hemoglobin Level			
Group	High	Medium	Low	Subtotal
A	40.0	50.0	10.0	100%
B	25.7	49.4	24.9	100%
C	63.6	27.3	9.1	100%
Overall	35.8	46.1	18.1	100%

Students can notice that the row pattern for these percentages is quite different among the three ethnic groups (or that each of them differs from the pattern for the combined data) suggesting that there is some evidence supporting the research hypothesis. PowerPoint is extremely helpful for these discussions on Tables 1 and 2 by illustrating the numerical interpretation of the problem.

Following this exploratory data analysis the hypothesis testing procedure for this problem is set-up. A discussion on the hypotheses is conducted where the null hypothesis Ho states that two categorical random variables are independent (not associated) against the research or alternative hypothesis Ha that they are dependent (associated). Next step consists of determining the strength of the statistical evidence collected. To that end the Chi-square test statistic and associated p-value must be calculated. The Chi-square test statistic measures the overall deviation of the observed frequencies relative to the null hypothesis Ho (no association/relation between the two categorical variables), and is given by the formula

$$X^2 = \Sigma\, (O - E)^2 / E$$

with

O = Observed frequency
E = Expected frequency (assuming no relation/association between the variables)
Σ = Summation involving all cells

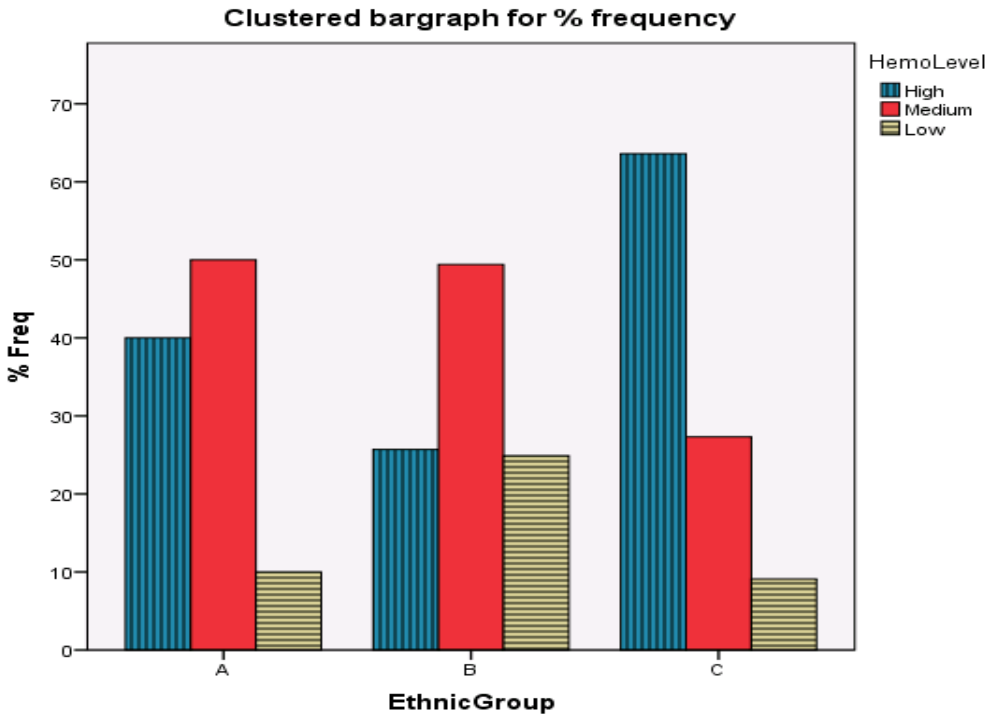
where the expected frequency for each cell is calculated using the observed frequencies from Table 1 as follows:

E = (row total) (column total) / overall sample size

The computation of this Chi-square test statistic using a hand calculator can result in a long and tedious process, even for a 3x3 contingency table. This old computational approach distracts students from the focus of the problem and may compromise their understanding. The present author uses the SPSS software as a substitute for this step. Thus, the SPSS output, previously ran by the Instructor, is inserted as part of the PowerPoint presentation. The SPSS output provides the observed frequencies (count and percentages) as well as expected frequencies by cell, followed by a table including the test statistic ($X^2 = 67.8$), degrees of freedom (df = 4), and p-value (less than .0005).

The statistically significant association or dependence between these categorical variables is graphically supported by the clustered bargraph (Figure 1), also provided by the software output.

Figure 1: Clustered bargraph by Ethnic group



The graph clearly indicates that the distribution of Hemoglobin level by Ethnic group is quite different. We can conclude that the Hemoglobin level of children in the given population is Ethnic dependent (p-value < .0005). All these analyses and discussions using technology resources contribute to an increased students' motivation and understanding that would be virtually impossible with a traditional teaching approach.

3. Results and Discussion

The STA3112 and STA3194 classes were taught by the present author every spring term of the 2010-2013 years. A problem involving the chi-square test for Independence-Dependence of two categorical variables was included in the SPSS take-home assignments and the cumulative final exam. Overall passing rates for these two courses during the given years were 90% for STA3112 (n = 211) and 94% for STA3194 (n = 63), with a combined retention rate of 98%. Moreover, students' satisfaction was high as demonstrated by the combined 88% of excellent/very good opinions about the overall quality of instruction, as assessed by the official university surveys.

The use of Power Point where text was presented in conjunction with tables, graphs and other pictorial representations assisted students, particularly those with a more visually oriented learning style. The course pack comprised of PowerPoint slides helped students to focus on class discussions by minimizing the note-taking process. Furthermore, the integration of computational technology provided an effective tool for this topic by generating more time for analysis of results and conceptual understanding. The use of SPSS output for statistical graphs and tables led to a deeper problem comprehension. Instructor's mobility in the classroom, granted by the use of a presenter device, also facilitated communication with the students.

4. Conclusions

This discussion indicates that the use of technology resources, in conjunction with an interactive approach that emphasizes the conceptual understanding, provides a highly effective teaching-learning approach for the Chi-square test of independence/dependence of two categorical variables. The quality of instruction and students' understanding improved as 1) Students learned effectively as indicated by the combined 91% passing rate of the courses; 2) Students' motivation was high as suggested by the 98% retention rate. 3) Students' satisfaction was also high as evidenced by the 88% of excellent/very good opinions about the overall quality of instruction.

References

American Statistical Association. (2005). Guidelines for assessment and instruction in statistics education (GAISE): College Report. www.amstat.org/education/gaise.

Cryer, J.D. (2005). Teaching statistics to business students. *Innovations in teaching statistics*, MAA Notes #65.

Daniel, W. W. & Cross, C. (2013). Biostatistics: A foundation for analysis in the health sciences, 10th edition. Hoboken, NJ: John Wiley & Sons

Fernando, H. (2012). Teaching an undergraduate statistics class with technology. *ICTCM-24 Proceedings*.

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

Gomez, R. (2010). Using technology in introductory statistics courses. *ICTCM-22 Proceedings*.

Gomez, R. (2010). Using technology in large statistics classes. *Review of Higher Education and Self Learning, Vol. 3, Issue 5, 109-113*

Gomez, R. (2010). Innovations in teaching undergraduate statistics courses for biology students. *Review of Higher Education and Self Learning, Vol. 3, Issue 7, 8-13*

Gomez, R. (2011). Teaching a second statistics course for undergraduate psychology students. *Review of Higher Education and Self Learning, Vol. 4, Issue 11, 1-6*

Gomez, R. (2013). A modern approach to teach business statistics at college. *Review of Higher Education and Self Learning, Vol. 6, Issue 18, 118-126*

Higazi, S. M. (2002). Teaching statistics using technology. *ICOTS6 Proceedings*.

Hulsizer, M. C. & Woolf, L. M. (2009). A guide to teaching statistics. Malden, MA: Wiley-Blackwell.

Lock, R. (2005). Teaching a technology-enhanced course in a liberal arts environment. *Innovations in teaching statistics*, MAA Notes #65, 31-38.

McClave, J. & Sincich, T. (2013). Statistics, 12th edition. Upper Saddle River, NJ: Prentice Hall.

Robinson, K. & Kimmel, J. (2013). Statistics education.-An evolving collaboration: Data, technology, and learning. *USCOTS-5 Proceedings*.