THREE GRAPHS TO PROMOTE STATISTICAL LITERACY

Milo Schield, Director of the W. M. Keck Statistical Literacy Project
Dept of Business Administration. Augsburg College. Minneapolis, MN

Abstract: Graphical techniques have been used in introductory statistics to teach three big statistical topics: (1) confounding (which can result in Simpson's Paradox), (2) statistical significance and (3) the vulnerability of statistical significance to confounding. These graphical techniques have been used to teach students as part of the W. M. Keck Statistical Literacy project. These graphs have transformed statistical education at Augsburg College; they can change statistical education everywhere.

1. THREE BIG PROBLEMS

Three big topics in teaching statistics are (1) confounding (Simpson's Paradox is an extreme case), (2) statistical significance and (3) the vulnerability of statistical significance to confounding. *Confounding involves three factors* and thus requires some form of multivariate analysis.

Utts (2003) upholds the importance of confounding. *"One lecture explaining the difference between an observational study and a randomized experiment, and the role of confounding variables in the interpretation of observational studies would do more to prepare students for reading the news than a dozen lectures on statistical inference procedures." "It is important for students … to understand … how the potential for confounding variables limits the conclusions that can be made from observational studies." "When illustrating this concept…, it is important to give many examples and to discuss how confounding variables may account for the relationship."*

Nicholson et al. (2004) reported on world class tests to assess the problem solving skills of high attaining students in mathematics, science and technology. These tests, developed by the MARS group at Durham and Nottingham, required students to work with three or more variables, often non-linearly related to each other. As part of this project, raters reviewed statistics exam papers at the AS level (the first level of advanced study). Raters found that none of the AS level questions required students to work with more than two variables. As mentioned previously, confounding involves a minimum of three variables, so confounding was not likely to be involved.

Some social sciences teachers argue that students should be taught multivariate analysis (regression or ANOVA) in the entry course. In 1993 Anagnoson and DeLeon held an NSF-sponsored workshop on Exploratory Data analysis at San Francisco State arguing this point.

Teaching ANOVA or multivariate regression without teaching the statistical inference associated therewith may be problematic for statistical educators in math/stat departments who have good reasons to believe the first course cannot be made to include all three topics. Aside from being told that "association is not causation" and listening to some well-chosen anecdotes, students are not educated on confounding. In many statistics texts, the term 'confounding' is not listed in the glossary or index. Instead students get a strong exposure to statistical significance in isolation.

As an unintended result, many students leave introductory statistics with three bad ideas: (1) statistical associations are immune to confounding (with few exceptions), (2) statistical significance is always 'cast in stone' and (3) if an association is statistically significant, then one can be very confident of having strong evidence of causation. From a process control perspective, students holding these conclusions are statistical defects. But since this outcome is consistent and systematic, the cause is the process. If the goal is to produce students who understand statistical significance and confounding, then statistical education is defective. Statistical education is unwittingly creating statistical defects: students who don't understand confounding, statistical significance and their interaction. This outcome is totally unacceptable – and now there is an alternative.

2. OVERVIEW

This paper presents three graphs that are used in teaching students majoring in business and the humanities at Augsburg College as part of the W. M. Keck Statistical Literacy Project. These graphs show the influence of confounding, the meaning of statistical significance, and the influence of confounding on statistical significance. Two of these graphs use standardization to determine the influence of confounding on associations. Preliminary results are presented.

3. TEACHING WEIGHTED AVERAGES GRAPHICALLY

Steen (2001) noted that understanding weighted averages is an important part of Quantitative Literacy. A survey of 24 Augsburg students (Appendix A) found that 5% had problems computing a simple average given the counts, 20% had problems computing a weighted-average of non-percentages using counts or percentages and 80% had problems computing a weighted-average of percentages using a weight measured in percent. More importantly, very few were able to explain what a weighted average is. Now the combination of the small sample size and the possibility of question-induced bias make these results preliminary, but they do support the claim that students lack an adequate understanding of weighted averages. The following figure presents two ways of showing weighted averages: a traditional graph and a new outcome-mixture graph.

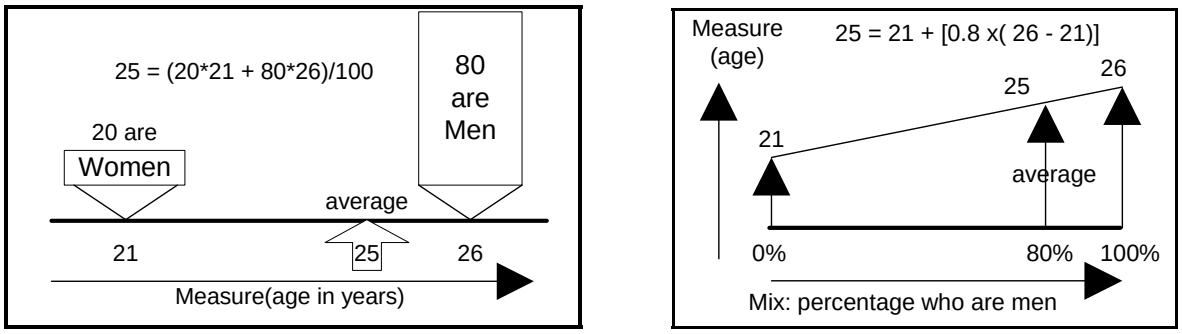

Illustration 1: Two Ways of Illustrating a Weighted Average

The traditional graph (left side) illustrates a weighted average using a see-saw where the measures run horizontally, the weight columns are placed at the appropriate horizontal locations and the balance point (calculated algebraically) is the weighted average. Sometimes the column heights indicate the size of the weights (in counts or percents). But since this approach can handle multiple values, there is no attempt to show the mixture as a single value.

The new outcome-mixture graph (right side) was developed by Jeon, Chung and Bae (1987), was independently developed by Baker and Kramer (2001) and was noted by Wainer (2002, 2004). The new approach works only for a confounder that is a binary variable and it assumes the outcomes obtained in the subgroups (0% and 100%) are independent of the mixture in the combined group. It transforms counts into percentages (which must add to 100% for a binary variable) so the mixture is measured using a single value. Mixtures run horizontally and measures run vertically. The weighted-average line connects the two outcomes. The value of the weighted average is the height of the weighted-average line at the value of a particular mixture.

The new approach has some difficulties. One difficulty is seeing that the right side consists entirely of people who are men (the left side is all women). Using "100% is the percentage of people who are men" (100% of people are men) to describe 'men' seems convoluted at best and irrational at worst. A second difficulty involves omitting the group and mistakenly seeing 80% as "individuals who are 80% men." Once past these problems, students' understanding of the weighted average as a mixture is straightforward. But at this point, they cannot see how this graph can identify the influence of confounding on an association between two groups.

4. STANDARDIZATION

What students can't foresee is the use of this outcome-mixture graph to standardize two groups. The nature and benefits of standardization are well known. According to Newell (1994), standardization is "*a general statistical technique and is used in many areas other than mortality analysis. The aim of standardization is to allow more precise comparison of two or more 'crude' rates by eliminating the effect of, say, the differences in age structure between two or more populations. Usually rates are age-standardized, but many other attributes can also be used*."

Standardization separates out the confounding influence of changes in an age distribution over time In the US, the crude death rate due to pneumonia was 7.4% higher in 1996 (33.4) than in 1990 (31.1) per 100,000 population. But the age-adjusted death rate due to pneumonia was 5.1% lower in 1996 (13.0) than in 1990 (13.7) per 100,000 population standardized to the 1940 US population distribution.[1] Standardizing reverses the direction of this association.

Standardization separates out the confounding influence of differences between age distributions of two groups at the same time. In 2001, the crude death rate was 43% lower in Mexico (5.0/1,000) than in the US (8.7/1,000). But the percentage of the population who were under 15 years old was 50% greater in Mexico (33.3%) than in the US (21.2%).[1] Given this difference in age distribution, standardization would decrease (if not reverse) this difference in death rates.

Standardization using external sources is always contra-factual: the standardized values do not exist – they are predicted.[2] Since standardization, per se, does not involve the use of a model, it bypasses much of the complexity associated with modeling. Standardizing does involve assumptions[3] (as does predicting that 95% of the values in a distribution are within 2 standard deviations of the mean). But the avoidance of modeling makes the use of standardizing particularly valuable in teaching statistical literacy where the primary focus is to help students obtain a conceptual appreciation of statistics in dealing with realistic situations. See Schield (2004).

5. GRAPH #1: STANDARDIZATION GRAPHS

A standardization graph is an outcome-mixture graph involving two groups (two predictors) having a common outcome and a common binary confounder where the mixture of the confounder in both groups has been standardized. Figure 1 shows two such rate-weight graphs.

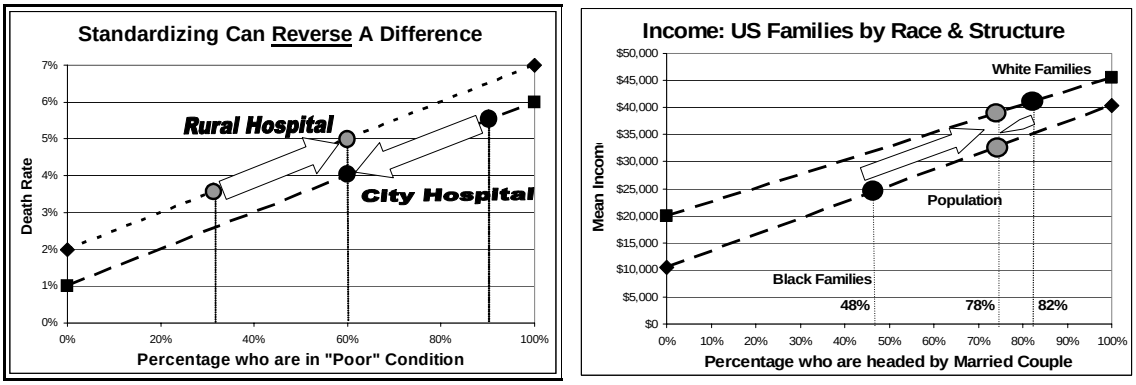

Figure 1 Confounder Influence on Associations

[1] 2001 U.S. Statistical Abstract. Tables 105 & 106; Tables 1330 and 1328.

[2] Standardization can be either direct or indirect. Direct standardization combines existing rates (rates by age group) with an external distribution of subjects (standard age distribution) to generate new standardized rates. Indirect standardization combines the existing distribution of subjects (distribution by age) to external standard rates (by age group) to obtain new expected rates against which the actual rates can be compared.

On the left side, the overall death rate is higher at the city hospital (5.5%) than the rural hospital (3.5%). But the percentage of patients who are in poor condition is much higher at the city hospital (90%) than at the rural hospital (30%). And given the slopes of these lines, being in poor condition is positively associated with a higher rate of death. Suppose we adjust the mix of patients by giving both groups a standard mix: the same percentage of patients in poor condition as in their combined patient load (60%). Standardizing the mix in both groups would have increased the expected death rate at the rural hospital and decreased it at the city hospital. In this particular case the association between hospital and patient death rate was reversed after controlling for patient condition. For more on Simpson's Paradox, see Schield (1999).

The graph on the right side shows that the mean income is 64% ($16,000) more for white families ($41,000) than for black families ($25,000). But the percentage of families who are headed by a married couple is much higher among whites (82%) than blacks (48%). Suppose we adjust the mix of family types in both groups to a standard mix: the same percentage of families who are married as found in their combined distribution (78%). This standardized family-income is 18% ($6,000) more for whites ($39,000) than for blacks ($33,000). Standardizing on family structure decreases the black-white income gap by 62% (from $16,000 to $6,000).

Students find both of these graphs all but self-explanatory. In both cases they see causality at work. In the hospital-death rate case, patients in poor condition are much more likely to die than those in good condition. In the race-income case, families headed by a married couple are more likely to earn higher incomes than those headed by a single parent. Since the value of the outcome (death rate or income) is intimately related to the confounder, students find this presentation to be most satisfying. This is not to say that they understand the role and importance of a weighted average as an idea. And they have difficulty distinguishing predictor from confounder and plotting outcomes at appropriate places in building these graphs. But they can articulate why an association can appear to be directed one way overall and yet change (increase, decrease, or even reverse) after standardizing for a confounder. Now for students who have studied differential equations, the fact that a total derivative may be larger or smaller (or even have the opposite sign) from a partial derivative is not surprising. But to students lacking this background, having a change in sign for an arithmetic comparison of ratios or means is very novel and challenging.

6. SELECTION OF A STANDARD VALUE FOR THE MIXTURE

At one level, the selection of a standard value for the confounder is arbitrary.[3] If one is interested in adjusted differences, then these will be the same independent of the choice if the rate-weight lines are parallel (the data is planar or the model is non-interactive), but they will vary with the choice if the rate-weight lines are not parallel (the data is non-planar and the model is interactive). If one is interested in adjusted ratios (relative risk) then the choice of a standard matters even if the rate-mixture lines are parallel.[4] The choice used herein is to set the percentage of the confounder in each group equal to that in their combined group. This has two benefits. The overall percentage of the confounder remains unchanged. Using the group average as the standard emulates the desired outcome in a randomized experiment where the goal is for each group (exposure and control) to have the same percentage of confounder as found in the overall population.

7. GRAPH #2: TEACHING STATISTICAL SIGNIFICANCE GRAPHICALLY

Giere (1996) proposed a graphical approach to the statistical significance of a difference in proportions. This approach introduces a statistically significant difference as involving the lack of

3 In some cases, the data being used may be very sensitive to the confounder level under which the data was obtained. In the hospital case, the actual death rate among patients in poor condition in the rural hospital might be much higher than 7% if the percentage of patients who are in poor condition were 100% rather than 30%.

4 One choice is to set the percentage of the confounder in one group equal to that in the other group, but results will often vary depending on which group is the standard (as will the percentage of the confounder overall).

overlap between confidence intervals (left side of Figure 2). This graph is not new nor is using this technique new. After presenting confidence intervals and statistical significance, statistical educators often note this lack of overlap can indicate a difference that is statistically significant. [5]

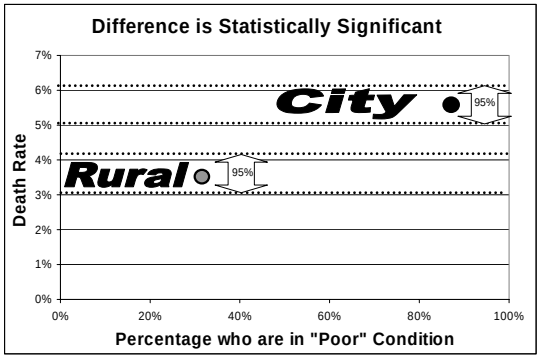

Figure 2 Statistical Significance: Non-overlapping Confidence Intervals

Giere's insight was to use this graph and related technique as the primary method of teaching statistical significance. Giere focuses entirely on binary outcomes (which bypassed mentioning standard deviation in the population), bypassed the derivation of the sampling distribution (which bypassed discussing standard error) and used the most conservative 50% confidence intervals (which are independent of the actual proportion Π in the population) so that non-overlapping confidence intervals had a fixed size and were sufficient (but not necessary) for statistical significance. This approach does substantially reduce the time involved in teaching statistical significance. But in so doing it omits or skims topics that many consider essential: sampling distributions of quantitative variables, null and alternate hypothesis, Type 1 and Type 2 error, one and two-tailed tests, p-values, two-t tests, etc. And Giere's test is so conservative it may fail to detect all too many cases of statistical significance. While this approach takes far less time to teach, statistical educators would have good reasons not to teach statistical significance in this manner.

8. GRAPH #3: CONFOUNDING AND STATISTICAL SIGNIFICANCE

The graph #1 approach to standardizing (Figure 1) has been combined with the graph #2 approach to teaching statistical significance (Figure 2) to create graph #3. Graph #3 (Figure 3) shows students how confounding can influence statistical significance.[6]

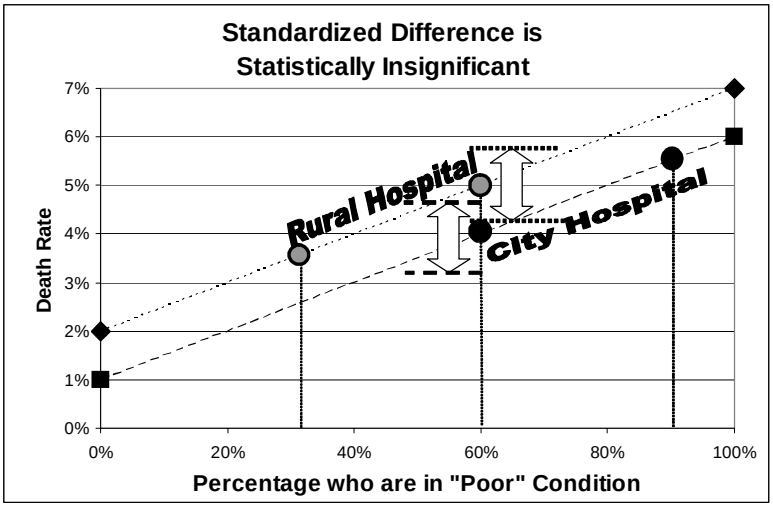

Figure 3 Vulnerability of Statistical Significance to Confounding

[5] If two (1-alpha) confidence intervals just touch, then the difference in population means is non-zero (statistically significant) at a level of significance at or below alpha in a two-tailed test.

[6] Envisioned by the author in January 2003 in the Czech Republic, this was first taught in the fall of 2003.

While the techniques are under development, students seem to understand that statistical significance can change on standardizing just as associations can change. At one level this graph (#3) is simply a combination of the previous two graphs (#1 and #2). But at another level, this graph is much greater than the sum of its parts. This simple graph is very potent; it takes on a major problem. One of the greatest misuses of statistics is to presume that statistical significance is immutable when the underlying associations are based on observational data. This simple graph may be a silver bullet for demolishing that mistaken conclusion. Presenting this graph may be the greatest contribution of statistical educators to helping the public understand the nature and limits of statistical significance in observational studies.

9. ANALYSIS

It may be possible to present all three graphs while still taking students through the traditional approach to statistical inference. If so, then students may have the best of both worlds. But if taking students through the long trek from the binomial theorem through the central limit theorem to statistical significance leaves them too brain-dead to appreciate the ideas presented in these three graphs then the goal will not have been achieved. At that point, statistical educators may be willing to consider teaching statistical significance using Giere's short-cut approach involving proportions. The best method of presenting these concepts is something that statistical educators will have to sort out.

Since the weighted average is central to Quantitative Literacy (MAA) while statistical significance is central to Quantitative Literacy (ASA), teaching the influence of confounders on statistical significance may serve to bring these two approaches closer together.

10. CONCLUSION

A primary goal of statistical education should be to introduce students to three topics: (1) the influence of confounding in observational studies, (2) the nature of statistical significance for an association, and (3) the vulnerability of statistical significance to confounding. The three graphs presented in this paper provide a way of teaching these three topics in a way that students can readily understand and with minimal development of auxiliary concepts. See Schield (2004).

If the goal is statistical literacy, then Giere's short-cut approach to statistical significance may be adequate. If the goal is statistical competence, then a traditional approach may be required. Either way, students must be taught that statistical significance is contextual (that it can vary depending on what is taken into account). The vulnerability of statistical significance to confounding seems at least as important as statistical significance alone. The introduction of these three graphical techniques gives statistical educators an opportunity to rethink the goals of the introductory course whether it is the traditional version or a statistical literacy survey course.

If using these graphs results in a better understanding among students for statistical significance and the vulnerability of statistical significance in the face of confounding, then these graphs will have produced a substantial improvement in statistical education.

REFERENCES

Baker, S. G. & Kramer, B. S. (2001). *Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies.* Journal of Women's health and gender-based medicine, 10, 867-872.

Giere, Ronald (1996). *Understanding Scientific Reasoning.* 4th ed., Holt, Rinehart and Winston.

Jeon, J. W., Chung, H. Y., and Bae, J. S. (1987). *Chances of Simpson's Paradox*. Journal of the Korean Statistical Society, 16, 117-125.

Newell, Colin (1994). *Methods and Models in Demography*. Wiley Press

Nicholson, James, Jim Ridgway and Sean McCusker (2004). *Uncovering and Developing Student Statistical Competencies via New Interfaces.* 2004 IASE Curriculum Design Roundtable.

Schield, Milo (1999). *Simpson's Paradox and Cornfield's Conditions.* 1999 ASA Proceedings of the Section on Statistical Education, p. 106-111.[7,8]

Schield, Milo and Thomas Burnham (2003). *Confounder-Induced Spuriousity and Reversal: Algebraic Conditions for Binary Data Using a Non-Interactive Model.* 2003 ASA Proceedings of the Section on Statistical Education, p 3690- 3697.[7,8]

Schield, Milo (2004). *Statistical Literacy Curriculum Design.* 2004 IASE Roundtable.[7,8]

Steen, Lynn (2001). *Mathematics and Democracy: The Case for Quantitative Literacy.* Prepared by the National Council on Education and the Disciplines (NCED). Published by the Woodrow Wilson National Fellowship Foundation.

Utts, Jessica (May 2003). *What Educated Citizens Should Know About Statistics and Probability?* The American Statistician, Vol. 57, No 2, p. 74-59.

Wainer, Howard (2002). "*The BK-Plot: Making Simpson's Paradox Clear to the Masses.*" Chance Magazine Vol. 15, No. 3, Summer 2002, pp. 60–62.

Wainer, Howard (2004). "*Three Paradoxes in the Interpretation of Group Differences"* Draft of a paper submitted to The American Statistician.[7]

ACKNOWLEDGMENTS

This paper was accepted for presentation by distribution at the International Conference of Mathematics Educators (ICME-10) in Copenhagen (2004) in Topic Study Group (TSG-11): Research and development in the teaching and learning of probability and statistics. This work was supported by a grant from the W. M. Keck Foundation to "support the development of statistical literacy as an interdisciplinary curriculum in the liberal arts." My colleague, Thomas Burnham, reiterated the distinction between adjusting the data and changing the parameters in a model). Project reviewer Dr. John Stein and project research assistant Lena Zakharova noted that having straight lines in the outcome-mixture graphs assumes the subgroup rates are independent of the confounder prevalence (see footnote 3).

Appendix A. Weighted Average Survey

These questions are extracted from a survey given to 24 students at Augsburg who are in majors (business and humanities) that do not require a previous course in college mathematics.[9] These are the answers and the percentage who answered correctly: #3, $16 (75%), #4, $1.90 (71%), #5, 80% (96%), #6, 50% (96%), #7, 60% (79%), #8, 30% (17%) and #9, 36% (21%).

---

[7] Copy posted at www.StatLit.org/Articles

[8] Copy posted at www.Augsburg.edu/ppages/~Schield

[9] #3. Jan buys 4 carving knives on E-bay at $10 each and 6 at $20 each. What is her average cost? $10, $11, $15, $16 or None of these

#4. John buys 80% of his gas at $2 per gallon and 20% at $1.50. What is his average cost? $1.60, $1.75, $1.80, $1.90 or None of these

A basketball player makes 8 shots in 10 tries in the first half; 10 shots in 20 tries in the second.

#5. During the first half, what percentage of tries are made? 8%, 18%, 80%, 125% or None of these.

#6. During the second half, what percentage of tries are made? 10%, 20%, 50%, 200%, None of these.

#7. During the entire game, what percentage of tries are made? 18%, 38%, 50%, 60%, None of these.

Ken has completed 40% of his task; Jan has completed 20% of her task.

#8. Ken's task is 50% of the project; Jan's task is the rest. How much of the project is complete? 20%, 30%, 36%, 60% or None of these.

#9. Ken's task is 80% of the project; Jan's task is the rest. How much of the project is complete? 20%, 30%, 36%, 60% or None of these.