Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report
DRAFT February 2016

Committee:

Robert Carver (Stonehill College), Michelle Everson (The Ohio State University), John Gabrosek (Grand Valley State University), Ginger Holmes Rowell (Middle Tennessee State University), Nicholas Horton (Amherst College), Robin Lock (St. Lawrence University), Megan Mocko (University of Florida), Allan Rossman (Cal Poly – San Luis Obispo), Paul Velleman (Cornell University), Jeffrey Witmer (Oberlin College), and Beverly Wood (Embry-Riddle Aeronautical University)

Contents

APPENDIX B: Multivariable Thinking

The 2014 ASA guidelines for undergraduate programs in statistics recommend that students obtain a clear understanding of principles of statistical design and tools to assess and account for the possible impact of other measured and unmeasured confounding variables (ASA, 2014). (possible footnote: See also Wild's "On locating statistics in the world of finding out", http://arxiv.org/abs/1507.05982.) An introductory statistics course cannot cover these topics in depth, but it is important to expose students to them even in their first course (Meng, 2011). Perhaps the best place to start is to consider how a third variable can change our understanding of the relationship between two variables.

In this appendix we describe simple examples where a third factor clouds the association between two other variables. Simple approaches (such as stratification) can help to discern the true associations. Stratification requires no advanced methods, nor even any inference, though some instructors may incorporate other related concepts and approaches such as multiple regression. These examples can help to introduce students to techniques for assessing relationships between more than two variables.

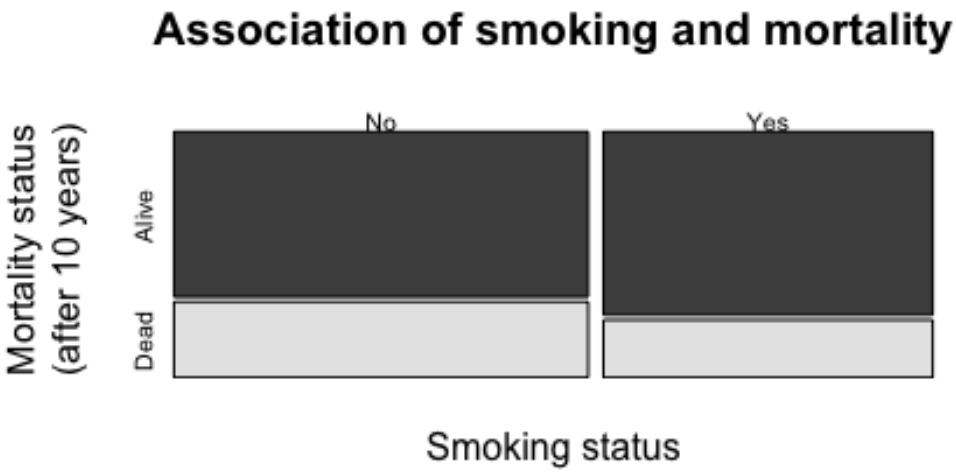
Including one or more multivariable examples early in an introductory statistics course may help to prepare students to deal with more than one or two variables at a time and examples of observational (or "found" data) that arise more commonly than results from randomized comparisons.

Smoking in Whickham

A follow-up study of 1,314 people in Whickham, England characterized smoking status at baseline, then mortality after 10 years (Appleton et al, 1996). The summary data are provided in the following table:

SMOKER	Alive	Dead
No	502 (68.6%)	230 (31.4%)
Yes	443 (76.1%)	139 (23.9%)

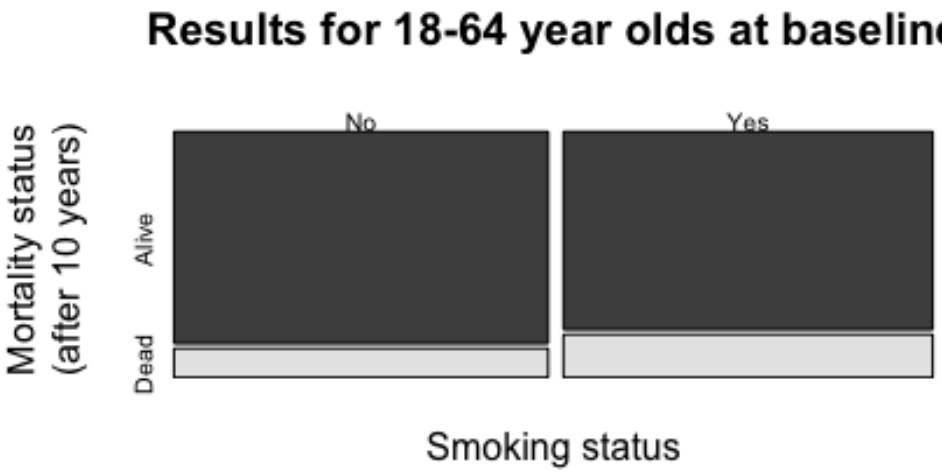
We see that the risk of dying is lower for smokers than for non-smokers, since 31.4% of the non-smokers died, but only 23.9% of the smokers did not survive over the ten year period. A graphical representation using a mosaicplot (also known as an *Eikosogram*) represents the cell probabilities as a function of area.

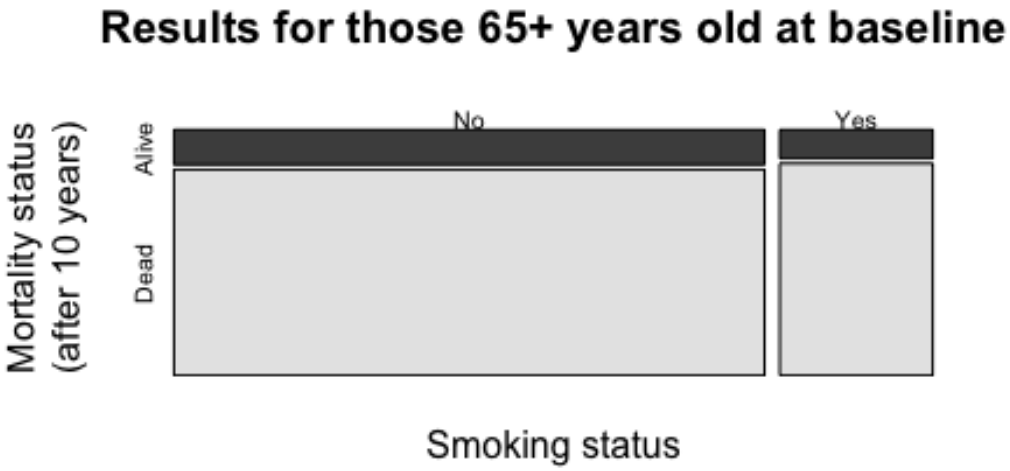


We note that the majority of subjects have survived, but that the number of the smokers who are still alive is greater than we would expect if there were no association between these variables. What could explain this result?

Let's consider stratification by age of the participants (older vs. younger). The following table and figure display the relationship between smoking and mortality over a 10-year period for two groups: those age 18-64 and subjects that were 65 or older at baseline.

Baseline age	SMOKER	Alive	Dead
18-64	No	474 (87.9%)	65 (12.1%)
18-64	Yes	437 (82.1%)	95 (17.9%)
65+	No	28 (14.5%)	165 (85.5%)
65+	Yes	6 (12.0%)	44 (88.0%)





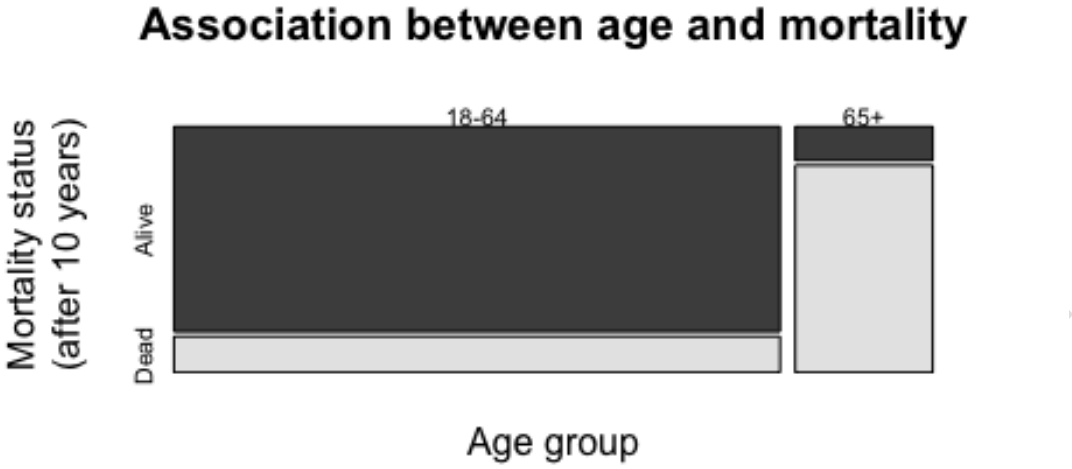
We see that mortality rates are low for the younger group, but the mortality rate is slightly higher for smokers than non-smokers (17.9% for smokers vs 12.1% for the non-smokers).

(possible foonote: Smoking is "bad" within both of the subgroups of age, while smoking is "good" overall.)

Almost all of the participants who were 65 or older at baseline died during the followup period, but the probability of dying was also slightly higher for smokers than non-smokers.

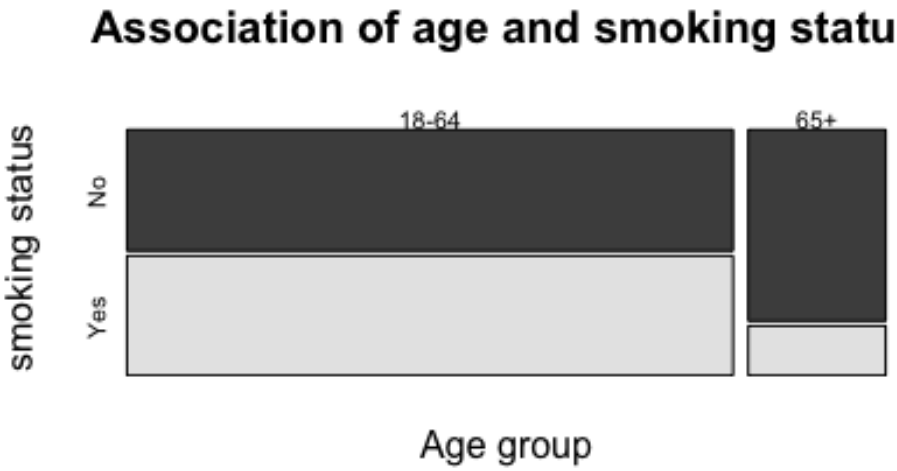
This example represents a classic example of *Simpson's paradox* (Simpson, 1951; Norton and Divine, 2015). For all of the subjects, smoking appears to be "protective", but within each age group smokers have a higher probability of dying than non-smokers.

How can this be happening? The following figure and table us to disentangle these relationships.



Not surprisingly, we see that mortality rates are highest for the oldest subjects.

We also observe that there is an association between age group and smoking status, as displayed in the following figure and table.



```
Age group  Non-smoker    Smoker
18-64      539 (50.3%)   532 (49.7%)
65+        193 (79.4%)    50 (20.6%)
```

Smoking is associated with age, with younger subjects more likely to have been smokers at baseline.

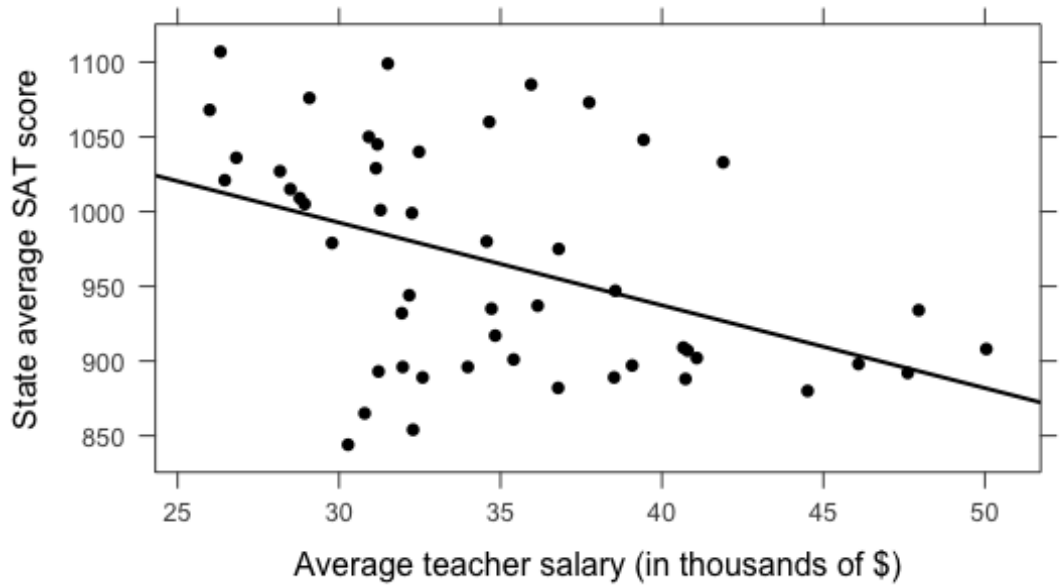
What should we conclude? After controlling for age, smokers have a higher rate of mortality than non-smokers in this study. This other factor is important when considering the association between smoking and mortality.

Simple methods such as stratification can allow students to think beyond two dimensions and reveal effects of confounding variables. Introducing this thought process early on helps students easily transition to analyses involving multiple explanatory variables.

SAT scores and teacher salaries

Consider an example where statewide data from the mid-1990's are used to assess the association between average teacher salary in the state and average SAT (Scholastic Aptitude Test) scores for students (Guber, 1999; Horton, 2015). These high stakes high school exams are sometimes used as a proxy for educational quality.

The following figure displays the (unconditional) association between these variables. There is a statistically significant negative relationship (β_1 hat =-5.54 points, $p = 0.001$). The model predicts that a state with an average salary that is one thousand dollars higher than another would have SAT scores that are on average 5.54 points lower.



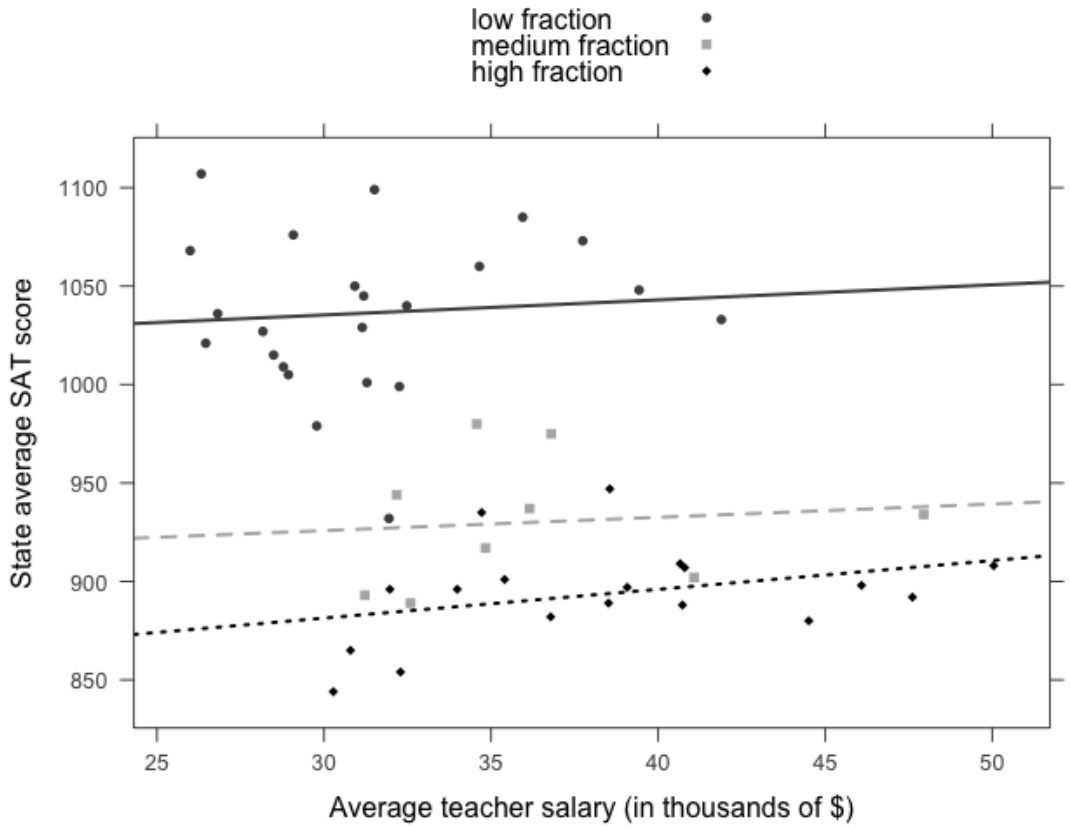
But the real story is hidden behind one of the "other factors" that we warn students about but do not generally teach how to address! The proportion of students taking the SAT varies

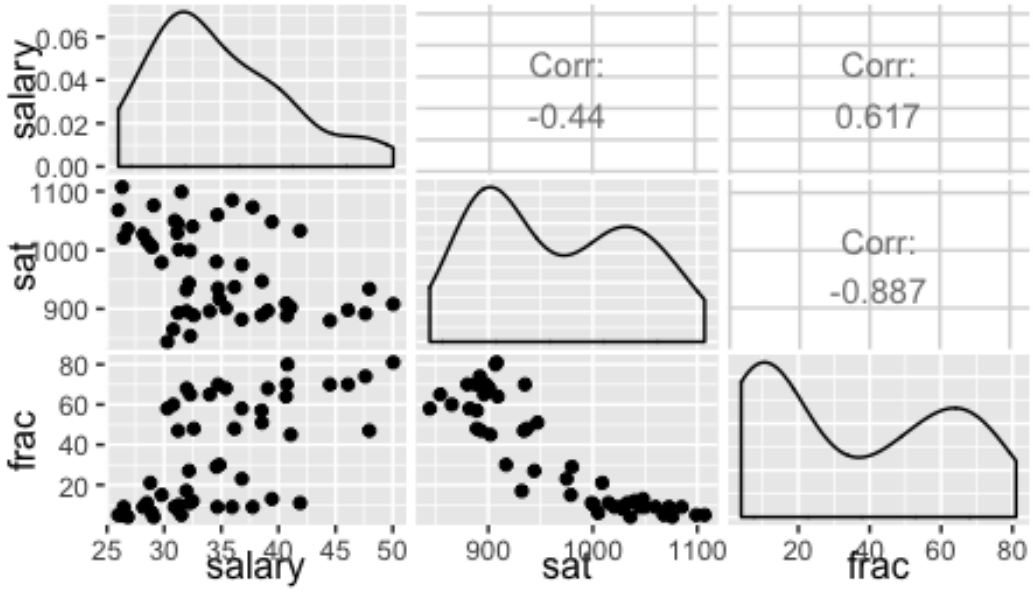
dramatically between states, as do teacher salaries. In the midwest and plains states, where teacher salaries tend to be lower, relatively few high school students take the SAT exam. Those that do are typically the top students who are planning to attend college out of state, while many others take the alternative standardized ACT test that is required for their state. For each of the three groups of states defined by the fraction taking the SAT, the association is non-negative. The net result is that the fraction taking the SAT is a confounding factor.

This problem is a continuous example of Simpson's paradox. Statistical thinking with an appreciation of Simpson's paradox would alert a student to *look for* the hidden confounding variables. To tackle this problem, students need to know that multivariable modeling exists, but not all aspects of how it can be utilized.

Within an introductory statistics course, the use of stratification by a potential confounder is easy to implement. By splitting states up into groups based on the fraction of students taking the SAT it is possible to account for this confounder and use bivariate methods to assess the relationship for each of the groups.

The scatterplot in the next figure displays a grouping of states with 0-22% of students ("low fraction", top line), 23-49% of students ("medium fraction", middle line), and 50-81% ("high fraction", bottom line). The story is clear: there is a positive or flat relationship between teacher salary and SAT score for each of these groups, but when we average over them, we observe a negative relationship.





Further light is shed via a matrix of scatterplots (see the above figure): we see that the fraction of students taking the SAT is negatively associated with the average statewide SAT scores and positively associated with statewide teacher salary.

Recall that in a multiple regression model that controls for the fraction of students taking the SAT variable, the sign of the slope parameter for teacher salary flips from negative to positive.

It's important to have students look for possible confounding factors when the relationship isn't what they expect, but it is also important when the relationship is what is expected. It's not always possible to stratify by factors (particularly if important confounders are not collected).

Multiple regression

The most common multivariable model is a multiple regression. Regression can be introduced as soon as students have seen scatterplots and thought about the patterns we look for in them. When students have access to a statistics program on a computer, they can find regression analyses for themselves. But even without computer access, they can learn about typical regression output tables. The point is to show students a model involving three (or more) variables and discuss some of the subtleties of such models. Here is one example.

Scottish hill races are scheduled throughout the year and throughout the country of Scotland (http://www.scottishhillracing.co.uk). The official site gives the current records (in seconds) for men and women in these races along with facts about each race including the distance covered (in km) and the total amount of hill climbing (in meters). Naturally, both the distance and the climb affect the record times. So a simple regression to predict time from either one would miss an important aspect of the races.

For example, the simple regression of time versus climb for women's records looks like this:

```
Response variable is:   Women's Record
R squared = 85.2%     R squared (adjusted) = 84.9%
s = 1126  with  70-2 = 68  degrees of freedom

Variable        Coefficient     SE(Coeff)     t-ratio    P-value
Intercept       320.528         222.2          1.44        0.1537
Climb             1.755           0.088       19.8       < 0.0001
```

We see that the time is greater, on average, by 1.76 seconds per meter of climb. The R^2 value of 85.2% assures us that the fit of the model is good with 85.2% of the variance in women's records accounted for by a regression on the climb.

But surely that isn't all there is to these races. Longer races should take more time to run. And although an R^2 of 0.852 is good, it does leave almost 15% of the variance unaccounted for.

It is straightforward for students to learn that multiple regression models work the same way as simple regression models, but include two or more predictors. Statistics programs fit multiple regressions in the same way as simple ones. Here is the regression with both `Climb` and `Distance` as predictors:

```
Response variable is:   Women's Record
R squared = 97.5%     R squared (adjusted) = 97.4%
s = 468.0  with  70 - 3 = 67  degrees of freedom
Variable      Coefficient  SE(Coeff)     t-ratio   P-value
Intercept     -497.656     102.8         -4.84     < 0.0001
Distance       387.628      21.45        18.1      < 0.0001
Climb            0.852       0.0621      13.7      < 0.0001
```

This regression model shows both the distance and the climb as predictors and has an R^2 of 0.975: a substantial improvement. More interesting, the coefficient of `Climb` has changed from 1.76 to 0.85. That's because in a multiple regression, we interpret each coefficient as the effect of its variable on y after allowing for the effects of the other predictors.

Closing thoughts

Multivariable thinking is critical to make sense of the observational data around us.
This type of thinking might be introduced in stages:

1. learn to identify observational studies,
2. explain why randomized assignment to treatment improves the situation,
3. learn to be wary of cause-and-effect conclusions from observational studies,
4. learn to consider potential confounding factors and explain why they might be confounding factors,
5. use simple approaches (such as stratification) to address confounding

Multivariable models are necessary when we want to model many aspects of the world more realistically. The real world is complex and can't be described well by one or two variables. If students do not have exposure to simple tools for disentangling complex relationships, they may dismiss statistics as an old-school discipline only suitable for small sample inference of randomized studies.

Simple examples are valuable for introducing concepts, but when we don't show students realistic models they are left with the impression that statistics is trivial and not really useful. This report recommends that students be introduced to multivariable thinking, preferably early in the introductory course and not as an afterthought at the end of the course.

References

Appleton, D. R., and French, J. M. and Vanderpump, M.P. (1996), "Ignoring a covariate: an example of Simpson's paradox", *The American Statistician*, 50(4):340-341.

American Statistical Association (2014), American Statistical Association Undergraduate Guidelines Workgroup. 2014. "2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science". Alexandria, VA: Author. Available at http://www.amstat.org/education/curriculumguidelines.cfm.

Guber, D. L. (1999), "Getting what you pay for: the debate over equity in public school expenditures", *Journal of Statistics Education*, 7(2).

Horton, N.J. (2015), "Challenges and opportunities for statistics and statistical education: Looking back, looking forward, *The American Statistician*, 69(2):138–145.

Meng, X.L. (2011), "Statistics: Your chance for happiness (or misery)", *The Harvard Undergraduate Research Journal*, 2(1), http://thurj.org/as/2011/01/1259.

Norton, H. J. and Divine, G. (2015), "Simpson's paradox, and how to avoid it", *Significance*, 40-43.

Simpson, E. H. (1951), "The interpretation of interaction in contingency tables", *Journal of the Royal Statistical Society, Series B*, 13:238-241.