SIMPSON'S PARADOX AND CORNFIELD'S CONDITIONS

Milo Schield, Augsburg College
Dept. of Business & MIS. 2211 Riverside Drive. Minneapolis, MN 55454

Abstract

Simpson's Paradox occurs when an observed association is spurious – reversed after taking into account a confounding factor. At best, Simpson's Paradox is used to argue that association is not causation. At worst, Simpson's Paradox is used to argue that induction is impossible in observational studies (that all arguments from association to causation are equally suspect) since any association could possibly be reversed by some yet unknown confounding factor. This paper reviews Cornfield's conditions – the necessary conditions for Simpson's Paradox – and argues that a simple-difference form of these conditions can be used to establish a minimum effect size for any potential confounder. Cornfield's minimum effect size is asserted to be a key element in statistical literacy. In order to teach this important concept, a graphical technique was developed to illustrate percentage-point difference comparisons. Some preliminary results of teaching these ideas in an introductory statistics course are presented.

Keywords: Statistical Literacy, Teaching; Epistemology; Philosophy of science; Observational studies.

1. STATISTICAL LITERACY

Statistical literacy studies the use of statistics and statistical associations as evidence in arguments (Schield, 1998). Many arguments involving statistical associations are based on observational studies and are directed at supporting claims on causation. Interpreting such associations is a major problem due to the possibility of confounding. Simpson's Paradox is a striking example of this problem.

2. WHAT IS SIMPSON'S PARADOX?

Simpson's Paradox is the reversal of an association between two variables after a third variable (a confounding factor) is taken into account. For an overview of association reversals, see Samuels (1993). A confounding factor is a factor — a lurking variable — which is found or mixed with another.

Simpson's paradox has been observed in several real-life situations. One well-known example occurred in the Graduate Division of the University of California at Berkeley. Women were rejected more often than men at the overall college level, but men were rejected more often than women at the individual departmental level. The confounding factor was the choice of department. Women were more likely to choose departments with higher rejection rates than were men (Freedman, Pisani, Purves, and Adhikari 1991, p. 16).

In another example, it was found that whites were more likely to be sentenced to death for murder than blacks. But after taking into account the confounding factor of the race of the victim, it was found that blacks were more likely to be sentenced to death than whites. A death sentence was more likely if the victim was white. Since blacks were more likely to kill blacks, they were less likely to be sentenced to death. But whether the victim was white or black, a death sentence was more likely for blacks than for whites. (Agresti 1984)

3. IS SIMPSON'S PARADOX IMPORTANT?

Simpson's Paradox is vitally important for several reasons. (1) It clearly demonstrates that correlation is not always causation. If the direction of an association can be reversed, any assertion about direct causation is clearly disputable. (2) It demonstrates that associations are sometimes conditional. Students often think of numerical associations as immutable—as unconditional. By studying Simpson's Paradox students overcome this mistaken perception. (3) It introduces the minimum effect size necessary for a confounder to explain a spurious association. The measurement of the minimum effect size is the point of this paper and is developed in a later section.

4. UNDERSTANDING SIMPSON'S PARADOX

It is not easy to understand the reasons for—much less the cause of—a reversal of an association, i.e., Simpson's Paradox. Consider three types of explanations:

Mathematical explanation: "Consider 8 variables: A, B, C, D, a, b, c and d. If it is true that A/B > a/b and C/D > c/d, is it also true that (A+C)/(B+D) > (a+c)/(b+d)?" The reply: "Not in general. For example, 1/1 > 3/4 and 1/4 > 0/1. Now (1+1)/(4+1) = 2/5 and (3+0)/(4+1) = 3/5. Now is 2/5 > 3/5? No!" (sci.stat.edu, 12/96). In this explanation, the reversal is just a consequence of the particular numbers involved.

Group inhomogeneity explanation. Suppose A, B, ... d are as above. Let A/B = P1 and C/D = P3. Let a/b = P2 and c/d = P4. Let the size of the groups being compared be illustrated by the number of "x" symbols.

xx A/B=P1 > P2=a/b xxxxxxxxxx
xxxxxx C/D=P3 > P4=c/d xx

The ⊕ symbol indicates the merging of the groups:

P1 ⊕ P3 = (A+C)/(B+D)
P2 ⊕ P4 = (a+c)/(b+d)

xxxxxxxx P1 ⊕ P3 < P2 ⊕ P4 xxxxxxxxxxxx

We see above that $P1 > \underline{P2}$ and $P3 > P4$, but $P1 \oplus \underline{P3} < \underline{P2} \oplus P4$: the Simpson's Paradox reversal.

The differing size of the groups explains the reversal in the association. This explanation is sometimes presented using baseball batting averages. Suppose in the first half of the season, Slugger (370 = P1) had a higher batting average than Bantam (330 = P2). In the second half Slugger (200 = P3) had a higher batting average than Bantam (190 = P4). Yet for the entire season, this association can reverse. Bantam can have a higher batting average than Slugger if Bantam bats much more than Slugger in the first half and much less than Slugger in the second half.

Confounding factor explanation: What we fail to take into account strongly influences our conclusions (Kelly 1994). Consider an example:

> A father and his young children were riding on the New York subway. The children were out of control. The father was slumped over with his head in his hands. When the father did nothing to control the children some of the passengers became irritated. One of them asked the father to control the children (implying the father was derelict in his responsibilities). The father lifted his head and explained that he and the children had left the hospital where his wife, their mother, had just died. The passengers immediately reversed their evaluation once they took account of the influence of the death on the family.

Students understand this principle: a more important factor can easily change one's standard for evaluation.

5. ANTICIPATING SIMPSON'S PARADOX

But even if Simpson's Paradox were readily understood, it is not easily anticipated. There is no test for determining whether an association is spurious (Pearl, 1999). Textbooks seldom indicate a way to estimate the likelihood of a Simpson's Paradox reversal.

After studying Simpson's Paradox, one student concluded one should never trust any association based on an observational study. And if there is no way to anticipate when a Simpson's Paradox reversal could occur, this student is absolutely right. One solution is to ignore observational studies and deal only with randomized experiments where the problem of confounding is minimized. However, experiments are not always possible, so students need to learn how to deal with associations based on observational studies.

6. FROM CORRELATION TO CAUSATION

A serious concern about the possibility of Simpson's Paradox arose in the late 1950s when several research projects found an association between smoking and lung cancer. But these associations were observational so it was possible that an unknown confounding factor might significantly change the associations.

Fisher (1958) noted that genetic factors might dispose one on whether to smoke or on what (cigarette, pipe, or cigar) to smoke. Although Fisher was a smoker, his article demonstrated his allegiance to the power of data. He did not just allude to the possibility of some confounding factor; he presented actual data on smoking choices among fraternal and identical twins. He calculated the percentage of twins in which there were distinct differences in smoking (smoker versus non-smoker or cigarette smoker versus pipe smoker). His data showed that there were distinct differences in smoking choice among 51% of the fraternal twins as opposed to 24% of the identical twins. He concluded, "There can be little doubt that the genotype exercises considerable influence on smoking, and on the particular habit of smoking adopted…"

Fisher used this association to suggest that perhaps lung cancer was not caused by smoking per se but was caused by that part of the genotype that also caused people to smoke. Thus people who are disposed to smoke would contract lung cancer at the same rate whether they smoke or not.

Cornfield et al (1959) countered Fisher's alternate explanation. They derived a necessary condition for a confounding factor to explain away an observed association—assuming the association was totally spurious.

7. CORNFIELD'S CONDITION

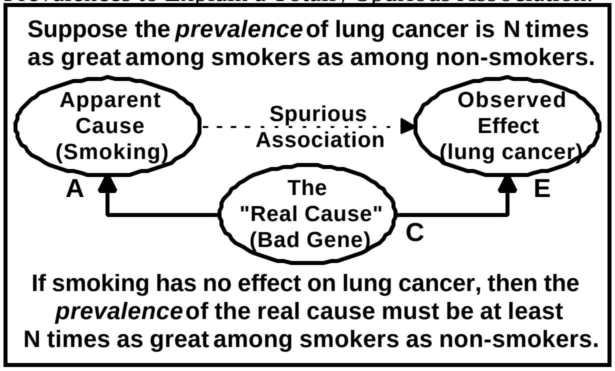
Cornfield et al deduced the minimum effect size necessary for a potential confounder to explain an observed association assuming the association is totally spurious. They wrote (Cornfield et al, 1959, Appendix A),

> If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r.
>
> Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X-producers among cigarette smokers must be at least 9 times greater than that of non-smokers. If the relative prevalence of hormone-X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect."

Cornfield's condition can be stated algebraically. P denotes a probability, A denotes the apparent cause, C denotes the common cause and E denotes an observable effect. A single quote following a letter is the complement of the condition(A' = 1-A). The vertical bar (|) denotes “given”. Thus P(C|A) is the probability of C given A; P(C|A') is the probability of C given the absence of A.

If factor A (smoking) had no effect on the likelihood of an observable effect E (lung cancer), Cornfield et al, proved that the prevalence of the actual cause (C) must satisfy: P(C|A)/P(C|A') > P(E|A)/P(E|A').

Figure 1. Necessary Relationship among Relative Prevalences to Explain a Totally Spurious Association.


Suppose the *prevalence* of lung cancer is N times as great among smokers as among non-smokers.
Apparent Cause (Smoking)
Spurious Association
Observed Effect (lung cancer)
A
E
The "Real Cause" (Bad Gene)
C
If smoking has no effect on lung cancer, then the *prevalence* of the real cause must be at least N times as great among smokers as non-smokers.

This necessary prevalence—Cornfield’s condition—blunted Fisher's argument. Fisher had noted a 2 to 1 relative prevalence (51% vs. 24%) in smoking behavior for the two types of twins. But Cornfield's condition required that Fisher show the prevalence of his genetic factor was nine times as great among smokers as among non-smokers. Fisher never replied.

[Actually, Fisher's comparison was of the form P(A|C)/P(A|C') – the relative prevalence of smokers among bad genes versus good genes -- instead of P(C|A)/P(C|A') – the relative prevalence of bad genes among smokers versus non-smokers.

The necessary condition of Cornfield et al is the positive side of Simpson's Paradox. It allowed statisticians to conclude that, to the best of their knowledge, smoking caused cancer – based on observational studies.

Cornfield's minimum effect size is as important to observational studies as is the use of randomized assignment to experimental studies. No longer could one refute an ostensive causal association by simply asserting that some new factor (such as a genetic factor) might be the true cause. Now one had to argue that the relative prevalence of this potentially confounding factor was greater than the relative risk for the ostensive cause. The higher the relative risk in the observed association, the stronger the argument in favor of direct causation, and the more the burden of proof was shifted onto those arguing against causation. While there might be many confounding factors, only those exceeding certain necessary conditions could be relevant.

Rosenbaum (1995) said of Cornfield's condition:

> Their statement is an important conceptual advance. The advance consists in replacing a general qualitative statement that applies in all observational studies by a quantitative statement that is specific to what is observed in a particular study. Instead of saying that an association between treatment and outcome does not imply causation, that hidden biases can explain observed associations, they say that to explain the association seen in a particular study, one would need a hidden bias of a particular magnitude. If the association is strong, the hidden bias needed to explain it is large.

8. METHOD OF DIFFERENCES

The minimum effect size can also be a simple difference of two percentages. Consider three approaches:

7.1 Conditional Probabilities

The influence of a confounding factor can be expressed using conditional probabilities and Bayes rule:

$$P(E|A) = P(E|C)\ P(C|A) + P(E|C')\ P(C'|A) \qquad 1a$$

$$P(E|A') = P(E|C)\ P(C|A') + P(E|C')\ P(C'|A') \qquad 1b$$

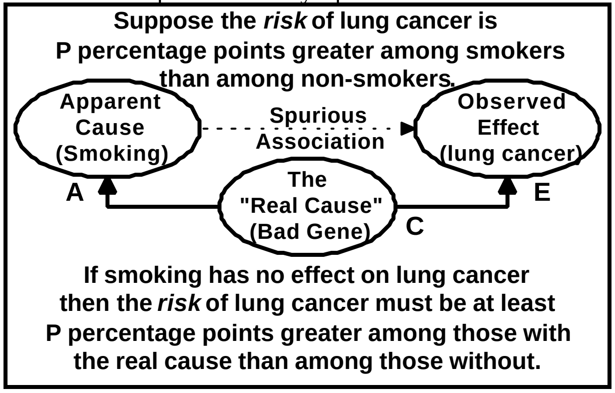
$$\boxed{P(E|A)-P(E|A') = [P(E|C)-P(E|C')][P(C|A)-P(C|A')] \quad 1c}$$

Since $[P(C|A) - P(C|A')] \le 1$,

$$\boxed{[P(E|C) - P(E|C')] \ge [P(E|A) - P(E|A')]} \qquad 1d$$

Cornfield et al derived the risk-difference condition in (1c) but dismissed it saying it “leads to no useful conclusion.” *This paper argues that this risk-difference condition is extremely useful (see Section 8).*

Figure 2. Necessary Relationship among Absolute Differences to Explain a Totally Spurious Association.


Suppose the *risk* of lung cancer is P percentage points greater among smokers than among non-smokers.
Apparent Cause (Smoking)
Spurious Association
Observed Effect (lung cancer)
A
E
The "Real Cause" (Bad Gene)
C
If smoking has no effect on lung cancer then the *risk* of lung cancer must be at least P percentage points greater among those with the real cause than among those without.

7.2 Regression Coefficients

The influence of a confounding factor can be expressed as a bias in the expected value of a regression coefficient (Wonnacott and Wonnacott 1990, p. 420). In the case of three variables: A, C and E, the expected change in the response variable E given a change in A can be biased whenever one ignores the influence of a confounding factor C. This bias is the product of two slope coefficients.

To illustrate, let the uncontrolled coefficient regressing E on A be b_0, the "whole effect". When regressing E on A and controlling for C, there are two coefficients, both involving E. Let b_1 be the coefficient involving A (the "direct effect"); let b_2 be the coefficient involving C. Let b_3 be the coefficient regressing C on A. The "indirect effect" is the product of b_2 and b_3. Wonnacott and Wonnacott show that the whole effect (b_0) is the sum of the direct effect (b_1) and the indirect effect (b_2 x b_3):

$$b_0 = b_1 + (b_2 \times b_3). \qquad 2a$$

If we fail to include C, the change in the expected value of E for a one unit change in A will be b_0, the whole effect. If C is a confounding factor, the change in expected value of E for a one-unit change in A should be b_1, the direct effect. This estimated change in E based on the whole effect will be biased by the amount of $b_2 \times b_3$, the indirect effect.

In relating this regression coefficient approach to Comfield's nullification, we can obtain the same result obtained earlier in (1d). With no direct effect ($b_1 = 0$), the direct association is completely spurious and

$$b_0 = b_2 \times b_3. \qquad 2b$$

The difference between the uncontrolled effect (b_0) and the direct effect (b_1) can be viewed as bias—an apparent influence due to a failure to take account of the confounding factor.

If all the variables are binary, then the regression slope coefficients are the difference in the associated percentages: $b_0 = P(E|A) - P(E|A')$,
$b_2 = P(C|A) - P(C|A')$ and
$b_3 = P(E|C) - P(E|C')$.
If $b_0 = b_2 \times b_3$, we obtain (1c).

Since these slopes are differences in probabilities, they have absolute values no greater than 1. Thus we can deduce that $b_2 \geq b_0$, as shown in (1d).

7.3 Partial Correlation Coefficients

The influence of a confounding factor can be expressed using partial correlation.

$$r_{AE,C} = \{r_{AE} - [r_{AC}\, r_{CE}]\}/ \sqrt{[(1-r^2_{AC})\,(1-r^2_{CE})]} \qquad 3a$$

If the apparent association between A and E (r_{AE}) is entirely spurious and is due entirely to associations with a common cause (C), then the association between A and E, conditioned on C, is zero ($r_{AE,C} = 0$). Thus,

$$r_{AE} = r_{AC}\, r_{CE} \qquad 3b$$

It follows that $|r_{AC}|$ and $|r_{CE}|$ must each be at least as large as $|r_{AE}|$. This relationship is well known, "For a confounding variable to explain an association of a given strength, it must have a much stronger association with both the possible causal factor and the disease" (Friedman 1994, p. 210 and 214).

When the variables involved are binary, the Pearson correlation coefficient reduces to phi (f):

$$f\,(E,C)=[P(E|C)-P(E|C')]\sqrt{[P(C)P(C')]/[P(E)P(E')]} \qquad 3c$$

Under (3b), f (E,A) = f (E,C) x f (C|A). Thus,

$$[P(E|A) - P(E|A')] \sqrt{\{[P(A)P(A')]/[P(E)P(E')]\}}$$
$$= \{[P(E|C) - P(E|C')]\sqrt{\{[P(C)P(C')]/[P(E)P(E')]\}}\}$$
$$\{[P(C|A) - P(C|A')]\sqrt{\{[P(A)P(A')]/[P(C)P(C')]\}}\} \qquad 3d$$

which reduces to (1c).

7.4 Comparison of Approaches

All three "difference" approaches give the same result as summarized by (1c) and (1d). The conditional probability approach is simplest. The regression approach is most powerful since it can be generalized to multiple confounding factors (Wonnacott and Wonnacott 1979, p. 415). Although the partial correlation coefficient approach is more theoretical, it can be shown to measure the strength of association without knowing the prevalence of C in A.

9. EXPLANATORY POWER

Equation (1d) gives a very simple method for determining whether a third variable (C) has the strength – the effect size – necessary to nullify or reverse an observed association between two other variables (A and E). Students need only compare two simple differences measured in percentage points. If,

$$\boxed{[P(E|C) - P(E|C')] \geq [P(E|A) - P(E|A')]} \qquad 1d$$

then one should be concerned about the possibility of a Simpson's Paradox reversal. This simple requirement establishes a minimum effect size for any confounding factor to bring about a nullification or a reversal of an observed association that is completely spurious.

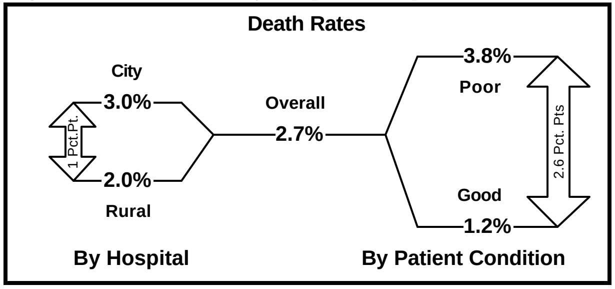
10. TEACHING SIMPSON'S PARADOX

For the past three years students in introductory statistics were taught to use simple differences -- differences in percentage points -- in comparing the explanatory powers of two binary variables. Students were cautioned that the truth of the percentage-point difference is not sufficient to imply a Simpson's Paradox reversal—it is only a necessary condition. Students have used these ideas as follows.

1. Consider two hospitals: a city hospital and a rural hospital. The death rate is 3% of cases at the city hos-

pital versus 2% at the rural. The combined death rate is 2.7%. Thus, it seems that the rural hospital is safer than the city hospital. See Figure 3.

Figure 3. Death rates by hospital and patient condition

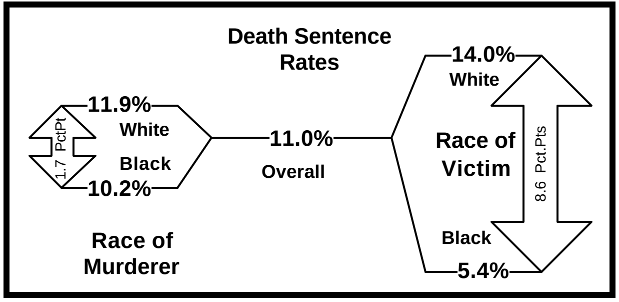


Now consider a plausible confounding factor: the condition of the patient's health. We find that overall the death rate among patients in poor condition is 3.8% while that among patients in good condition is 1.2%.

Here the simple difference in death rates by patient condition (2.6 percentage points) is greater than the simple difference in death rates by hospital (1 percentage point). Thus we have strong reason to be concerned about a possible Simpson's Paradox reversal of the association between hospital and death rate. To guard against such a reversal we can take into account (control for) patient condition when comparing the death rates for these two hospitals.

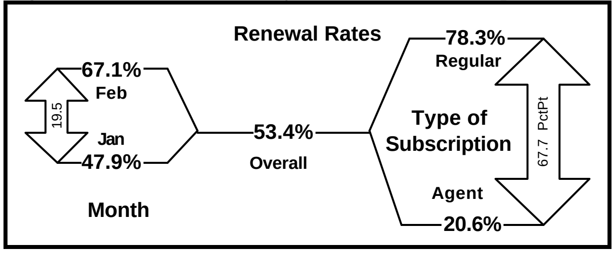
2. In a group of convicted murderers, the death penalty was given for 11.9% of white murderers and 10.5% of black murderers (Agresti 1984). Based on this data, one might argue that the legal system is biased against whites. However, when the sentences are classified by the race of the victim, the death penalty was given in 14.0% of the cases with a white victim and 5.4% of the cases with a black victim. The difference in the rate of death sentences by race of victim (8.6 percentage points) is greater than the difference in rate of death sentences by race of murderer (1.4 percentage points). To guard against a Simpson's Paradox reversal we must take into account the race of the victim when studying the association between the death penalty and the race of the murderer. See Figure 4.

Figure 4. Death sentence rates by race of murder and race of victim



3. Cryer and Miller discuss renewal rates of magazine subscriptions (1991, p. 93). In one year the overall renewal rate was increased between January and February. Yet the renewal rate in every category went down. With six kinds of subscriptions, the cause is difficult to see. But if we eliminate all types of subscriptions except the two largest groups, we find the overall renewal rate was 53.4%. The overall rate was 47.9% in January and 67.1% in February. The two-month renewal rate for regular renewal was 78.3% while that for subscription agents was 20.6%. The difference in renewal rates by type of subscription (67.7 percentage points) is much greater than the difference in renewal rates by month (19.5 percentage points). Thus to understand the month-to-month difference, we must take into account the type of subscription. This example shows that even a time difference is susceptible to Simpson's Paradox.

Figure 5. Renewal rates by month and subscription



11. RESULTS OF TEACHING

Following are some observations of the results of teaching students about the Cornfield conditions:

1. *Students found the algebraic form of Cornfield's condition to be unintuitive.* To better illustrate these differences, a graphical technique was developed. The overall probability of the effect, P(E), was used as the base line. The four probabilities being compared were grouped so that percentage point differences were visually evident (See Figures 3, 4 and 5).

2. *Students found this graphical device (Figures 3, 4 and 5) to be visually intuitive.* It seems to be a simple and sensible way to measure the importance -- the explanatory power -- of a confounding factor.

3. *Students need extensive reinforcement to see that associations can be conditional.* They don't see arithmetic as conditional, so why should statistics be different.

4. *Students seem to understand the problem of association reversal better when describing the association as "spurious" rather than as "biased".* The concept of 'bias' implies error, whereas the concept of 'spurious' better captures the spirit of being true in one sense, but not in another. 'Spurious' indicates 'real' but lacking in authenticity. Knowing an association can be unbiased (true) but still be spurious gives them a more powerful way of evaluating an association. Furthermore,

knowing an association can be spurious sets the stage for partial correlation and multivariate regression.

5. *Students who are trained in this way seem better able to appreciate the distinction between total and partial correlations in multivariate regression.* They are more concerned about what is taken into account.

12. POSSIBLE OBJECTIONS

Following are some arguments against featuring Simpson's Paradox in teaching introductory statistics:

1. *Simpson's Paradox is unimportant.* It is omitted from many introductory texts. When present, it is sometimes just a problem or an optional section. Even if in the text, teachers often skip it. Reply: True, these are signs of unimportance, but they are not arguments.

2. *In observational studies, Simpson's Paradox is always possible. There is no known statistical test for confounding*. Reply: True but with Cornfields' minimum effect-size conditions we can eliminate many confounders and thus strengthen an inductive argument.

3. *Simpson's Paradox is seldom encountered in doing real statistical studies.* Reply: Yes, but the observational studies most susceptible to Simpson's Paradox are often the studies used as evidence for important changes in public policy (c.f., second hand smoke).

4. *There are too many other statistical concepts that are more fundamental.* Reply: Fundamentality depends on the goal. Simpson's Paradox is not fundamental if the goal is to reason deductively about statistical inference, but is most fundamental if the goal is to reason *inductively* from association to direct causation.

5. *Typically, there are multiple confounding factors.* Reply: True. This single-factor emphasis should be used as an introduction to multiple regression.

13. CONCLUSION

If students are to understand proper inductive reasoning about causality in observational studies, they must understand Simpson's Paradox. Understanding the necessary condition (minimum effect size) for a reversal of a spurious association is the key to proper understanding of Simpson's Paradox. Without proper understanding of the necessary condition, Simpson's Paradox can be a doorway to subjectivism (i.e., if an argument is not deductively valid, nothing is certain and anything is possible). Thus Cornfield's conditions are a most important statistical contribution to human thought.

REFERENCES

Agresti, A. 1984. *Analysis of Ordinal Categorical Data.* John Wiley.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959). *Smoking and lung cancer: Recent evidence and a discussion of some questions.* Journal of the National Cancer Institute, 22, 173-203.

Cryer, Jon and Robert Miller (1991), *Statistics for Business: Data Analysis and Modelling.* PWS-Kent.

Fisher, Ronald (1958). *Letter to the Editor.* Nature.

Freedman, David, Robert Pisani, Roger Purves and Ani Adhikari. *Statistics* 2nd ed. W.W. Norton & Co.,

Friedman, Gary (1994). *Primer Of Epidemiology* 4th ed., McGraw Hill

Kelley, David (1994). *The Art of Reasoning.* 2nd ed. W.W. Norton & Co.

Pearl, Judea (1999). *Why There Is No Statistical Test For Confounding, Why Many Think There Is, And Why They Are Almost Right.* Technical Report R-256. http://singapore.cs.ucla.edu/jp_home.html.

Rosenbaum, Paul R. (1995). *Observational Studies.* Springer-Verlag. P. 88.

Rosenbaum, Paul R. *Cornfield's Inequality.* Encyclopedia of Biostatistics.

Samuels, Myra (1993). *Simpson's Paradox and Related Phenomena*. Journal of the American Statistical Association, March, 1993, p. 81.

Schield, Milo (1998). *Statistical Literacy and Evidential Statistics.* ASA Proceedings of the Section on Statistical Education, p. 137.

Wonnacott, Thomas H. and Ronald J. Wonnacott (1979). *Econometrics,* 2nd ed., John Wiley.

Wonnacott, Thomas H. and Ronald J. Wonnacott (1990). *Introductory Statistics, 5th ed.* John Wiley

Acknowledgments: Paul Rosenbaum called attention to Cornfield's essay in his talk at the 1998 ASA JSM. Neil Thomason, University of Melbourne, provided David Lewis's graphical explanation of Simpson's Paradox. Bob Gibberd, University of Newcastle, suggested the use of simple conditional expansions in obtaining a simple difference comparison. Tom Wonnacott, University of Western Ontario, noted the use of regression to analyze bias due to omitted confounding factors. This paper is based on talks given in Australia at the University of Newcastle and at the Statistical Society of Australia, New South Wales. Donald Macnaughton made numerous suggestions on earlier drafts.

The author can be reached at schield@augsburg.edu. This paper is at www.augsburg.edu/ppages/schield.

Appendix I: Quotes from Cornfield et al (1959).

Measures of Difference:

"…we now discuss the use of relative and absolute measures of differences in risk. When an agent has an apparent effect on several diseases, the ranking of the diseases by the magnitude of the effect will depend on whether an absolute or a relative measure is used. …

Both the absolute and the relative measures serve a purpose. The relative measure is helpful in 1) appraising the possible noncausal nature of an agent having an apparent effect; 2) appraising the importance of an agent with respect to other possible agents inducing the same effect; and 3) properly reflecting the effects of disease misclassification or further refinement of classification. The absolute measure would be important in appraising the public health significance of an effect known to be causal.

The first justification for use of the relative measure can be stated more precisely, as follows:

> If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r.

Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X-producers among cigarette smokers must be at least 9 times greater than that of non-smokers. If the relative prevalence of hormone-X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect (Appendix A).

The second reason for using a relative measure may be phrased as follows:

> If two uncorrelated agents, A and B, each increase the risk of a disease, and if the risk of the disease in the absence of either agent is small (in a sense to be defined), then the apparent relative risk for A, r, is less than the risk for A in the absence of B.

The presence of other real causes thus reduces the apparent relative risk. If, for example, the relative risk of developing either disease I or disease II on exposure to A is the same in the absence of other causes, and if disease I, but not disease II, also has agent B present, then the apparent relative risk of developing disease I on exposure to A will be less than that for disease II (Appendix B).

The third reason for using a relative measure is:

> If a causal agent A increases the risk for disease I and has no effect on the risk for disease II, then the relative risk of developing disease I, alone, is greater than the relative risk of developing disease I and II combined, while the absolute measure is unaffected.

Appendix A

We feel obliged to give proof of the rather obvious statement on the magnitudes of relative risk because it has been suggested that the use of a relative measurement is merely "instinctive" and lacking in rational justification. Let the disease rate for those exposed to the causal agent B, be r_1 and for those not exposed, r_2, each rate being unaffected by exposure or nonexposure to the noncausal agent, A. Let $r_1 > r_2$. Of those exposed to A, let the proportion exposed to B be p_1, and of those not exposed to A, let the proportion exposed to B be p_2. Because of the assumed positive correlation between A and B, $p_1 > p_2$. Then

R_1 = rate for those exposed to A = $p_1r_1 + (1-p_1)r_2$

R_2 = rate for those not exposed to A = $p_2r_1 + (1-p_2)r_2$

(1) $R_1/R_2 = \{p_1r_1 + (1-p_1)r_2\} / \{p_2r_1 + (1-p_2)r_2\}$

Since $p_1 > p_2$ and $r_1 > r_2$, it follows that $R_1/R_2 > 1$.
From (1) we obtain

$$p_1/p_2 = R_1/R_2 + [r_2 /(p_2\, r_1)]\, [(1-p_2)(R_1/R_2)-(1-p_1)]$$

Since $p_1 > p_2$ and $R_1/R_2 > 1$, the second term on the right is positive and $p_1/p_2 > R_1/R_2$.

Since p_1/p_2 is the ratio of the prevalence of B among those exposed to A relative to that among those not so exposed, and R_1/R_2 is the apparent relative risk, r, the statement is proved.

On the other hand, if the absolute difference, $R_1 - R_2$, is used, the relationship, $(R_1 - R_2) = (r_1 - r_2)(p_1 - p_2)$, leads to no useful conclusion about $p_1 - p_2$.

Appendix B

The proof again is simple. Let r_{11} denote the risk of the disease in the presence of both A and B, r_{12}, the risk in the present of A and absence of B, $r_{21,}$ the risk in the absence of A and presence of B, and r_{22}, the risk in the absence of both A and B. It is reasonable to assume $r_{22} = 0$, but the less restrictive specification, $r_{22} < r_{12}\ r_{21} / r_{11}$ is sufficient for what follows. The proportion of the population exposed to B is denoted by p, and this, by hypothesis, is the same whether A is present or absent. Then

R_1 = rate for those exposed to A = $pr_{11} + (1-p)r_{12}$

R_2 = rate for those not exposed to A = $pr_{21} + (1-p)r_{22}$

and R_1/R_2 = apparent relative risk

$$\frac{R_1}{R_2} = \frac{r_{12}\,[1+\{[p/(1-p)](r_{11} / r_{12})\}]}{r_{22}\,[1+\{[p/(1-p)](r_{21} / r_{22})\}]}$$

Since $r_{22}/r_{21} < r_{12}/r_{11}$, the second factor is less than unity and $(R_1/R_2) < (r_{12}/r_{22})$ which proves the proposition."

[Editorial comment: In Appendix A, the result following (1) is obtained by multiplying (1) by the right-hand denominator, dividing by p_2r_1, and canceling r_1.]