BINARY CONFOUNDERS AS MATHEMATICAL OBJECTS: CONFOUNDER INFLUENCE AND CONFOUNDER INTERVALS

Milo Schield, Augsburg College and Thomas V.V. Burnham, Cognitive Consulting.
Dept. of Business Administration. Minneapolis, MN 55454

Abstract: Confounding is present in most observational studies. Yet by its nature, confounding is not generally present in the data. In order to use statistical associations as evidence for causal connections, one must try to take into account the influence of confounding. This paper reviews the role of confounding in the epic debate between Cornfield and Fisher on the statistical association between smoking and lung cancer and Cornfield's measure of the influence of an unobserved confounder in terms of a necessary condition. This paper extends the approach of Cornfield and Gastwirth to obtain defining conditions under which a binary confounder will nullify – render spurious – an association between binary variables when using a non-interactive (NI) linear OLS regression model. These defining conditions are used to derive necessary conditions for NI spuriosity and reversal. From these necessary conditions, simple tests are obtained to infer whether an association will be increased, decreased or reversed after controlling for a confounder. Using this non-interactive linear model, families of confounders are identified as mathematical objects based on their ability to nullify an observed relative prevalence. This paper also identifies the numerical properties of a binary confounder that would nullify a given association. Associations that can withstand a certain size confounder without being nullified are considered confounder resistant. This paper also identifies conditions under which the influence of a confounder can be shown as confounder intervals for an observed ratio and a given size confounder. Formulas for the upper and lower limits of confounder intervals are determined. In order to highlight the influence of potential confounders on relative risks or prevalences in observational studies, data analysts should accompany these measures with some measure of their susceptibility to confounding using either the size confounder that would nullify the association or the interval for a given size confounder.
Keywords: Epidemiology, Simpson's Paradox, nullify.

INTRODUCTION

Statistics studies the use of statistical associations as evidence for causal connections. While statistical associations may be a sign of causation, they can also be influenced by confounding in non-randomized studies and by randomness in any study. Statistics studies variation – random variation and systemic variation. Random variation is the basis for statistical inference; systemic variation is the basis for modeling.

Statistics has developed a vast literature on randomness as the basis for statistical inference, and on various kinds of models for systemic variation. But statistics has said very little to date about the influence of confounding since by their nature confounders are typically not present in the data and there is no theoretical model for a distribution of confounders. For a solid background on confounding, see Rosenbaum (2005).

The study of confounding is presented in five parts.

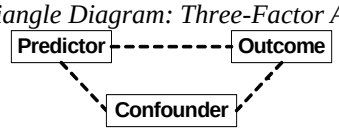
1. The first part reviews the epic debate between Cornfield and Fisher in the late 1950s involving smoking and lung cancer. To set the stage we first review the meaning of terms such as confounding and nullification. We then review the well-known defining condition for nullification by a confounder when the variables are continuous. Studying confounding involving continuous variables may indicate how confounding might be handled with binary variables. Then we return to the Fisher-Cornfield debate involving binary variables.
2. The second part reviews relationships involving differences in prevalences and develops associated necessary conditions.
3. The third part reviews relationships involving two or three binary variables, derives the slopes in an ordinary least squares non-interactive model, and then derives the conditions under which a ratio of prevalences is nullified by taking into account the influence of a confounder. Necessary conditions for nullification are derived from these defining conditions and are related to those derived by Cornfield and Gastwirth.
4. The fourth part introduces confounder resistance: the ability of an observed association to resist nullification by confounders of a given size.
5. The fifth part introduces confounder intervals to indicate the influence of a given size confounder.

1.1 CONFOUNDING

In statistics a **confounder** is a factor that associated with both the predictor and outcome in an association and that is not present in their analysis.[1]

A **triangle diagram**, see Figure 1, shows the relationships between these three related factors: a predictor, an outcome and a third related factor.

Figure 1. Triangle Diagram: Three-Factor Association



[1] In epidemiology a confounder is defined to exclude a mechanism: a third factor that is causally influenced by the predictor.

1.2 CONTINUOUS OUTCOMES

The influence of confounding involving continuous variables is well known. Consider modeling *E* on two continuous predictors *A* and *C*. When the partial correlation coefficient between *A* and *E* is zero then the relationship between A and E is said to be 'spurious' with respect to *C*. When the model is linear and non-interactive (NI), the regression coefficient relating *E* and *A* is proportional to $r_{AE,C}$, the partial correlation coefficient between *A* and *E* after controlling for *C*:[2,3]

Eq. 1 $$r_{AE,C} = (r_{AE} - r_{AC}\, r_{CE}) / \sqrt{(1 - r_{AC}^2)(1 - r_{CE}^2)}$$

NI spuriosity occurs when $r_{AE,C} = 0$. This implies that:

Eq. 2 $$r_{AE} = r_{AC}\, r_{CE}.$$

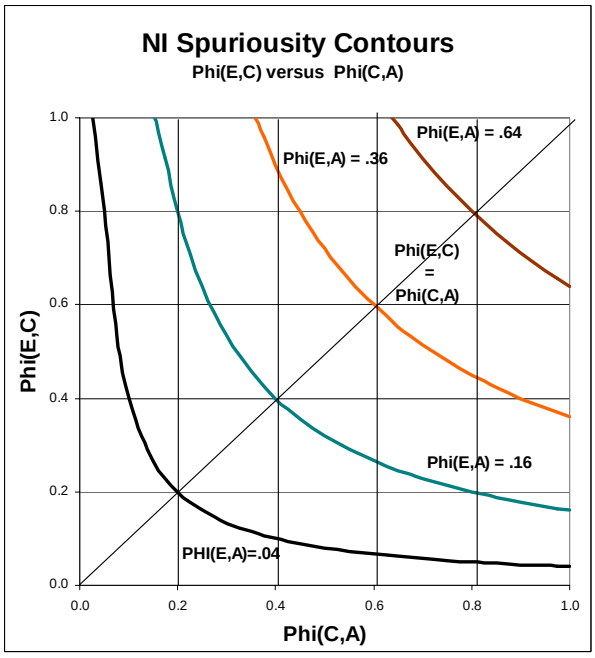
From this defining condition we obtain two necessary conditions:

Eq. 3 $$r_{AC} \geq r_{AE} \text{ and } r_{CE} \geq r_{AE}$$

But these two necessary conditions are not jointly sufficient; taken together they do not define the condition for nullification.

We can obtain a family of confounders by grouping them based on their ability to nullify an association of a given size. Figure 2 shows these families using *phi* instead of *r* to describe the correlation coefficient.

Figure 2: Families of Equal Nullification Power



Note that the point at which both correlation coefficients are jointly the smallest corresponds to the point at which they are equal. Thus, we might model this family by a single factor: the size of the correlation coefficient at that point. We could say that associations of size r_{AE} are resistant to confounders in the family having a "Confounder Size" less than Sqrt(r_{AE}).

Note what has been accomplished. First, a family of confounders has been identified based on their common effect: the power to nullify a given association. Second, this family of confounders has been summarized by a single value. Third, this value is in some sense representative of the values found in the family. Fourth, this value has the same units as the original association. In short, this integrative process results in greater comprehension by using simple metrics that are representative of the data at hand. Understanding this process is important for it foreshadows the process that will be used in identifying families of confounders involving binary variables.

1.3 FISHER-CORNFIELD DEBATE

In the 1950s, 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 – a minimum relative prevalence – for a confounding factor to explain away an observed association—assuming the association was 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

[2] Note: r_{AE} is the Pearson correlation coefficient between *E* and *A*.

[3] If $r_{CE} = 0$ then $|r_{AE,C}| > |r_{AE}|$. So, the association between *A* and *E* can not be nullified or reversed by such a confounder.

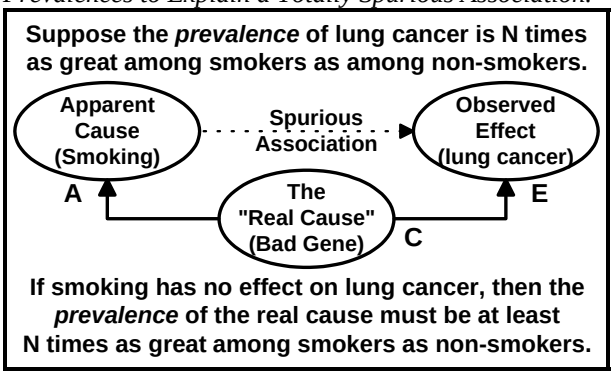
those exposed to A, relative to the prevalence among those not so exposed, must be greater than r.

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 tilde (~) preceding a letter is the complement of the condition (~A = non-A) so P(~A) = 1 - P(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 3 illustrates this for the case of smoking and lung cancer.

Figure 3. Necessary Relationship among Relative Prevalences to Explain a Totally Spurious Association.



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.[4]

1.4 IMPACT OF CORNFIELD'S CONDITION

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.

Schield (1999) said of Cornfield's condition:

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.

With this introduction to the debate, let us review the elements involved: the use of relative risk rather than correlation, and the mathematical model that was used to obtain Cornfield's necessary condition.

1.5 CORRELATION AND RELATIVE RISK

The dispute between Fisher and Cornfield involved data with binary outcomes. Subjects were labeled as either smokers or non-smokers. Their deaths were either due to lung cancer or they were not.

Of course it is possible to obtain a Pearson Correlation coefficient using binary variables. This form is commonly known as *Phi* (ϕ). But epidemiologists seldom use Pearson Correlation coefficients; they use relative risk (*RR*) and the odds ratio (*OR*). In the Dictionary of Epidemiology (1985), the article on the Pearson correlation coefficient notes that special varieties "have occasional uses in Epidemiology."

Statisticians might argue that correlation should not be used in 2×2 tables since correlation is properly defined only for continuous data where correlations can be generalized from samples to populations.

Abramson and Gahlinger (2001) give reasons why epidemiologists prefer other measures. "Unlike the odds ratio and Yule's Q, phi and lambda vary with the relative sizes of the case and control groups, and should

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

in general be used only if the cases and controls together make up a defined population, or comprise a representative sample of a defined population. The values of phi and lambda are then applicable to this specific population…. Misleading results may be obtained if the marginal totals are determined arbitrarily, as in case-control or cohort studies in which samples of arbitrary sizes are compared."

Yet even when the entire population is surveyed or when the samples are representative of the entire population, epidemiologists seem to avoid using correlations. One epidemiologist remarked that the *proportion of the variance* which the factor explains is obviously less relevant to the issue than the *proportion of the disease rate* which is explained. Granting that this is so, one wonders why. To see this we need to explore the world of binary variables.

1.6 NOTATION

This paper deals with confounder-induced spuriosity.[5] An association between two variables is *confounded* by a third if the third has an influence on their association. An association is *spurious* – of no effect – if it vanishes after taking a confounder into account. Let *E* be a binary effect and let *A* and *C* be binary predictors. The goal of this paper is to identify the conditions when the association between *A* and *E* becomes spurious (is nullified) or reverses (changes sign) after taking into account a confounder, *C*.

The variable name is used to indicate the values (e.g., *A* and *non-A*). Non-*A* is indicated by ~A. If *E* is cancer and *A* is smoker, then *P(E|~A)* is the prevalence of cancer for non-smokers.[6] In order to study double ratios (differences between, and ratios of, prevalences), this notation is also used:

Eq. 4 $DP(Y:X) \equiv P(Y|X) - P(Y|\sim X)$

Eq. 5 $RP(Y:X) \equiv P(Y|X) / P(Y|\sim X)$

$XRP(Y:X) \equiv RP(Y:X) - 1$

Eq. 6 $AFG(Y:X) \equiv [P(Y|X) - P(Y|\sim X)] / P(Y|X)$

$AFP(Y:X) \equiv [P(Y) - P(Y|\sim X)] / P(Y)$

The colon indicates that the following value and its complement are involved. Consider cancer *(E)*, smoking (*A*) and a cancer gene *(C)*. *DP(E:A)* is the differential prevalence of cancer for smokers vs. non-smokers. *RP(E:A)* is the relative prevalence, *XRP(E:A)* is the excess relative prevalence, of cancer for smokers vs. non-smokers. *AFG(E:A)* is the fraction of cancer cases in the exposure group (smokers) that are attributed to smoking. *AFP(E:A)* is the fraction of cancer cases in the sampled population that are attributed to smoking.

The selection of *A* vs. ~*A* and of *C* vs. ~*C* is arbitrary. This paper assumes they are selected so *DP(E:A)* > 0 and *DP(E:C)* > 0.[7] These selections do not determine whether *DP(C:A)* is positive or negative.

1.7 SMOKING AND LUNG-CANCER DEATHS

Consider the case of smoking and deaths due to lung cancer. Epidemiologists viewed the high relative risk of lung cancer for smokers ($RR \geq 9$) as strong evidence of a non-spurious association. See Cornfield (1959).

Suppose that Table 1 were a random sample of deaths.[8] We see that 5% of these deaths are due to lung cancer, that 10% of those who died are smokers, and that among the deceased the relative risk (*RR*) of dying due to lung cancer for smokers is 9.

Table 1: Deaths (hypothetical)

Deceased	~E:Other	E: Lung Cancer	Total
~A: Non-smokers	875	25	900
A: Smokers	75	25	100
Total	950	50	1000

The left column in Appendix A summarizes the algebraic identities between many of the common measures of association between two binary variables.[9] One form of the relation between *RR* (where *XRP*(*E*:*A*) = *RR*-1) and Phi (ϕ) is:

Eq. 7 $$\phi^2 = \left[\frac{P(E)}{P(A)}\frac{[1-P(A)]}{[1-P(E)]}\right]\left[\frac{P(A)XRP(E:A)}{P(A)XRP(E:A)+1}\right]^2$$

Given the data in Table 1, $\phi^2 = (9/19)(0.8/1.8)^2 = 0.094$, so $\phi = 0.306$. How could epidemiologists consider $RR \geq 9$ strong evidence of a non-spurious association if $\phi \approx 0.3$? Perhaps the problem is not the use of ϕ per se, but the use of $\phi^2 = 1$ as the standard.

1.8 RELATIVE PHI (ϕ)

What is the maximum value of ϕ given a certain prevalence, P(A), for an exposure factor? Eq. 7 specifies the relationship between the binary correlation coefficient ϕ, the relative risk *RR* (where *XRP*(*E*:*A*) = *RR*-1), the prevalence of the exposure factor P(A), and

[5] A spurious association can also be chance-based: due to sampling variability when there is no association in the population.

[6] Note that *P*(*X*) signifies prevalence or percentage – not probability.

[7] If *DP*(*E*:*A*) = 0 then reversal is not meaningful. If *DP*(*E*:*C*) = 0 or *DP*(*C*:*A*) = 0, then spuriosity and reversal are impossible (Eq. 45).

[8] This hypothetical data is not totally unrealistic. In the US in 1998, 7% (160,000) of all deaths (2.3 million) were due to lung cancer. In the US in 1999, 26% of those 12 and older smoked cigarettes. Statistical Abstract of the United States: 2001, Tables 105 and 190.

[9] *Cases* are subjects having the outcome of interest. Subjects are classified in the exposure or non-exposure groups, and in the case or non-case groups. In this discussion of *AFP*, *P(E)* and *P(A)*, the whole group is the population or a random sample thereof. *Prevalence* is a rate that doesn't involve a time interval (e.g., the unemployment rate, the exchange rate). Note that the Attributable Fraction in the Population (which has population in its name) can be calculated for a sample as well as for an entire population. Abramson (1994) discusses these measures. *RR*, *AFG* and *OR* are independent of the relative size of the exposure group, P(A), assuming *P(E|A)* and P(E|~A) are constant. Similarly, *RP* and *OR* are independent of the relative size of the cases, P(E). *AFP* and ϕ are dependent on the prevalence of the exposed subjects, P(A), and the prevalence of the cases, P(E).

the prevalence of the outcome P(E). When *RR* is infinite, Eq. 7 gives the maximum value of $|\phi|$ as

Eq. 8 $\phi^2_{Max} = [P(E)/P(A)]\{[1-P(A)]/[1-P(E)]\}$ [10]

In Table 1, the maximum value of ϕ^2 is 0.8.

Suppose we compare the observed ϕ with the maximum ϕ possible *given the observed prevalence of the exposure: P(A)*. This would compare the observed factor with the factor having the same *prevalence and* having the maximum relative risk*: RR* = ∞.

Eq. 9 $$\frac{\phi^2}{\phi^2 \max} = \left[\frac{P(A)XRP(E:A)}{P(A)XRP(E:A)+1}\right]^2 = AFP(E:A)^2$$

The attributable fraction in the population (*AFP*) is the fraction of cases that would be eliminated if that exposure factor were a necessary condition for the rate of cases above the base rate, *P*(*E*|~*A*), and if that exposure factor were eliminated.[11] Thus the attributable fraction in the population (*AFP*) with an exposure prevalence, P(A), is the same as this **relative correlation**: the observed ϕ relative to the ϕ of a genuinely necessary factor – P(E|~A) = 0 which implies RR = ∞ – that has the same exposure prevalence, *P(A)*.

Using equation j in Appendix A, we can see that for a given predictor prevalence, *P*(*A*), the relative risk, *RP(E:A),* increases monotonically with *AFP(E:A)*:

Eq. 10 $RP(E:A) = 1+AFP(E:A) / \{P(A)[1-AFP(E:A)]\}$

Thus the relative risk, *RP*(*E:A*) increases monotonically with ϕ taken relatively – relative to the maximum value of ϕ possible given a predictor prevalence of *P*(*A*), and an outcome prevalence of *P*(*E*). *RP*(*E:A*) has two benefits: it is independent of the size of the exposure group, *P(A)* and it increases monotonically with *AFP(E:A)* where *AFP(E:A)* is a relative correlation $|\phi/\phi_{max}|$: the observed correlation relative to that of a genuinely necessary factor having the same exposure prevalence, *P*(*A*).

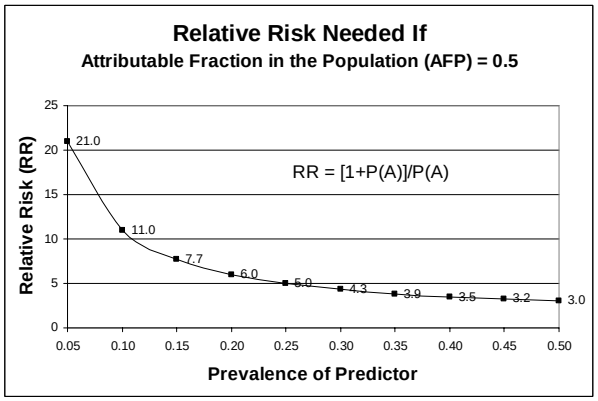
This limit on the maximum value of ϕ may explain why epidemiologists prefer to focus on the proportion of the case rate that is explained, AFP(E:A), rather than on the proportion of the variance that is explained, ϕ^2.

1.9 COMMENTS

The attributable fraction of cases among the exposed, *AFG,* has been misrepresented as the chance that a case among those exposed was caused by their exposure.[12] Suppose that *RP(E:A)* = 3 and *AFG(E:A)* = 67%. It has been claimed this means, "if a person has the disease in question and was exposed to the chemical in question, the probability that the exposure caused the person's disease is 67%." We disagree. If the exposure had caused 67% of the deaths among those exposed, then *AFG*(*E:A*) would be 67% and thus RR would be 3. But arguing the reverse begs the question.

On the other hand, the attributable fraction in the population is a relevant value for decision making. If the exposure were in fact a causal factor and a necessary condition, and if eliminating that factor eliminated only a small fraction of the cases, then doing so might not be justifiable if there are significant costs associated with that decision. The following graph shows the relative risk needed for the attributable fraction to be 50%:

Figure 4. Relative Risk Needed for Attributable Fraction in the Population



1.10 NECESSARY CONDITIONS

Cornfield et al. (1959) worked out the first necessary condition for nullification of an association between binary variables in terms of relative risk or prevalence. Indeed it was Cornfield (1951) who either created or popularized both Relative Risk and the Odds Ratio to better measure associations in Epidemiology.

The derivation of the first necessary conditions for spuriosity involving binary variables arose in the argument about whether smoking causes lung cancer. A clear association had been demonstrated. But was smoking a *direct cause* of cancer or was the association spurious – due to some confounder? In 1958, Fisher, a leading statistician and a smoker, argued that the smoking-cancer association might be confounded by genetics. He found an association for twins between the degree of twinship (identical or fraternal) and smoking preference. To reply, Cornfield modeled spuriosity by assuming smoking (*A*) had "no effect":

Eq. 11 $P(E|C,A) = P(E|C,\sim A) = P(E|C)$
Eq. 12 $P(E|\sim C,A) = P(E|\sim C,\sim A) = P(E|\sim C)$

We call these conditions "**cross-A rate equalities**" because the rates are equal across A (conditionally independent of A). Note that the restrictions are not on C, but on *P*(*E*|*C*) and *P*(*E*|~*C*). In equations derived

[10] This maximum value is like an odds ratio using margin ratios.
[11] ϕ^2=*AFP* times its *diagonal exchange* partner. Appendix A, Eq. A9.
[12] Source: www.toxictorts.com/relrisk.htm. "*Relative Risk: Proving Causation by the Numbers*" by Raphael Metzger, Esq.

from Eq. 11 and Eq. 12, *C* is replaced by c to indicate these restrictions. Cornfield et al derived a variation of this equation: [13]

Eq. 13 $$RP(E:A) = \frac{[P(c \mid A)\ XRP(E:c)] + 1}{[P(c \mid \sim A)\ XRP(E:c)] + 1}$$

From their variation, Cornfield et al (1959) proved that if the observed association is spurious then the confounder must satisfy this necessary condition:

Eq. 14 $RP(E:A) < RP(c:A)$ or $\frac{P(E \mid A)}{P(E \mid \sim A)} \leq \frac{P(c \mid A)}{P(c \mid \sim A)}$

Cornfield et al. (1959) replied to Fisher (italics added):

> "Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer [$RP(E:A) = 9$], 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 [$RP(C:A) > 9$]."[14]

Fisher never replied. Based on a wide range of epidemiological data, public health officials then asserted that smoking was "causally related" to lung cancer.

Using the cross-A rate equality conditions (Eq. 11, Eq. 12), Cornfield also derived a difference equality:

Eq. 15 $DP(E:A) = DP(E:c)\ DP(c:A)$

Thus, if the association between smoking and cancer is spurious, then the differential cancer prevalence for smokers vs. non-smokers, $DP(E:A)$, must equal the differential cancer prevalence for cancer-gene carriers vs. non-carriers, $DP(E:c)$, times the differential cancer-gene prevalence for smokers vs. non-smokers, $DP(c:A)$. Cornfield did not see this as useful.[15]

Gastwirth (1988) used Cornfield's "no effect" assumption to derive an expression for spuriosity:

Eq. 16 $$RP(c:A) = RP(E:A) + \frac{XRP(E:A)}{P(c \mid \sim A) XRP(E:c)}$$

Cornfield's condition follows from this since the fraction is positive. From a form of this equation Gastwirth derived a second necessary condition:[16]

Eq. 17 $RP(E:A) \leq RP(E:c)$ *or* $\frac{P(E \mid A)}{P(E \mid \sim A)} \leq \frac{P(E \mid c)}{P(E \mid \sim c)}$

This necessary condition can also be derived from Eq. 13 which has the form, $Z = [U(Y\text{-}1)+1]/[V(Y\text{-}1)+1]$. U>0, V>0, Y>1. Since [V(Y-1)+1] > 1, [U(Y-1)+1]/[V(Y-1)+1] < [U(Y-1)+1]. So, $Z < [U\cdot(Y\text{-}1)+1]$. Since $U < 1$, $Z < (Y\text{-}1)+1$. So $Z < Y$ and RP(E:A) < $RP(E:C)$.

If the smoking-cancer association is due to a gene, this condition means that the relative prevalence of cancer among smokers vs. non-smokers [$RP(E:A)$] must be less than or equal to the relative prevalence of cancer among those with vs. without the gene [$RP(E:c)$].

Note that these two necessary equations (Eq. 14 and Eq. 17) were derived on the basis that the predictor had "no effect" as defined in Eq. 11 and Eq. 12. Note also the similarity between these two necessary conditions (Eq. 14 and Eq. 17) and the two necessary conditions derived earlier (Eq. 3) for the continuous case involving correlations.

2. SPURIOUS DIFFERENCES

Although Cornfield et al saw little value in his difference conditions, they are easy to derive and they are easy for students to calculate and understand.

The conditions under which a difference in two ratios can be nullified can be derived in various ways. The following presents three different approaches: the Cross-Rate equality approach, a regression approach, and a partial-correlation coefficient approach. While all three generate the same equations, the different approaches have their unique strengths and weaknesses.

2.1 Cross-Rate Equality

A sufficient condition for "no effect" is cross-A rate equality: P(E|C,A) = P(E|C,~A) = P(E|C). And P(E|~C,A) = P(E|~C,~A) = P(E|~C). These give P(E|A) = P(E|C,A) P(C|A) + P(E|~C,A) P(~C|A). P(E|~A) = P(E|C,~A) P(C|~A) + P(E|~C,~A) P(~C|~A). P(~C|A) = 1 - P(C|A). These give Eq. 18. Subtraction gives Eq. 19..

Eq. 18 P(E|A) = P(E|C) P(C|A) + P(E|~C) P(~C|A)
P(E|~A) = P(E|C) P(C|~A) + P(E|~C) P(~C|~A)

Eq. 19 DP(E:A) = DP(E:C) DP(C:A)

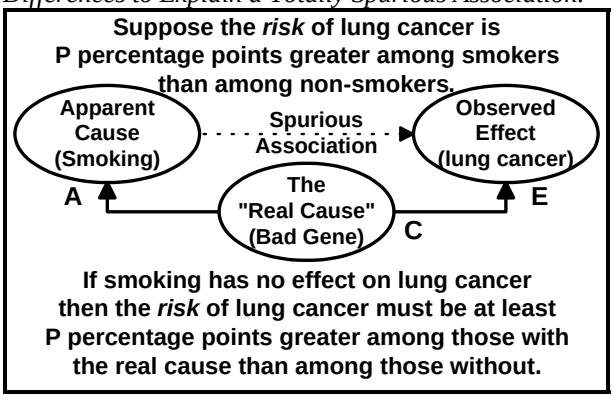
Since DP(C:A) ≤ 1,

Eq. 20 DP(E:C) ≥ DP(E:A)

Cornfield et al derived the risk-difference condition in Eq. 19 but dismissed it saying it "leads to no useful conclusion." This paper argues that this risk-difference condition is extremely useful as shown in Figure 5.

[13] In Eq. 13, $P(c|A) > P(c|\sim A)$ since $RP(E:A) > 1$ and $XRP(E:c) > 0$.
[14] Appendix A of Schield (1999) replicates Cornfield's derivation.
[15] "if the absolute difference, R1 - R2, is used, the relationship, R1-R2 = (r1-r2)(p1-p2), leads to no useful conclusion about p1-p2."
[16] Gastwirth (1988) attributed this condition to Cornfield. But Gastwirth first published it, so we call it the Gastwirth condition.

Figure 5. Necessary Relationship among Absolute Differences to Explain a Totally Spurious Association.



2.2 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, let b_1 be the coefficient involving A (the "direct effect"); let b_2 be the coefficient involving C and let b_3 be the coefficient regressing C on A so that E = b1 A + b2 C + b3 C(A).

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 b_3):

Eq. 21 $b_0 = b_1 + (b_2\, b_3)$

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 Cornfield'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

Eq. 22 $b_0 = b_2\, b_3$

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:

Eq. 23 $b_0 = P(E|A) - P(E|\sim A) = DP(E:A)$
Eq. 24 $b_3 = P(C|A) - P(C|\sim A) = DP(C:A)$
Eq. 25 $b_2 = P(E|C,A) - P(E|\sim C,A)$
Assuming A has "no effect" on E, $b_2 = P(E|C) - P(E|\sim C) = DP(E:C)$. If $b_1 = 0$, then $b_0 = b_2\, b_3$ and we obtain Eq. 19.

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 Eq. 20.

2.3 Partial Correlation Coefficients

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

Eq. 26 $r_{AE,C} = \{r_{AE} - [r_{AC}\, r_{CE}]\} / \sqrt{[(1-r^2_{AC})\,(1-r^2_{CE})]}$

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,

Eq. 27 $r_{AE} = r_{AC}\, r_{CE}$

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 are binary, the Pearson correlation coefficient reduces to phi (ϕ). See Eq. A10:

Eq. 28 $$\phi(E,A) = DP(E:A)\sqrt{P(A)P(\sim A)/[P(E)P(\sim E)]}$$

From Eq. 27, $\phi(E,A) = \phi(E,C)\,\phi(C,A)$. Thus,

Eq. 29 $$DP(E:A)\sqrt{P(A)P(\sim A)/[P(E)P(\sim E)]}$$

$$= DP(E:C)\,DP(C:A)\sqrt{\frac{P(C)P(\sim C)P(A)P(\sim A)}{P(E)P(\sim E)P(C)P(\sim C)}}$$

which reduces to Eq. 19

2.4 Comparison of Approaches

All three "difference" approaches give the same result as summarized by Eq. 19. The cross-rate equality approach is simplest. The regression approach is most powerful since it can be generalized to multiple confounding factors (Wonnacott and Wonnacott 1979, p. 415). The partial correlation coefficient approach clearly shows the binary form of a well known relationship involving continuous variables.

Eq. 20 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). Stu-

dents need only compare two simple differences measured in percentage points as shown in this equation:

Eq. 30 $[P(E|C) - P(E|\sim C)] \geq [P(E|A) - P(E|\sim A)]$

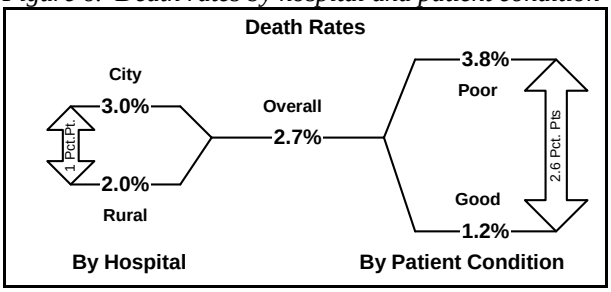
If this holds, then one has a definite reason to be concerned about a possible Simpson's Paradox reversal.

2.5 TEACHING SIMPSON'S PARADOX

For several years students in introductory statistics have been 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 hospital 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 6.

Figure 6. Death rates by hospital and patient condition

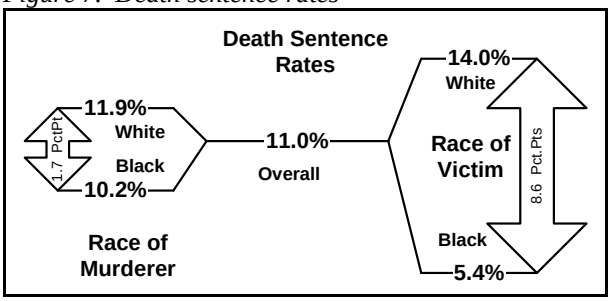


Now consider a plausible confounding factor: the condition of the patient's health. We find that overall the death rate among patients in poor condition is 3.8% while that among patients in good condition is 1.2%.

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.

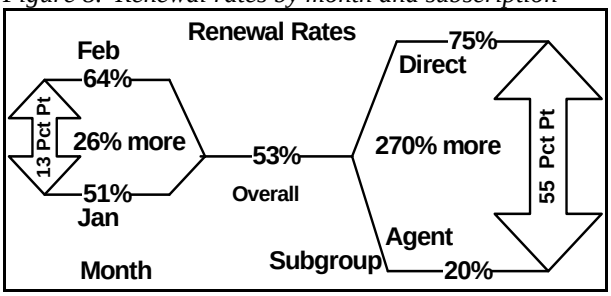
2. In a group of convicted murderers, the death penalty was given for 11.9% of white murderers and 10.5% of black murderers (Agresti 1984). Based on this data, one might argue that the legal system is biased against whites. However, when the sentences are classified by the race of the victim, the death 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 7.

Figure 7. Death sentence rates



3. Cryer and Miller (1991, p. 93) discuss renewal rates of magazine subscriptions. 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 confounder is difficult to see. But if we eliminate all types of subscriptions except the two largest groups, we obtain the results shown in Figure 8. The combined renewal rate for both months combined was 53%. The combined rate was 51% in January and 64% in February. The two-month renewal rate for regular renewal was 75% while that for subscription agents was 20%. The difference in renewal rates by type of subscription (55 percentage points) is much greater than the difference in renewal rates by month (13 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 8. Renewal rates by month and subscription



3. CHOICE OF MODELS

Recall that two the three approaches used in the preceding section involved the use of a linear model. Since we could obtain the same results without assuming linearity that feature was less critical. But in the coming sections, the choice of the model is more critical.

In deciding how to model an association between three binary variables, there are three items to consider.

First, the outcome variable is binary. Second, the data points involved are actually averages from which group averages are linear combinations based on the weights involved. Third, the predictors are binary.

The first point (the binary nature of the outcome variable) supports the use of a logistic model rather than an ordinary least squares (OLS) linear model. But the second point supports the opposite. The OLS linear model always models group averages as weighted averages of the sub-groups whereas the logistic model does so only in a few special cases. Thus the OLS linear model is superior for modeling the relation between a weighted average and its components. The third point supports the use of an OLS linear model for two reasons. First, since the predictor variables are binary their range is limited so that only a small segment of the full logistic model would be used. Secondly, the ability of a linear model to give unacceptable outcomes is limited by the restriction on the predictor variables.

For these reasons, this paper uses an OLS linear model to summarize the association between three binary variables.

3.1 NON-INTERACTIVE SPURIOSITY

In the following models, the values of the variables are treated as continuous. Rather than use new notation, we ask readers to recognize that *E*, *A* and *C* can be continuous in Eq. 31, Eq. 32, Eq. 34 - Eq. 36, and in Figure 9, Figure 10 and Figure 11.

Consider modeling *E* on two continuous predictors *A* and *C*. When the regression coefficient between *A* and *E* is zero, that relationship is said to be 'spurious' with respect to *C*. When the model is linear and non-interactive (NI), the regression coefficient relating *E* and *A* is proportional to $r_{AE,C}$, the partial correlation coefficient between *A* and *E* after controlling for *C*:[17,18]

Eq. 31 $$r_{AE,C} = (r_{AE} - r_{AC}\, r_{CE}) / \sqrt{(1-r_{AC}^2)(1-r_{CE}^2)}$$

NI spuriosity occurs when $r_{AE,C} = 0$. This implies that:

Eq. 32 $$r_{AE} = r_{AC}\, r_{CE}$$

Schield (1999) applied this well-known condition for spuriosity to binary data and obtained this condition:

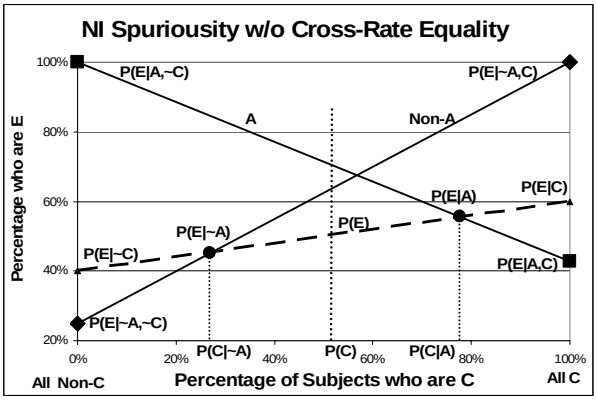
Eq. 33 $$DP(E:A) = DP(E:C)\, DP(C:A)$$

This condition (Eq. 33) is similar to the condition in Eq. 15, but without the cross-A rate equality assumption. $DP(C:A) > 0$ for NI spuriosity (since $DP(E:A) > 0$ and $DP(E:C) > 0$) and for NI reversal (defined in 3.3) as proven in 3.6 after Eq. 45.

3.2 CROSS-A VERSUS NI SPURIOSITY

Since both the cross-A rate equality condition and the NI model give similar results (Eq. 15 and Eq. 33), it may be worth explicating their difference. The difference equation (Eq. 33) can be written as equal slopes: $\Delta Y/\Delta X = DP(E:A)/DP(C:A) = DP(E:C)/(1\text{-}0)$. For other forms see Equations F9 in Appendix F. Figure 9 shows data that satisfies this slope condition.

Figure 9: Non-Interactive (NI) Spuriosity



In *Figure 9*, $P(\sim A,\sim C) = 8/20$, $P(\sim A,C) = 3/20$, $P(A,\sim C) = 2/20$ and $P(A,C) = 7/20$ so n = 20. $P(E|\sim A,\sim C) = 2/8$, $P(E|\sim A,C) = 3/3$, $P(E|A,\sim C) = 2/2$ and $P(E|A,C) = 3/7$ so P(E) = 10/20 = 50%.

In cross-A rate equality, $P(E|\sim A,C) = P(E|A,C) = P(E|C)$ and $P(E|\sim A,\sim C) = P(E|A,\sim C) = P(E|\sim C)$. So Figure 9 does not involve cross-A rate equality. $P(E|C)$ is always a weighted average of two rates: $P(E|A,C)$ and $P(E|\sim A,C)$. For cross-A rate equality, these rates are equal, so the weights don't matter. For non-interactive spuriosity, these rates can be unequal so the weights do matter.

3.3 NON-INTERACTIVE REVERSAL

Non-interactive (NI) reversal is readily seen using the regression approach presented by Wonnacott and Wonnacott (1990, Appendix 13-5). A regression model generates a line, $E(A)$ or $C(A)$, or a surface, $E(A,C)$:

Eq. 34 $E(A) = k_1 + b_0(E|A)\, A$
Eq. 35 $E(A,C) = k_2 + b_1(E|A,C)A + b_2(E|A,C)C$
Eq. 36 $C(A) = k_3 + b_3(C|A)\, A.$

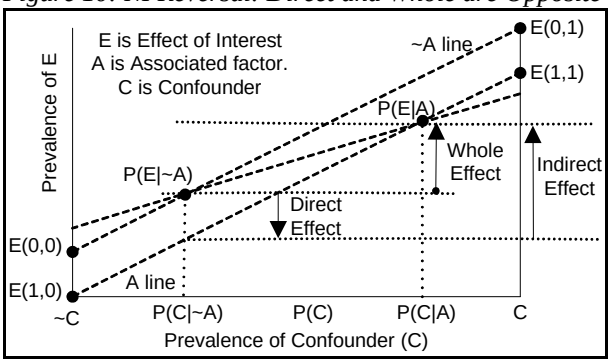
They showed these four slopes are related as follows:

Eq. 37 $b_0(E|A) = b_1(E|A,C) + [b_2(E|A,C)\ b_3(C|A)]$.
Eq. 38 *whole effect = direct effect + indirect effect.*

An NI model with two binary predictors generates a surface, $E(A,C)$, that has two parallel edges: The ~A line, $E(A=0,C)$; the A line, $E(A=1,C)$. See Figure 10.

[17] Note: r_{AE} is the Pearson correlation coefficient between *E* and *A*.
[18] If $r_{CE} = 0$ then $|r_{AE,C}| > |r_{AE}|$. So, the association between *A* and *E* can not be nullified or reversed by such a confounder.

Figure 10: NI Reversal: Direct and Whole are Opposite



The *~A* line always runs through $P(E|\sim A)$; the *A* line always runs through $P(E|A)$.

The **whole effect** of *A* on *E*, $b_1(E|A)$, is $DP(E:A)$ while $b_1(C|A)$ is $DP(C:A)$. The **direct effect** of *A* on *E* is the vertical distance between the two lines.

Non-interactive (NI) reversal of the association between A and E occurs when the signs of their coefficients are opposite in the one and two factor models:

Eq. 39 $b_0(E|A)\, b_1(E|A,C) < 0.$

Thus, NI reversal occurs when the sign of the whole effect is opposite the sign of the direct effect. Since $DP(E:A) > 0$ the whole effect is positive and the direct effect is negative. If $DP(C:A) > 0$, the *A* line lies beneath the *~A* line: a geometric condition for NI reversal.

3.4 DEFINING AND NECESSARY CONDITIONS

Non-interactive (NI) spuriosity is also defined by:

Eq. 40 $b_1(E|A,C) = 0.$

Although correlation ($r_{AE,C} = 0$ or Eq. 32) is a primary defining condition, Eq. 40 is a direct corollary.[19] Appendix F contains consequences of Eq. 40. Appendices B through E give details on NI modeling.

If the association between *A* and *E* is NI spurious, then $b_2(E|A,C) = DP(E:C)$ as shown in footnote 64, the direct effect is zero, the whole effect equals the indirect effect, and we obtain Eq. 33.

NI spuriosity and NI reversal are closely related. The defining condition for NI spuriosity (Eq. 40) is a boundary of the defining condition for NI reversal (Eq. 39). Since $b_1(E|A) = DP(E:A)$ and since we are assuming that $DP(E:A) > 0$, we can state the defining condition for NI reversal as:

Eq. 41 $b_1(E|A,C) < 0.$

3.5 BENEFIT OF NECESSARY CONDITIONS

When little is known about the confounder, necessary conditions can be weaker but more useful.

Textbooks seldom indicate a way to estimate whether an observed association is spurious. After studying Simpson's Paradox, one student concluded one should never place any trust in 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, and most students studying statistics are in fields that deal primarily with observational studies. Furthermore, an increasing amount of health data is obtained from observational studies, so students need to learn how to deal with associations based on observational studies.

To help students better understand how a statistically significant association can still be spurious, we need to focus on the necessary conditions for spuriosity.

3.6 JOINT NECESSARY CONDITIONS

A condition that is necessary for NI spuriosity, $b_1(E|A,C) = 0$ may not be necessary for NI reversal, $b_1(E|A,C) < 0$.[20] One condition necessary for both is the following:[21]

Eq. 42 $b_1(E|A,C) \le 0.$

Any condition that satisfies this is necessary for both $b_1(E|A,C) = 0$ and for $b_1(E|A,C) < 0$.[22]

From Eq. E1c in Appendix E, it follows that:[23]

Eq. 43 $b_1(E|A,C) = K1\ [DP(E:A) - DP(C:A)\, DP(E:C)]$
where $K1 = 1/[1-DP(C:A)DP(A:C)]$.

Since $K1 > 0$, combining the joint condition (Eq. 42) with this form of b_1 gives this necessary condition:

Eq. 44 $DP(E:A) \le DP(C:A)\, DP(E:C).$

Since $DP(E:A) > 0$ and $DP(E:C) > 0$, it follows that $DP(C:A) > 0$, so $RP(C:A) > 1$, for both NI spuriosity and NI reversal. Since $0 < DP \le 1$,[24]

Eq. 45 $DP(E:A) \le DP(C:A)$ and $DP(E:A) \le DP(E:C).$

Similarly structured relations involving correlation coefficients are obtained from Eq. 31.[25]

From Eq. E2c in Appendix E, it follows that:

[19] If r = 0, then b = 0 since b = r (Sy/Sx).

[20] Necessary conditions exist for one that are not necessary for the other. $RP(C:A) < P(E|A) / [P(E|\sim A)-P(E|\sim C)]$ (from Eq. F12) is necessary for NI spuriosity, but not for all NI reversals.

[21] Necessity can be confused with sufficiency. Note that b1>0 is sufficient to make both b1=0 and b1<0 false. Hence ~(b1>0) or b1≤0 is necessary for both. The "joint" applies to being necessary for NI spuriosity and NI reversal. This joint necessity does not apply to b1=0 and to b1 < 0 simultaneously – since that is impossible.

[22] If a joint necessary condition is L ≤ R then an increase in R or decrease in L makes NewL < NewR a necessary condition for both. If a necessary condition is false, then the conclusion is false.

[23] Recall that the whole effect is $b_1(E|A) = DP(E:A)$. If $K1 = 1$, the indirect effect is $DP(C:A)\, DP(E:C)$, but this is a degenerate case.

[24] If $DP(C:A) = 1$, we have co linearity: a non-useful degenerate case.

[25] $DP(E:A) > 0$ and $DP(E:C) > 0$, so $r_{AE} > 0$ and $r_{CE} > 0$. Since $b_1(E|A,C)$ is proportional to $r_{AE,C}$, applying Eq. 42 to Eq. 1 gives $r_{AE} \le r_{AC}\ \ r_{CE}$ as a necessary condition for NI spuriosity and reversal. So $r_{AE} \le r_{AC}\, r_{CE,}$ $r_{AE} \le r_{AC,}$ and $r_{AE} \le r_{CE}$ are necessary for NI spuriosity and reversal. These are analogs of Eq. 44 and Eq. 45.

Eq. 46 $b_1(E|A,C)=K2[AFP(E:A)-AFP(C:A)AFP(E:C)]$

where $K2 = P(E)/\{P(A)[1-DP(C:A)DP(A:C)]\}$.

Since $K2 = P(E)K1/P(A)$, $K2 > 0$. Combining the joint condition (Eq. 42) with this form of b_1 gives this necessary condition:

Eq. 47 $AFP(E:A) \le AFP(C:A)\, AFP(E:C)$.

$AFP(E:A)$ is the fraction of E attributable to A in the population. Since $0 < AFP < 1$, [26]

Eq. 48 $AFP(E:A) < AFP(E:C)$;
$AFP(E:A) < AFP(C:A)$.

From Eq. E3c in Appendix E, it follows that:

Eq. 49

$$b_1(E \mid A,C) = K3\{XRP(E:A)[P(C \mid\sim A)\, XRP(E:C)+1] - [P(C \mid\sim A)\, XRP(C:A)\, XRP(E:C)]\}$$

when $K3 = P(E)/\{[1-DP(C:A)DP(A:C)][P(A)XRP(E:A)+1][P(C)XRP(E:C)+1]\}$

$K3 = P(A)K2/\{[P(A)XRP(E:A)+1][P(C)XRP(E:C)+1]\}$ so K3 > 0. Combining the joint condition (Eq. 42) with this form of b_1 gives this necessary condition:

Eq. 50 $$XRP(E:A) \le \frac{XRP(C:A)\, P(C \mid\sim A)\, XRP(E:C)}{1+[P(C \mid\sim A)\, XRP(E:C)]}$$

The denominator is more than 1; the product of the first two factors in the numerator is less than 1.[27] Replacing both with 1 gives a necessary condition that is a generalization of the Gastwirth-Cornfield condition (Eq. 17):

Eq. 51 $XRP(E:A) < XRP(E:C)$, $RP(E:A) < RP(E:C)$.

In Eq. 50, the denominator is greater than 1, so the inequality remains if we replace it with 1. This generates:

Eq. 52 $XRP(E:A) < XRP(C:A)\, P(C|\sim A)\, XRP(E:C)$.

If $XRP(C:A)\, P(C|\sim A) < 1$, Eq. 52 is stronger than Eq. 51. If $XRP(E:C)\, P(C|\sim A) < 1$, Eq. 52 is stronger than Eq. 55.

From Eq. E4c in Appendix E:

Eq. 53 $b_1(E|A,C) = -K4\{[P(C)\, XRP(E:C)\, XRP(C:A)] - XRP(E:A)[P(C)\, XRP(E:C)+1+P(A)\, XRP(C:A)]\}$

if $K4 = P(E)/\{[1-DP(C:A)DP(A:C)][P(A)XRP(E:A)+1][P(C)XRP(E:C)+1][P(A)XRP(C:A)+1]\}$

Since $K4 = K3/[P(A)XRP(C:A)+1]$, $K4 > 0$. Combining the joint condition (Eq. 42) with this form of b_1 gives this necessary condition:

Eq. 54 $$XRP(E:A) \le \frac{P(C)XRP(C:A)XRP(E:C)}{P(C)XRP(E:C)+P(A)XRP(C:A)+1}.$$

Since the items being added in the denominator are positive, we can retain the inequality by retaining any one of them. Doing this from left to right gives these three necessary conditions:

Eq. 55 $XRP(E:A) < XRP(C:A)$,
Eq. 56 $XRP(E:A) < [P(C)/P(A)]\, XRP(E:C)$,
Eq. 57 $XRP(E:A) < P(C)\, XRP(C:A)\, XRP(E:C)$.

Eq. 55 is a generalization of Cornfield's condition (Eq. 14). Eq. 56 is more restrictive than the generalized Gastwirth-Cornfield condition (Eq. 51) if $P(C) < P(A)$. Eq. 57 is less restrictive than Eq. 52 but might be more useful.[28] Since $P(C) \le 1$, Eq. 57 yields Eq. 58:[29]

Eq. 58 $XRP(E:A) < XRP(C:A)\, XRP(E:C)$.

For the case of smoking and cancer, the generalization (Eq. 51) of the Gastwirth-Cornfield condition means that if this association were spurious and $RP(E:A)$ were 9, then $RP(E:C)$ must be greater than 9 for a hypothetical genetic confounder. But if the prevalence of such a genetic confounder, $P(C)$, was 10%, and the smoker prevalence, $P(A)$, was 40%, then this new condition (Eq. 56) would require $RP(E:C) > 33$.

3.7 "NO EFFECT" SPURIOSITY

Under NI spuriosity, the two cross-A rate differences, $DP(E:A|C)$ and $DP(E:A|\sim C)$[30], must either be opposite in sign (Figure 9) or zero (cross-A rate equality).

In Appendix D, it is shown that any instance of cross-A rate equality must involve NI spuriosity. Since Figure 9 is an example of NI spuriosity which does not involve cross-A rate equality, we infer that cross-A rate equality is a special case of NI spuriosity.

3.8 GEOMETRY OF NI REVERSAL

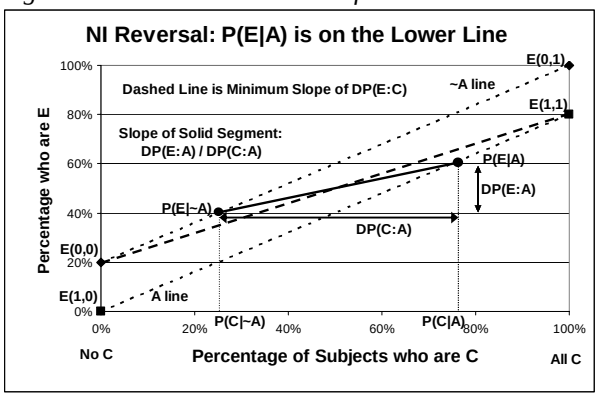
Eq. 41 gives a defining condition for NI reversal. Using Eq. 43 with Eq. 41 gives this form:

Eq. 59 $DP(E:A) / DP(C:A) < DP(E:C)$.

Figure 11 illustrates this condition graphically. The light dotted lines are the edges of the $E(A,C)$ surface for A and $\sim A$ where the A line lies below the $\sim A$ line. $P(E|C)$ is between $E(0,1)$ and $E(1,1)$; $P(E|\sim C)$ is between $E(0,0)$ and $E(1,0)$. See Eq. D6. $DP(E:A)/DP(C:A)$ is the slope of the dark solid segment. The slope of the dashed line, $[E(1,1)-E(0,0)]/1$, is the maximum of $DP(E:A)/DP(C:A)$ and the minimum of $DP(E:C)/1$.[31]

[26] $DP(E:A)=AFP(E:A)\, P(E)/P(A)$. $DP(E:A)>0$ implies $AFP(E:A)>0$.
[27] $[XRP(C:A)\, P(C|\sim A)] = [P(C|A) - P(C|\sim A)] < P(C|A) < 1$.
[28] This is more restrictive than Eq. 51 and Eq. 55 if both XRPs < 1. It is more useful than Eq. 52 or Eq. 56 if $P(C|A)$ or $P(C)$ are unknown.
[29] $RP(E:A)-1 < [RP(C:A)-1][RP(E:C)-1] = [RP(C:A)\, RP(E:C) - RP(C:A) - RP(E:C) + 1] < RP(C:A)\, RP(E:C) - 1 - 1 + 1$.
[30] $DP(Z:X|\sim Y) \equiv [P(Z|X,\sim Y) - P(Z|\sim X,\sim Y)]$ is analogous to Eq. 4.
[31] The maximum of $DP(E:A)/DP(C:A)$ and minimum of $DP(E:C)/1$ are achieved simultaneously only under NI spuriosity.

Figure 11: Geometric Condition for NI Reversal



A geometric condition for NI reversal is that the *A* line lies below the ~*A* line so *P*(*E*|*A*) lies on the lower line.

3.9 SIMPSON'S REVERSAL

Simpson's Paradox exists when the sign of association in *each* sub-group (*C* and ~*C*) is opposite the sign in the composite group. We define **Simpson's reversal** as the reversal occurring in Simpson's Paradox. When *DP(E:A) > 0*, this gives:[32]

Eq. 60 $DP(E:A|C) < 0, DP(E:A|\sim C) < 0$

Not all NI reversals involve a Simpson's reversal. Figure 9 illustrates an NI reversal but not all the signs of the sub-group differences are opposite that in the composite: $DP(E:A|C) < 0$ but $DP(E:A|\sim C) > 0$.

Simpson's reversal cannot occur without NI reversal as shown using this identity (Eq. B7 in Appendix B):

Eq. 61 $DP(E:A) = DP(C:A)\, DP(E:C) + X,$

Eq. 62 where $X = [P(E|A) - P(E|C)\, P(C|A) - P(E|\sim C)\, P(\sim C|A)] / P(\sim A)$.

In Eq. 62, $X < 0$ is another form of the defining condition for NI reversal (see Eq. 44). As defined in Eq. 60, a Simpson's reversal is sufficient to make $X < 0$ in Eq. 61. So, all instances of Simpson's reversal must involve an NI reversal. But not vice versa since a Simpson's reversal is not necessary for $X < 0$ in Eq. 62.

3.10 INFERENCES

The influence of a confounder, *C*, on an observed association between *A* and *E* can be inferred without doing the regression provided one has information on comparisons of single-predictor prevalences: *P*(*X*|*Y*). Assume as usual that values of *A* and *C* are selected so *DP*(*E:A*) > 0 and *DP*(*E:C*) > 0. We describe three cases given the (1) signs of three comparisons, (2) three relative differences, or (3) three simple differences.

#1: Direction of Change

Since $b_1(E|A) = DP(E:A)$, Eq. 43 can be rewritten as:

Eq. 63 $b_1(E|A,C) = K1[b_1(E|A) - DP(C:A)\, DP(E:C)]$.

The direction of change in the association between A and E can be inferred from the sign of *DP*(*C:A*):

Eq. 64 Decrease: $b_1(E|A,C) < b_1(E|A)$ if $DP(C:A) > 0$.

Eq. 65 Increase: $b_1(E|A,C) > b_1(E|A)$ if $DP(C:A) < 0$.

Since *XRP* has the same sign as *DP*, the sign of *XRP(C:A)* can be used to infer the direction of change.

#2: Non-Reversal[33,34]

If $XRP(C:A) > 0$, then $b_1(E|A,C) < b_1(E|A)$. In this case, an NI reversal, $b_1(E|A,C) < 0$, is precluded if any of the following are true:

Eq. 66 $XRP(E:A) > XRP(E:C)$, $XRP(E:A) > XRP(C:A)$ or $XRP(E:A) > XRP(C:A)\, XRP(E:C)$.

Eq. 66 follows from Eq. 51, Eq. 55 and Eq. 58. If all of the known elements of Eq. 66 are false, then an NI reversal is not precluded.

#3: Reversal

When rearranged, Eq. 59 gives this form of the defining condition for NI reversal:

Eq. 67 $DP(E:A) < DP(C:A)\, DP(E:C)$.

If Eq. 67 is true, then an NI reversal holds after taking the confounder into account; otherwise it does not.

3.11 AN EXAMPLE

The relevant outcome (*E*) is death, *A* is hospital (city vs. rural), and *C* is patient condition (poor vs. good).

(#1) Consider these qualitative comparisons. Death is more prevalent among patients at city hospitals that among those at rural hospitals; death is more prevalent among patients admitted in poor condition than among those admitted in good condition; and admission in poor condition is more prevalent among patients at city hospitals than among those at rural hospitals. It follows that the association between city hospitals and higher death rates is decreased after controlling for patient condition because all three *DP*s or *XRP*s are positive.

(#2) Consider these percentage comparisons. Death is 57% more prevalent among patients at city hospitals than among those at rural hospitals, so *XRP*(*E:A*) = 0.57. Death is 230% more prevalent for patients admitted in poor condition than for patients admitted in good condition, so *XRP*(*E:C*) = 2.3. And admission in poor condition is 200% more prevalent among patients at city hospitals than among patients at rural hospitals, so *XRP*(*C:A*) = 2.0. As in #1, the association between city hospitals and higher death rate is decreased by taking

[32] If the underlying rates were coplanar with cross-A rate difference equality, $DP(E:A|C) = DP(E:A|\sim C)$, then $E(1,1) = P(E|A,C)$, etc., See Eq. D2e. If so, Figure 11 would illustrate Simpson's reversal: $DP(E:A|C) = [P(E|A,C) - P(E|\sim A,C)] = [E(1,1) - E(0,1)] < 0$.

[33] Skip this step if *DP*(*E:A*), *DP*(*E:C*) and *DP*(*C:A*) are available.

[34] If *DP(C:A)* or *XRP*(*C:A*) are not available, they can be derived from a number of other statistics. For example
- $P(C|A) = [P(E|A) - P(E|A,\sim C)]/[P(E|A,C) - P(E|A,\sim C)]$
- $P(C|\sim A) = [P(E|\sim A) - P(E|\sim A,\sim C)]/[P(E|\sim A,C) - P(E|\sim A,\sim C)]$.

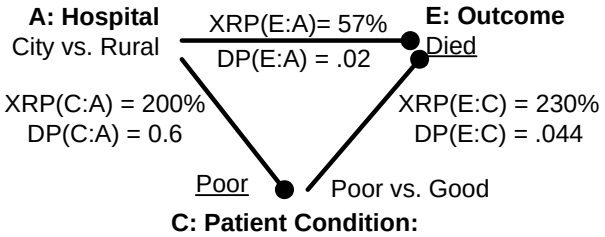
They can also be derived using *Phi*(*C,A*), *P*(*C*) and *P*(*A*):
- $[DP(C:A)]^2 = Phi^2(C,A)\{P(C)[1-P(C)]\}/\{P(A)[1-P(A)]\}$.

into account patient condition. In addition, it follows that a reversal of the association is not precluded, because *XRP*(*E:C*), *XRP*(*C:A*), and *XRP*(*E:C*) *XRP*(*C:A*) are each larger than the observed difference, *XRP*(*E:A*).[35]

(#3) Consider these percentage-point differences. Death is 2 percentage points more prevalent among patients at city hospitals than among those at rural hospitals, so *DP*(*E:A*) = 0.02. Death is 4.4 percentage points more prevalent for patients admitted in poor condition than for patients admitted in good condition, so *DP*(*E:C*) = 0.044. And admission in poor condition is 60 percentage points more prevalent among patients at city hospitals than among patients at rural hospitals, so *DP*(*C:A*) = 0.6. It follows that this association between city hospitals and higher death rates is reversed by taking patient condition into account, because the product of the two confounder-related simple differences, 0.6 times 0.044, is greater than the observed simple difference, *DP*(*E:A*) = .02.[36]

Figure 12 summarizes these comparisons. An underscore on a value or a dot at the end of connecting line indicates a common part numerator.

Figure 12: Comparison Triangle



Now consider similar data in which admission in poor condition is just 30 percentage points more prevalent among patients at city hospitals than among those at rural hospitals. It follows that the association between city hospitals and higher death rates is not reversed by taking into account patient condition, because the confounder linkages are not strong enough to reverse the association: (0.30)(0.044) is less than 0.02.[36]

COLLINEARITY

The problem of confounding is closely related to the problem of co-linearity. Both involve a third factor that is correlated with the two factors in an association. There is statistical test for either.

Unspecified alternate vs. specified. With an unspecified alternative hypothesis, there is no way to talk about power, whereas with a specified alternate involving both a separation distance and a standard deviation, there is a way to talk about power. The greater the separation, the smaller the standard deviation and the larger the sample size, the greater the statistical power.

Confounding is different. Sample size is not a factor

4.1 MINIMUM SIZE

Confounding seems omnipresent in observational studies. Things are tangled up and mingled together; everything seems to be connected to everything. An association between two variables is *confounded* by a third if the third is entangled with both these variables. While random assignment can statistically break such entanglements, most studies can not (or do not) involve random assignment. Without random assignment, there is no known statistical test for confounding. (Pearl, 1998)

Without knowing the distribution of confounders, there seems to be no way to say, "*there is a 20% chance that this observed association is due to confounding*" or "*If this association were entirely spurious, there is less than a 5% chance of seeing an association this big or bigger due to confounding*."

Finally, there seems to be no generally accepted way to talk about the nature or size of a confounder. We have no way to eliminate a variable in a regression by saying the association is beneath some minimum threshold for susceptibility to confounding. There is nothing comparable to the "5% level of significance." As a result there is no way to determine what size relative risk constitutes strong evidence for saying an association is not spurious.

Operationally, epidemiologists tend to disregard relative risks of less than three as being generally inadequate to withstand the influence of confounding.

Taubes (1995) noted the following:

> Sir Richard Doll of Oxford University, who once co-authored a study erroneously suggesting that women who took the anti-hypertension medication reserpine had up to a fourfold increase in their risk of breast cancer, suggests that no single epidemiologic study is persuasive by itself unless the lower limit of its 95% confidence level falls above a threefold increased risk. Other researchers, such as Harvard's Trichopoulos, opt for a fourfold risk increase as the lower limit. Trichopoulos's ill-fated paper on coffee consumption and pancreatic cancer had reported a 2.5-fold increased risk. "*As a general rule of thumb*," says Angell of the New England Journal, "*we are looking for a relative risk of three or more [before accepting a paper for publication], particularly if it is biologically implausible or if it's a brand-new finding*." Robert Temple, director of drug evaluation at the Food and Drug Administration, puts it bluntly: "*My basic rule is if the relative risk isn't at least three or four, forget it*."

While a relative risk of three may be a rule of thumb in some areas, lower ratios are being used. In concluding that second-hand smoke caused health problems,

[35] Note: to multiply percentages they must first be converted to fractions.

[36] If *DP*(*E:A*)/*DP*(*C:A*) < *DP*(*E:C*), the *A* line is below the ~*A* line so NI reversal will happen. When *DP*(*E:A*)/*DP*(*C:A*) > *DP*(*E:C*), the *A* line is above the ~*A* line so NI reversal is impossible.

the EPA relied on a relative risk of 1.2.[37] A relative prevalence of 1.25 is used to monitor adverse impact in hiring practices involving members of protected classes as identified by Title 7 of the 1964-1965 Civil Rights Act.[38] And in calculating the number of deaths attributable to various factors, epidemiologists are using relative risks less than 2. See Mokdad et al, (2004).

But as John Bailar, an epidemiologist at McGill University and former statistical consultant for the NEJM, points out, "*there is no reliable way of identifying the dividing line*." Taubes (1995). Thus, any rule of thumb such as RR > 3 requires justification.[39]

An important goal of science is to quantity the properties of entities. Since unmeasured confounders are difficult to deal with, one approach is to identify assumptions under which the properties of a confounder are completely determined by a single value. Given the complete specifications of a confounder one can then determine its' effects on a given association. Schield and Burnham (2003) have shown that specifying a binary confounder involves three values when using relative prevalences.

The first task is to identify a simple way to determine all the properties of a confounder by specifying just a single parameter: the confounder size. A second task is to identify what size confounder is required to nullify an observed association. Nullification is confounder-induced spuriosity.[40] An association is *spurious* – of no effect – if it vanishes after taking a confounder into account. A third task is to generate intervals for an observed relative risk based on the influence of a binary confounder of a given size.

4.2 DEFINING CONDITIONS FOR CONFOUNDER-INDUCED SPURIOSITY

Schield and Burnham (2003) obtained defining conditions under which an observed association would be made spurious by a confounder when using a non-interactive OLS model for binary data. The OLS non-interactive model has the form:

Eq. 68 $E(A,C) = b0 + b1\ A + b2\ C.$

Recall that E is the outcome of interest, A is the binary predictor and C is the binary confounder. Note that b1 is the partial regression coefficient between the outcome (E) and the binary predictor (A) after taking into account the influence of the confounder (C) using a non-interactive model.

If b1 = 0 then any association between A and E is spurious. There are many forms of this spuriosity condition as shown in Appendix E. The main problem is that at least three values must be specified for a confounder in order to determine its influence on an observed association.

Can we summarize these characteristics in the same way that we summarize a distribution by its center and spread? A first step is to see how these characteristics interact in rendering a given association spurious. Hopefully this will help us identify a single value that might be used to determine more than one property of a confounder. The goal is to identify summary characteristics that will identify confounders having the same nullifying strength on a relative prevalence in ways that are meaningful and useful.

4.3 ERROR

In almost every case where a proxy is used to measure the presence or absence of a related condition, there is error. Suppose we use a certain size confounder, S, as a fixed level. All associations which are not nullified or reversed by that confounder are considered to be S confounder resistant. All others are classified as non-resistant.

Assume that all these relative prevalences involve groups that are jointly exhaustive so as to exclude comparisons involving the top group with the bottom group. Assume also that the outcome is a natural binary and not a binary variable created by an arbitrary cutoff on a continuous distribution.

As mentioned, relative prevalences are either S confounder resistant or not. In some cases the predictor is spurious to some other factor while in other cases it is not. These situations can be summarized in a 2x2 table.

Table 2: Classification of Associations

SIZE S CONFOUNDER	Spurious	Non-Spurious
Non-Resistant	Non-causal	Type 2 Error
Resistant	Type 1 Error	Causal

A predictor that in fact is spurious and is also non-resistant to an S confounder is a true negative (non-causal). A predictor that is in fact non-spurious and resistant to an S confounder is a true positive (causal).

A predictor that is in fact spurious and is resistant to an S confounder is a false positive – Type 1 error. A predictor that is in fact non-spurious and is not resistant to an S confounder is a false negative – Type 2 error.

Consider a type 2 error involving a weak association between a causal factor and the outcome. How might

[37] www.forces.org/evidence/ets-whop/index.htm

[38] On August 25, 1978, four federal agencies (Department of Labor, Equal Employment Opportunity Commission, Office of Personnel Management and Department of Justice) issued the Adoption by Four Agencies of Uniform Guidelines on Employee Selection Procedures (1978). The Uniform Guidelines provide standards for fair selection procedures for EEO protected classes. Adverse impact in the selection process is presumed when the pass rate of applicants from a protected class with a low pass rate is less than 80 percent of the pass rate of applicants from the group with the highest selection rate. This is also referred to as the "four-fifths" rule.

[39] If one had a distribution of confounders, then one might be able to make probabilistic statements. Of the 24 cases cited by Taubes (1995), 80% have RR ≤ 3.

[40] A spurious association can also be chance-based: due to sampling variability when there is no association in the population.

this occur? Suppose that there are two causal factors. One is present throughout the population while the other is present only among a subgroup. For either of these factors to cause the outcome there must be a trigger – a fourth factor that is seldom present and thus the outcome is somewhat rare. Since those in the subgroup contain both causal factors, the risk of the outcome is greater in the subgroup than in the rest of the population. The size of the relative risk is determined entirely by the relative prevalence of the local causal factor to the prevalence of the background causal factor.

Now consider a type 1 error involving a strong association between a spurious factor and the outcome. How might this occur? Consider the population of those who read either fashion magazines or sport magazines. Suppose those who read fashion magazines are 100 times as likely to become pregnant as are those who read sport magazines. Assume that reading a magazine has no causal power to produce pregnancy so that both associations are spurious. In this case we recognize the confounder – gender. Women are more likely to read fashion magazines while men are more likely to read sport magazines. And pregnancy is much more prevalent among women than among men.

Following the use of terms in medical tests, specificity is 100% minus the percentage of Type 1 errors among those associations that are spurious while sensitivity is 100% minus the percentage of Type 2 errors among those associations that are non-spurious. A type 2 error is well known as the problem of co-linearity. In this type 2 error mentioned, there is a high correlation between reading fashion magazines and being a woman and in turn there is a much higher prevalence of pregnancy among women than among men.

The smaller the value of S, the greater the prevalence of positives: associations that can resist nullification or reversal by a confounder of that size. As S decreases, the greater the percentage of Type 1 errors among those associations that are spurious and the smaller the percentage of Type 2 errors among those associations that are non-spurious. As S decreases, the sensitivity increases, but the specificity decreases.

The greater the value of S, the greater the prevalence of negatives: associations that cannot resist nullification or reversal by a confounder of that size. As S increases, the smaller the percentage of Type 1 errors among those associations that are spurious and the greater the percentage of Type 2 errors among those associations that are non-spurious. As S increases, the sensitivity decreases, but the specificity increases.

If our goal were to minimize the prevalence of type 1 errors among those associations that are resistant to an S confounder, then we should set S higher. But if our goal were to minimize the prevalence of type 2 errors among those associations that are not resistant to an S confounder, then we should set S lower.

It would seem that in exploratory work, S would be set lower so as not to exclude any promising candidates. But in those areas involving compulsion such as regulations, legal liability, etc., S should be set higher so as to minimize the probability of type 1 error among those associations that are S resistant.

4.4 LOW PREVALENCE PREDICTORS

Note that as the prevalence of those treated or exposed decreases, the size confounder needed to nullify a given association increases.

For example, suppose we are looking for associations with death due to bulimia. We take the exposure group to be people who read early-20s fashion magazines while the non-exposure group is everyone else. The prevalence of the exposure group is small – probably less than 1% of the adult population. [No, this is a type 2 error. I need a good example of a type 1 error.]

4.5 INFLUENCE ON RELATIVE PREVALENCE

One form of the condition needed for a binary confounder to nullify an observed excess relative prevalence, $XRP(E:A)$, is given by Eq. F4b:

Eq. 69 $$P(C)\,XRP(E:C) = \frac{XRP(E:A)\{1+[P(A)\,XRP(C:A)]\}}{[XRP(C:A)-XRP(E:A)]}$$

For an observed excess relative prevalence $XRP(E:A)$ and predictor prevalence $P(A)$, this condition involves three other factors: $P(C)$, $XRP(C:A)$ and $XRP(E:C)$.

Notice how $XRP(E:C)$ is directly influenced by $P(C)$ for given values of $XRP(E:A)$, $P(A)$ and $XRP(C:A)$. If $P(C)$ is small, then $XRP(E:C)$ must be large and vice versa. If we have no knowledge of $P(C)$, then it seems unwarranted and opportunistic to pick values that yield smaller values for either the confounder size, $RP(E:C)$ or the confounder linkage with the predictor, $RP(C:A)$.

To avoid opportunism and to simplify things, assume that $P(C) = P(A)$. This restricts confounders to those in the same prevalence class as the exposure, just as the Attributable Fraction of Cases in the Population (AFP) measures the correlation between exposure and cases – relative to the maximum possible for exposures in the same prevalence class: e.g., $P(C) = P(A)$. See Schield and Burnham (2002).

Since the confounder is hypothetical, there is no claim that this assumption or stipulation is realistic. Only that it is one way of achieving the stated goal of specifying all the properties of a confounder given the observed data and a single value.

4.6 NULLIFICATION WHEN $P(C) = P(A)$

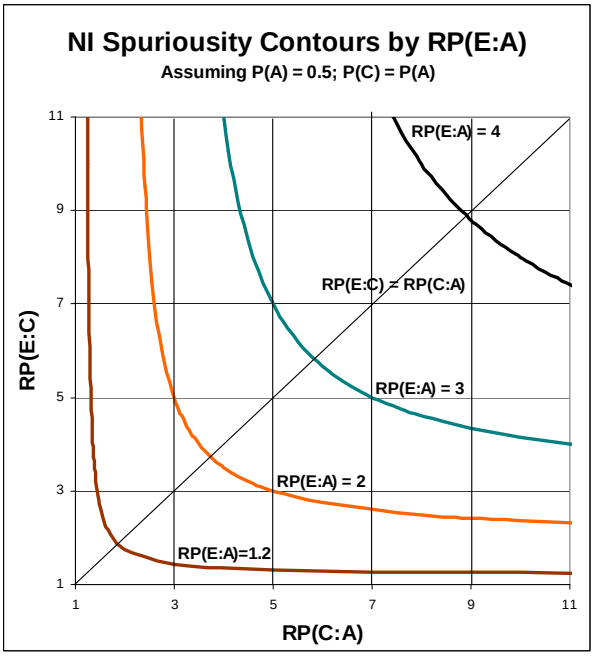
When $P(C) = P(A)$, the nullification condition is:

Eq. 70 $$XRP(E:C) = \frac{XRP(E:A)\{[1/P(A)]+XRP(C:A)\}}{XRP(C:A)-XRP(E:A)}$$

When $P(A) = 0.5$, we obtain the contours of equal strength shown in *Figure 13*. Although there are a wide

range of combinations for *RP*(*E*:*C*) and *RP*(*C*:*A*) it can be shown that there is a symmetry around the line, *RP*(*E*:*C*) = *RP*(*C*:*A*).[41] When a function, y = f(x), has one point closest to the origin, that point is given by dy/dx = -x/y. Since these contours are symmetric about the diagonal, they are closest when their slope is -1, so that the closest point is x = y or *RP*(*E*:*C*) = *RP*(*C*:*A*).[42]

Figure 13: RP(E:C) vs RP(C:A) Spuriosity Contours



When P(C) = P(A), we can describe a strength contour using a single value, S, where RP(C:A) = RP(E:C) = S. For a given value of RP(E:A), this combination gives the point closest to the origin. This doesn't say that either *RP*(*E*:*C*) or *RP*(*C*:*A*) is smallest at this point. There are combinations where either is smaller, but this is the point at which the sum of their squares is smallest – they are jointly minimal.

All the other combinations can be derived given this one value of S since *P*(*C*) = *P*(*A*). One advantage of using this minimal Cartesian distance point, *XRP*(*C*:*A*) = *XRP*(*E*:*C*), is that it avoids extremes.

a. A very weak confounder *XRP*(*E*:*C*), minimally more than *XRP*(*E*:*A*), can still nullify an association provided *XRP*(*C*:*A*) is very large. Focusing on the relatively small size of *XRP*(*E*:*C*) needed for nullification makes the observed association seem weak.

b. A very strong confounder *XRP*(*E*:*C*) is required to nullify an association provided the excess confounder prevalence, *XRP*(*C*:*A*) is minimally greater than the observed association: *XRP*(*E*:*A*). Focusing on the large size of *XRP*(*E*:*C*) in this pair makes the observed association *XRP*(*E*:*A*) seem very strong.

4.7 S CONFOUNDER NULLIFICATION

An **S confounder** is hereby defined as a binary confounder where *P*(*C*) = *P*(*A*) and where *RP*(*C*:*A*) = *RP*(*E*:*C*) = S. Using Eq. 70, it follows that an S confounder will nullify the association *RP*(*E*:*A*) when

Eq. 71 $$S - 1 = \frac{XRP(E:A)\{1 + [P(A)(S-1)]\}}{P(A)[(S-1) - XRP(E:A)]}$$

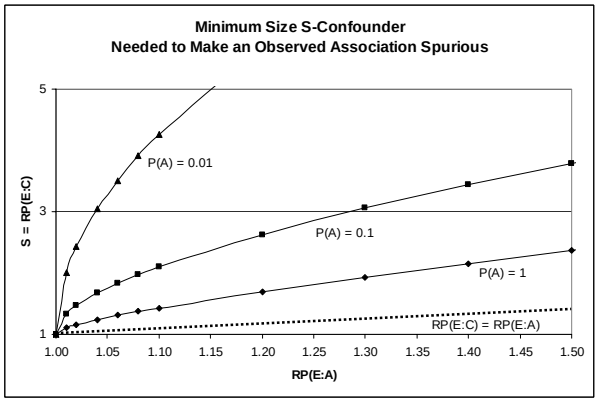
Collecting terms, solving the quadratic and taking the root for $S > RP(E:A)$[43] gives:

Eq. 72 $$S = RP(E:A) + XRP(E:A)\sqrt{1 + 1/[P(A)XRP(E:A)]}$$

This equation identifies the size of an S confounder needed to nullify an observed association having a prevalence, *P*(*A*), and a relative prevalence, *RP*(*E*:*A*).

Figure 14 illustrates the size of an S confounder needed to nullify an observed association – given the prevalence *P*(*A*) and the relative prevalence, *RP*(*E*:*A*).

Figure 14: Minimum S Needed to Nullify Association



Consider those exposed to second hand smoke. If their prevalence is 25% and their relative risk of lung cancer is 1.2, then this association would be made spurious by an S confounder of size 2.1.

Note that the smaller the prevalence of the predictor, *P*(*A*), the larger the confounder size, S, needed to nullify an observed association, *RP*(*E*:*A*). For low-prevalence predictors, very large confounders are required to nullify the observed association.

Disciplines, not statisticians, must decide what size S confounder is considered small – just as with p-values.

4.8 NULLIFICATION BY S CONFOUNDERS

The largest *XRP(E:A)* that is made spurious by an S confounder as a function of *P(A)* is given by:[44]

[41] Let x = XRP(E:C), y = XRP(C:A), z = XRP(E:A) and k = 1/P(A). Eq. 70 yields, x = z(k+y)/(y-z) so z = xy/(k+x+y). The latter shows the symmetry between x and y for a given z.

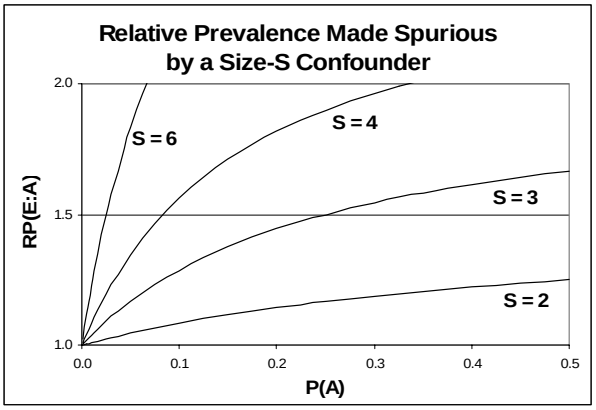
[42] This can be proven. Let $D^2 = x^2 + y^2 = y^2 + [z(k + y)/(y - z)]^2$. To minimize D for a given value of z, let dD/dy = 0. So, y = z + SQRT[z (z + k)] plus a negative and two imaginary roots. Substituting the positive solution into the equation for x gives: x = z + SQRT[z (z + k)]. So D is minimized when x = y, which means when XRP(C:A) = XRP(E:C) or RP(C:A) = RP(E:C)

[43] RP(E:C) > RP(E:A) is a necessary condition for nullification. See Schield and Burnham (2003).

Eq. 73 $$XRP(E:A) = P(A)(S-1)^2 / \{1+[2P(A)(S-1)]\}$$

As shown in Figure 15, relative prevalences under 1.5 are made spurious by S confounders with S < 4 when $P(A) > 0.1$.

Figure 15: Relative Prevalence Made S-Spurious



If S = 5, then $XRP(E:A) = 16\ P(A)/[1+8\ P(A)]$. If P(A) = 0.5, RP(E:A) = 2.6; if P(A) = 0.1, RP(E:A) = 1.9. If epidemiologists were to require relative prevalences to withstand nullification by an S confounder of size 5, many relative prevalences would need to be flagged as being vulnerable to confounding.

5.1 IDEA OF CONFOUNDER INTERVALS

We now set aside the topic of nullification and turn to the question of influence. Suppose that an S confounder was tangled up in the observed association, $RP(E:A)$. What value of $RP(E:A)$ would be expected if that confounder were removed?

To repeat, note that we are not saying anything about nullification – just about influence – so we are not starting from the prior nullification equations. But certainly it seems useful to include the conditions $P(C) = P(A)$ and $RP(C:A) = RP(E:C)$ as determining the lower limit of a confounder interval. An S confounder may decrease without reversal, nullify or reverse an observed association. The latter is Simpson's Paradox. If the confounder interval for an observed relative risk included unity, we would say that for that size confounder the observed association was not 'confounder resistant'; otherwise it is 'confounder resistant.'

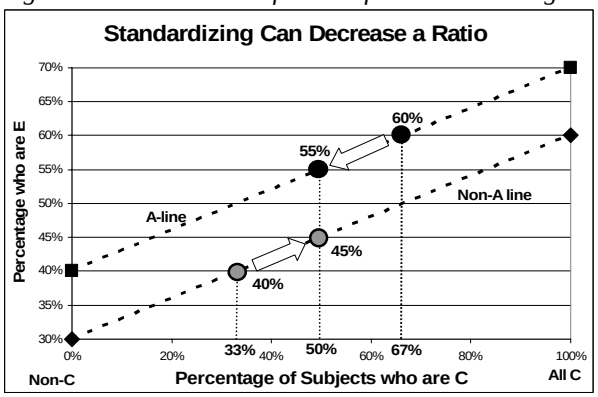
As presented in Schield (2004), standardization involves moving weighted averages along the lines connecting the actual data points: the rates. There is no necessity that these lines be parallel. But when viewing the results of a non-interactive model, there is typically no mention of the actual rates and the associated lines are necessarily parallel.

Standardizing using the four corner values from a non-interactive model (which are co-planar) give the same results as computing the expected values using the model. Using these four data points, the standardizing approach illustrates graphically what the non-interactive model does algebraically.

5.2 LOWER LIMIT

For example, what is the influence of an S confounder of size 2 on an observed association with $P(A)$ = 50% and $RP(E:A)$ = 1.5? To obtain the lower limit, the non-interactive model must fit four requirements: (1) $RP(E:A) = 1.5$, (2) $P(C) = P(A) = 50\%$, (3) $RP(C:A) = 2$, and (4) $RP(E:C) = 2$. Figure 16 illustrates the non-interactive model fitting these values and the standardization from which one can obtain the lower limit.

Figure 16: Lower Limit of Ratio after Standardizing



To see (1) note that $P(E|A)$ = 60% and $P(E|\sim A)$ = 40%, so $RP(E:A)$ = 1.5. To see (2) note that $P(C)$ = $P(A)$ = 50% as specified. To see (3) note that $P(C|A)$ = 66.7% and $P(C|\sim A)$ = 33.3%, so $RP(C:A)$ = 2 (66.7% / 33.3%) as specified. Although this figure does not show $P(E|C)$ or $P(E|\sim C)$ directly, note that if $P(E|C)$ = 2/3 and $P(E|\sim C)$ = 1/3, then $RP(E:C)$ = 2 as specified.

On the right, the weighted average of P(E|C,A) and P(E|C,~A) is P(E|C) = P(A|C) P(E|C,A) + P(~A|C) P(E|C,~A). If P(C) = P(A), then P(A|C) = P(C|A) and P(~A|C) = P(~C|A), so P(E|C) = P(C|A) P(E|C,A) + P(~C|A) P(E|C,~A). If P(E|C,A) = 0.7, P(E|C,~A) = 0.6 and P(C|A) = 2/3, then P(~C|A) = 1/3 and P(E|C) = (2/3)(70%) + (1/3)(60%) = +46.67% + 20% = 2/3. On the left, the weighted average of P(E|~C,A) and P(E|~C,~A) is P(E|~C) = P(A|~C)P(E|~C,A) + P(~A|~C)P(E|~C,~A). Since P(C) = P(A), P(A|~C) = P(C|~A) and P(~A|~C) = P(~C|~A), so P(E|~C) = P(C|~A) P(E|~C,A) + P(~C|~A) P(E|~C,~A). If P(E|~C,A) = 0.4, P(E|~C,~A) = 0.3 and P(C|~A) = 1/3, then P(~C|~A) = 2/3 and P(E|~C) = (1/3)(40%) + (2/3)(30%) = 13.3% + 20% = 1/3.

After taking into account the influence of this confounder, the standardized value of $P(E|A)$ is 55% and the standardized value of $P(E|\sim A)$ is 45%. Thus, the standardized value of the relative prevalence of E for A, the lower limit of this confounder interval, is 1.22.

[44] Since RP is continuous, this also "equals" the minimum RP(E:A) that can withstand being made spurious by a size S confounder.

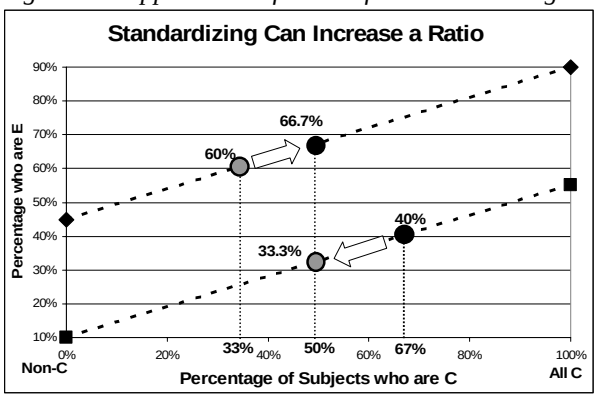
5.3 UPPER LIMIT

Now we turn to obtaining the upper limit of an S confounder interval. Recall that the removal of a confounder can increase the value of an association, *RP*(*E*:*A)* as well as decrease the value. Since we know almost nothing about the confounder it seems inappropriate to assume that it must always decrease the observed association. Suppose that *A* and *C* are chosen so that *RP*(*E*:*A*) and *RP*(*C*:*A)* are both greater than unity. Schield and Burnham (2003) showed that in this case the direct effect is greater than the whole effect only if $0 < RP(C{:}A) < 1$.

What value of *RP*(*C*:*A)* less than 1 can be readily determined given *RP*(*E*:*C)*? An obvious choice is $RP(C{:}A) = 1/RP(E{:}C)$. What is happening is that the confounder groups, *C* and ~*C*, are being exchanged – not in relation to the outcome *E* but in relation to the predictor groups, *A* and ~*A*. In this sense, $RP(C{:}A) = 1/RP(E{:}C)$ used for the upper limit is closely related to $RP(C{:}A) = RP(E{:}C)$ used for the lower limit.

Using a non-interactive model, Figure 17 illustrates the standardization from which one can obtain the upper limit of a size S=2 confounder interval for $P(A) = 50\%$ and $RP(E{:}A) = 1.5$.

Figure 17: Upper Limit of Ratio after Standardizing



Note the four requirements this non-interactive model must fit: (1) $RP(E{:}A) = 1.5$, (2) $P(C) = P(A) = 50\%$, (3) $RP(C{:}A) = 1/2$, and (4) $RP(E{:}C) = 2$. To see (1) note that $P(E|A) = 60\%$ and $P(E|{\sim}A) = 40\%$, so $RP(E{:}A) = 1.5$. To see (2) note that $P(C) = P(A) = 50\%$ as specified. To see (3) note that $P(C|A) = 33.3\%$ and $P(C|{\sim}A) = 66.7\%$, so $RP(C{:}A) = 1/2$. Although $P(E|C)$ and $P(E|{\sim}C)$ are not shown directly, note that if $P(E|C) = 2/3$ and $P(E|{\sim}C) = 1/3$ then $RP(E{:}C) = 2$.

P(E|C) is a weighted average on the right. Since P(C)=P(A), P(E|C) = P(C|A) P(E|C,A) + P(~C|A) P(E|C,~A). 2/3 = (1/3)(90%) + (2/3)(55%) = 30% + 36.67%. P(E|~C) is a weighted average on the left. Since P(C) = P(A), P(E|~C) = P(C|~A) P(E|~C,A) + P(~C|~A) P(E|~C,~A). 1/3 = (2/3)(45%) + (1/3)(10%) = 30% + 3.33%.

After taking into account the influence of this confounder, the standardized value of $P(E|A)$ is 2/3 and the standardized value of $P(E|{\sim}A)$ is 1/3. Thus the standardized value of the relative prevalence of *E* for *A*, the upper limit of this confounder interval, is 2.

So having obtained both the lower and upper limits of this S confounder interval, we can state the size of this particular confounder interval as follows. Given an observed relative prevalence of 1.5 and a predictor prevalence of 50%, the confounder interval due to a size 2 confounder is given by [1.22, 2.0].[45,46]

To summarize, for the S confounder interval proposed herein, both lower and upper limits presume that $P(C) = P(A)$. The lower limit is that determined by $RP(C{:}A) = RP(E{:}C) = S$ while the upper limit is that determined by $RP(C{:}A) = 1/RP(E{:}C) = 1/S$.

5.4 CONFOUNDER INTERVAL FORMULAS

Appendix G derives the standardized values for $P(E|A)$ and $P(E|{\sim}A)$ when $P(C|A) = P(C|{\sim}A) = P(C)$ in terms of the slope b1 in the non-interactive model. Various combinations of these standardized values are also obtained. The spuriosity conditions obtained earlier can be obtained from these formulas.

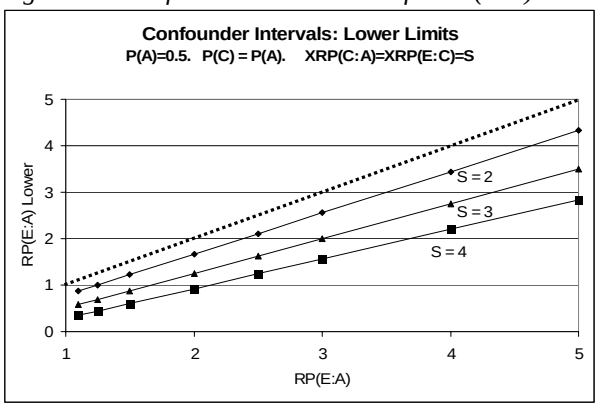
It may be useful to see these standardized values in terms of the conditions specifying the predictor and the confounder – without including b1. The limits of S confounder intervals when $P(C) = P(A)$ are derived for $P(E)/P(E|{\sim}A)$, in Appendix H and for the relative prevalence $RP(E{:}A)$ in Appendix I. In both cases, the formulas seem to conceal more than they reveal. Hopefully they contain analytical relationships that enable a better understanding of the underlying dynamics.

Figure 18 illustrates the lower limit of relative prevalence confounder intervals involving S confounders of size 2, 3 and 4. As a function of *RP*(*E*:*A*), these lower limits are nearly linear.

[45] The first of these confounder intervals was obtained on 12/23/2003 using Excel with co-planar rates for RP(E:A) = 1.25, P(A) = 0.5.

[46] We avoid using 'model' in talking about an S confounder to emphasize that standardized values are based on a model – not the specifications of the S confounder. We avoid using 'adjusted' to emphasize that the data itself is not being adjusted.

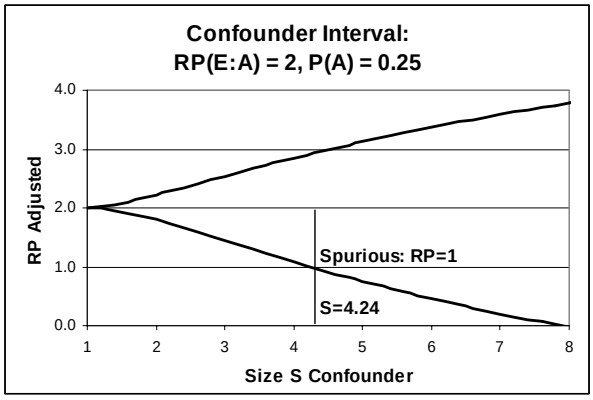
Figure 18: Confounder Lower Limits for RP(E:A)



Although the language of direct, whole and indirect effects is properly used only for differences, the terms can be appropriated for ratios provided the equation relating these items is set aside. Consider a whole effect given in terms of a ratio, *RP*(*E:A*), the observed relative prevalence association between *A* and *E*. Eq G6b in Appendix G presents just the direct effect in terms of a ratio given the whole effect and an S confounder.

Figure 19 shows the upper and lower limits of a confounder interval as a function of confounder size.

Figure 19: Confounder Interval for Size S



Some combinations of values may drive these equations into regions involving unacceptable values where the upper limit goes infinite or drops below 1, or where the lower limit goes below zero.

5.5 DISCUSSION

These relationships may be useful in educating data analysts and journalists on the possible influence of unobserved confounders even though these relationships just restate the size of the original association, RP(E:A), in different terms (since there is not yet any objective basis for selecting confounder sizes).

Furthermore, these relationships may be useful in setting rules or standards for publication by journal editors. They may even be useful in modeling to set a minimum criterion for including a predictor so as to avoid including predictors that may be statistically significant, but are too weak to withstand nullification by a confounder of a given size.

A deeper question involves the ability of relative prevalence to measure the causal status of the predictor. More work may be needed on this foundational issue.

5.6 RECOMMENDATIONS

The following are recommendations for handling count-based associations obtained from observational studies.

1. Those presenting relative risks or prevalences should indicate the minimum size S confounder that would nullify the observed association.
2. Those using relative risks or prevalences to make decisions for publication or for action should set minimum standards of the confounder size which an acceptable association must resist without being nullified or reversed – or what size confounder one should use in giving confounder intervals.
3. More analysis is needed on the use of a double ratio such as relative risk to measure the strength of evidence on the causal status of the predictor.

6 CONCLUSION

This model of confounding is primitive and the choice of a value to use in discriminating between spurious and non-spurious associations is somewhat ad hoc. On an absolute scale, this model is far from the stature of the binomial model of chance. But on a relative scale, given that there are only a few rules of thumb, this proposed measure may be of some value for those who are trying to set aside those statistical associations that provide the weakest support for causal connections.

REFERENCES

Abramson, J.H. (1994). *Making Sense of Data*. Oxford University Press, Second Edition.

Abramson J.H. and Gahlinger P.M. (2001): *Computer Programs for Epidemiologists: PEPI v. 4.0*. Sagebrush Press. Salt Lake City, Utah.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* 11, 1269-1275.

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.* J. of National Cancer Institute, 22, pp. 173-203.

Fisher, Ronald (July, 1958). Letter to the Editor, *Lung Cancer and Cigarettes.* Nature, 182. p. 108.[47]

[47] www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafreader.htm

Gastwirth, Joseph L. (1988). *Statistical Reasoning in Law and Public Policy.* pp. 296-297. Academic Press

Hill, Austin Bradford. *The Environment and Disease: Association or Causation*? *Evolution of Epidemiologic Ideas: Annotated Readings on Concepts in Methods.* Sander Greenland Ed. Epidemiology Resources Inc., Massachusetts 1987, pp 7-12. This article is on-line at: www.med.utah.edu/dfpm/SirAustinBradfordHill.htm

Last, John M. (1985). *A Dictionary of Epidemiology.* Oxford University Press.

Lazarsfeld, Paul F. "*Algebra of Dichotomous Systems"* in *Studies in Item Analysis and Prediction* edited by Herbert Solomon (1961), p. 121, Stanford Univ. Press.

Mokdad, Ali H, James S. Marks, Donna F. Stroup and Julie L. Gerberding (2004). *Actual Causes of Death in the United States, 2000.* Journal of the American Medical Association, Vol. 291, N. 10, p. 1238-1245.

Pearl, Judea (1998). *Why there is no statistical test for confounding, why many think there is, and why they are almost right.* UCLA Computer Science Department, Technical Report (R-256), July 1998.[48]

Rosenbaum, Paul R. (2002). *Observational Studies.* Springer-Verlag. Second Edition, P. 106.

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

Schield, Milo (1999). *Simpson's Paradox and Cornfield's Conditions.* 1999 ASA Proceedings of the Section on Statistical Education, pp. 106-111.[49]

Schield, Milo (2004). *Three Graphs for Statistical Literacy.* International Conference on Mathematics Education (ICME-10) in Copenhagen.[50]

Schield, Milo and Thomas V.V. Burnham (2002). *Algebraic Relations between Relative Risk, Phi and Measures of Necessity and Sufficiency in 2x2 Tables.* 2002 ASA Proceedings of the Section on Statistical Education. [CD-ROM], 3089-3094.[51]

Schield, Milo and Thomas Burnham (2003). *Confounder-induced Spuriosity and Reversal: Algebraic Conditions for Binary Data Using a Non-Interactive Model.* ASA Proceedings of the Section on Statistical Education. [CD-ROM], 3690-3697.[52]

Schield, Milo and Thomas Burnham (2004). *Confounder Resistance and Confounder Intervals for a Binary Confounder.* ASA Proceedings of the Section on Statistical Education. [CD-ROM], 2781-2788.[53]

Shoukri, M. M. (2000). *Agreement, Measures of.* Encyclopedia of Epidemiological Methods, p. 38-49.

Tabues, Gary (1995). *Epidemiology Faces Its Limits.* Science Magazine 269 (July 14): 164–69.[54]

Wonnacott, Thomas A. and Ronald Wonnacott (1979). *Econometrics (Probability and Mathematical Statistics).* John Wiley & Sons.

Wonnacott, Thomas A. and Ronald Wonnacott (1990). *Introductory Statistics, 5th ed.* John Wiley & Sons.

Acknowledgments: This research was supported by a grant from the W. M. Keck Foundation to Augsburg College "to support the development of statistical literacy as an interdisciplinary curriculum in the liberal arts." Thanks to Dr. Shoukri for equation A8 (which we describe as the *Shoukri equation*) since this equation gave us our starting point. Thanks to Romney Schield (Olomouc University, Czech Republic) for his insight in validating this equation.[55] Dr. Schield is the project director of the W. M. Keck Statistical Literacy Project. He can be contacted at schield@augsburg.edu. This paper is posted at www.StatLit.org.

[48] On web at http://bayes.cs.ucla.edu/frl_papers.html.

[49] At www.StatLit.org/pdf/1999SchieldASA.pdf.

[50] At www.StatLit.org/pdf/2004SchieldICME.pdf.

[51] At www.StatLit.org/pdf/2002SchieldBurnhamASA.pdf.

[52] At www.StatLit.org/pdf/2003SchieldBurnhamASA.pdf.

[53] At www.StatLit.org/pdf/2004SchieldBurnhamASA.pdf.

[54] On web at www.agcom.purdue.edu/AgCom/homepages/tally/Science%20in%20Society%20web/96Taubesarticle.html

[55] Romney Schield validated equation A9 and created A2a and A2b. A11 was derived in Schield (1999). A6 is Bayes rule. Using Derive, Thomas Burnham created and validated all the other equations excluding those that are definitions (e.g., a-j in Appendix A), those that are continuous (e.g., Eq. 1-3, Eq. 21-27, Eq. 31-32 and Eq. 34-38) and those inequalities lacking an equal sign. Validation begins after converting higher-level constructs (such as AFP, XRP) to lower-level constructs (such as P) and then converting the lower level constructs to rates (Ra-Rd) and counts (xa-xd) for submission to Derive. Validated equations involve one of four conditions: Normal (no additional constraint), NI spuriosity (b1=0), general standardization [P(C|A) = P(C|~A)] or P(C) standardization [P(C|A) = P(C|~A) = P(C)].

Appendix A. 2×2 COUNT TABLE IDENTITIES[56]

Counts	~E:Non-Case	E:Case	TOTAL
~A: Control	*a*	*b*	*g=a+b*
A: Exposure	*c*	*d*	*h=c+d*
TOTAL	*e=a+c*	*f=b+d*	*n=e+f=g+h*

Definitions and Basic Relationships:[57]

a. Phi: $\phi = r \equiv (a\,d - b\,c) / w$ *where* $w^2 = e\,f\,g\,h$[58]
$\phi^2(d,f,h,n) = (d\,n - f\,h)^2/[f\,(n-f)\,h\,(n-h)]$
$Phi(E,A) = Phi(A,E) = \phi$

b. Margin Fractions:
Of Cases: $P(E) \equiv f/n$
Of Exposures: $P(A) \equiv h/n$

c. Body Fractions:
Of Cases: $P(E|A) \equiv d/h$ $P(E|\sim A) \equiv b/g$
Of Exposures: $P(A|E) \equiv d/f$ $P(A|\sim E) \equiv c/e$

d. Prevalence:
Case: $P(E) = P(E|A)\,P(A) + P(E|\sim A)[1-P(A)]$
Exposure: $P(A) = P(A|E)\,P(E) + P(A|\sim E)[1-P(E)]$

e. Differential Prevalence
Of Cases: $DP(E:A) \equiv P(E|A)-P(E|\sim A)$;
Of Exposures: $DP(A:E) \equiv P(A|E)-P(A|\sim E)$

f. Relative Prevalence:
Of Case $RP(E:A) \equiv P(E|A) / P(E|\sim A)$
Of Exposure $RP(A:E) \equiv P(A|E) / P(A|\sim E)$

g. Excess Relative Prevalence
Of Cases $XRP(E:A) \equiv RP(E:A) - 1$
Of Exposures $XRP(A:E) \equiv RP(A:E) - 1$

h. Relative Lift
Of Case: $RL(E,A) \equiv P(E|A) / P(E)$
Of Exposure: $RL(A,E) \equiv P(A|E) / P(A)$

i. Attributable Fraction in Group (*AFG*)[59]
Of Cases in Exposure Attributable to Exposure:
$AFG(E:A) \equiv [P(E|A) - P(E|\sim A)] / P(E|A)$
Of Exposures in Case Attributable to Case:
$AFG(A:E) \equiv DP(A:E) / P(A|E)$

j. Attributable Fraction in Population (*AFP*):
Of Cases in Population Attributable to Exposure:
$AFP(E:A) \equiv [P(E) - P(E|\sim A)] / P(E)$
Of Exposures in Population Attributable to Case:
$AFP(A:E) \equiv [P(A) - P(A|\sim E)] / P(A)$

For more relationships (e.g., Odds Ratio), see Schield and Burnham (2002).

Identities Using Existing Factors[56]

A1a $RP(E:A) = \dfrac{P(A|E)[1-P(A)]}{[1-P(A|E)]P(A)} = \dfrac{P(E|A)[1-P(A)]}{P(E) - P(E|A)P(A)}$

A1b $\dfrac{RP(E:A)}{RP(A:E)} = \dfrac{[1-P(E|A)][1-P(A)]}{[1-P(A|E)][1-P(E)]}$

A2a $XRP(E:A) = \dfrac{[1-P(E)]}{[1-P(A|E)]} \dfrac{XRP(A:E)}{[P(E)\,XRP(A:E)+1]}$

A2b $XRP(E:A)/[1-P(E|A)] = XRP(A:E)/[1-P(A|E)]$

A3a $XRP(E:A) \equiv RP(E:A) - 1 = DP(E:A)/P(E|\sim A)$

A3b $AFG(E:A) = DP(E:A)/P(E|A) = 1-P(E|\sim A)/P(E|A)$

A3c $AFG(E:A) = XRP(E:A)\,P(E|\sim A) / P(E|A)$

A4 $[XRP(E:A)/RP(E:A)] / [XRP(A:E)/RP(A:E)]$
$= [1-P(E)]/[1-P(A)]$

A5a $AFP(E:A)=P(A)\,AFG(E:A) / \{1-[1-P(A)]AFG(E:A)\}$
$= P(A)\,XRP(E:A) / [P(A)\,XRP(E:A)+1]$

A5b $AFG(E:A)=AFP(E:A)/\{P(A)+ [1-P(A)]AFP(E:A)\}$
$= P(A)\,XRP(E:A) / \{[P(A)\,XRP(E:A)+1]P(A|E)\}$

A6 $P(E|A)/P(E) = P(A|E)/P(A) = RL(E,A) = RL(A,E)$
Bayes Rule Comparison

A7 $AFG(E:A)/AFG(A:E) = [1-P(E)]/[1-P(A)]$

Identities Involving Phi. Some are over-specified.

A8 $a = (e\,g + \varphi\,w)/n$ $b = (f\,g - \varphi\,w)/n$
$c = (e\,h - \varphi\,w)/n$ $d = (f\,h + \varphi\,w)/n$

A9 $\phi^2 = \dfrac{[P(E|A)-P(E)][P(E)-P(E|\sim A)]}{P(E)[1-P(E)]}$[60]

A10 $\phi = DP(E:A)\sqrt{\dfrac{P(A)P(\sim A)}{P(E)P(\sim E)}}$ Proportions test[61]

A11

$$\phi^2 = \frac{P(A)[DP(E:A)]^2[1-P(A)]}{[1-P(E|\sim A)-P(A)DP(E:A)][P(A)DP(E:A)+P(E|\sim A)]}$$

A12 $\phi^2 = \dfrac{[P(A)XRP(E:A)+1]-[RP(E:A)/RP(A:E)]}{[P(A)XRP(E:A)+1]^2/[P(A)XRP(E:A)]}$

A13 $\phi = \sqrt{\dfrac{P(E)[1-P(A)]}{P(A)[1-P(E)]}}\left[\dfrac{P(A)XRP(E:A)}{P(A)XRP(E:A)+1}\right]$

A14a $\phi^2 = \dfrac{P(E)}{P(A)}\dfrac{[1-P(A)]}{[1-P(E)]}AFP(E:A)^2$

A14b~ $\phi^2 = P(E|A)AFG(E:A)P(A|E)AFG(A:E)$

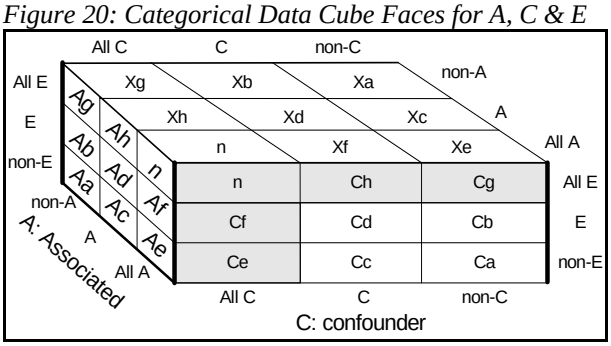
A15 $\phi^2 = AFP(E:A)\,AFP(A:E) = DP(E:A)\,DP(A:E)$

[56] Over-specified equations allow inconsistent inputs.
[57] Lower case indicates counts; upper case indicates ratios.
[58] $X^2 = \Sigma[(\text{actual value} - \text{expected value})^2 / \text{expected value}] = n\,\phi^2$
[59] The term 'group' includes both exposure and control groups and case and non-case groups.
[60] Note that the right-hand term, [P(E) - P(E|~A)]/P(E), is *AFP(E:A)*.
[61] When squared and multiplied by *n*, this is a test for independence.

Appendix B: DISCRETE DATA SUMMARY

The data values involving the three binary variables E, A and C are simply counts within the various categories indexed by the associated index values: *A, ~A, C, ~C, E and ~E*. The data is completely specified by the 8 body counts or by various ratios from which one can obtain these 8 body counts. Figure 20 shows the faces of the categorical cube for binary variables.

Figure 20: Categorical Data Cube Faces for A, C & E



A useful set of ratios involves four rates and four weights. The four rates are

B1. $Ra = P(E|\sim A,\sim C) = n(E,\sim A,\sim C) / n(\sim A,\sim C)$
$Rb = P(E|\sim A,C) = n(E,\sim A,C) / n(\sim A,C)$
$Rc = P(E|A,\sim C) = n(E,A,\sim C) / n(A,\sim C)$
$Rd = P(E|A,C) = n(E,A,C) / n(A,C)$.

Four associated weights are given by

B2 $Xa = n\,P(\sim A,\sim C)$, $Xb = n\,P(\sim A,C)$,
$Xc = n\,P(A,\sim C)$, $Xd = n\,P(A,C)$.[62]

As expected, the sum of these four weights is n.

From the four rates and weights, one can obtain the counts in any of the eight body cells.

B3. $n(E,\sim A,\sim C) = Ra\,Xa$, $n(\sim E,\sim A,\sim C) = (1-Ra)Xa$,
$n(E,A,\sim C) = Rb\,Xb$, $n(\sim E,A,\sim C) = (1-Rb)Xb$,
$n(E,\sim A,C) = Rc\,Xc$, $n(\sim E,\sim A,C) = (1-Rc)Xc$,
$n(E,A,C) = Rd\,Xd$, $n(\sim E,A,C) = (1-Rd)Xd$.

As expected, the sum of the eight body cell counts is n.

From the weights in B2, one gets these prevalences.

B4. $P(C|A) = Xd/(Xc+Xd)$, $P(C|\sim A) = Xb/(Xa+Xb)$.
$P(A|C) = Xd/(Xb+Xd)$, $P(A|\sim C) = Xc/(Xc+Xa)$.
$P(C) = (Xb+Xd)/n$, $P(A) = (Xc+Xd)/n$.

From these four rates and weights, one can obtain various weighted averages:

B5. $P(E|A) = (Rd\,Xd + Rc\,Xc) / (Xd+Xc)$,
$P(E|\sim A) = (Rb\,Xb + Ra\,Xa) / (Xb+Xa)$,
$P(E|C) = (Rd\,Xd + Rb\,Xb) / (Xd+Xb)$,
$P(E|\sim C) = (Rc\,Xc + Ra\,Xa) / (Xc+Xa)$.

A more interesting set of coordinates involves the total count and a series of orthogonal ratios involving perpendicular binary cuts in the 2x2x2 cube. The first cut separates A and ~A with P(A) as the ratio of interest. The second cut separates C from ~C with P(C|A) and P(C|~A) as the two ratios of interest. The third cut separates E from ~E with the four ratios (Ra, Rb, Rc and Rd) as shown above. The total count and these seven ratios completely specify the eight counts.

B6. $P(E|A) = Rd\,P(C|A) + Rc[1-P(C|A)]$
$P(E|\sim A) = Rb\,P(C|\sim A) + Ra[1-P(C|\sim A)]$
$P(E) = P(E|A)\,P(A) + P(E|\sim A)[1-P(A)]$

Given these ratios as summaries of the underlying data, general identities such as these can be derived:

B7. $DP(E:A) = DP(C:A)\ DP(E:C) + [P(E|A) - P(E|C)\ P(C|A) - P(E|\sim C)\ P(\sim C|A)]/P(\sim A)$

B8. $DP(E:A) = DP(C:A)\ DP(E:C) + \{[P(C|A)(Rd-Rb)P(C|\sim A) / P(C)]\} + \{[1-P(C|A)](Rc-Ra)[1-P(C|\sim A)] / [1-P(C)]\}$

B9. $DP(E:A) = \{(Rd-Rb)P(C) + (Rc-Ra)[1-P(C)]\} + DP(C:A)\ \{(Rd-Rc)[1-P(A)]+(Rb-Ra)P(A)\}$

Note the association between B8 and the Lazarsfeld accounting formula. See Lazarsfeld (1961).

Table X: Cross-prevalence between *A* and C

Table X	*Non-C*	*C*	TOTAL
Non-A	*Xa*	*Xb*	*Xa+Xb*
A	*Xc*	*Xd*	*Xc+Xd*
TOTAL	*Xa+Xc*	*Xb+Xd*	*n*

Table R: Rate of *E* classified by *A* and C.

Table R	*Non-C*	*C*	TOTAL
Non-A	*Ra*	*Rb*	*P(E|~A)*
A	*Rc*	*Rd*	*P(E|A)*
TOTAL	*P(E|~C)*	*P(E|C)*	*P(E)*

Table E: Distribution of *E* by *A* and C.

Table E	*Non-C*	*C*	TOTAL
Non-A	*Ea=Ra·Xa*	*Eb=Rb·Xb*	*Ea+Eb*
A	*Ec=Rc·Xc*	*Ed=Rd·Xd*	*Ec+Ed*
TOTAL	*Ea+Ec*	*Eb+Ed*	*Ea+Eb+Ec+Ed*

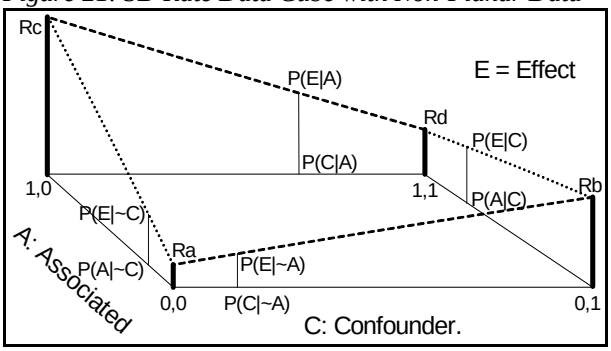
Note that this paper does not include a comprehensive analysis of, or treatment for, $P(E|\sim A) = 0$ or $P(E|A) - P(E|\sim A) = 1$.

Appendix C: RATE DATA CUBE

To model this data, the values of variables *A, C* and *E* are treated as continuous. Their extreme values (*A* and *~A*) are 0 and 1. See Figure 21. In the A-C plane, location 0, 0 is *~A, ~C*. Instead of having a pair of data points (at $E = 0$ and $E = 1$) for each of the four corners, each pair is replaced by its rate as defined in the previous appendix. E.g., $Rd = P(E|A,C)$

[62] $P(A) = [P(E)-P(E|\sim A)] / [P(E|A)-P(E|\sim A)]$
$= [P(C)-P(C|\sim A)] / [P(C|A)-P(C|\sim A)]$
$P(C) = [P(E)-P(E|\sim C)] / [P(E|C)-P(E|\sim C)]$
$= [P(A)-P(A|\sim C)] / [P(A|C)-P(A|\sim C)]$

Figure 21: 3D Rate Data Cube with Non-Planar Data



Noteworthy values of *C* are *0*, P(C|~A), P(C), P(C|A) and 1. As shown in Figure 21, *P(E|A)* is a weighted average of *Rc* and *Rd*: $P(E|A) = Rc[1-P(C|A)] + Rd\ P(C|A) = Rc + (Rd-Rc)P(C|A)$.

Appendix D: NON-INTERACTIVE MODEL

A linear non-interactive regression model involving two predictors is:

D1. $E(A,C) = b_0 + b_1A + b_2C$.

Coefficients are obtained by minimizing OLS variance. These coefficients can have many forms.

(1) One form involves rates and weights. Let b_3 indicate non-planarity where $b_3 = Rd-Rc-Rb+Ra$.
Let $D1 = Xa[Xb(Xc+Xd)+(Xc\ Xd)]+(Xb\ Xc\ Xd)$,

D2a. $b_0 = Ra - (b_3\ Xb\ Xc\ Xd)/D1$,
D2b. $b_1 = (Rc-Ra) + [b_3\ Xb(Xa+Xc)Xd]/D1$,
D2c. $b_2 = (Rb-Ra)+[b_3\ Xc(Xa+Xb)Xd]/D1$.

Under cross-A rate equality, $Ra = Rc$ and $Rd = Rb$. So, $b_3 = 0$, $b_1 = 0$, and we have NI spuriosity.

If the data is planar, b_3 is zero, so $(Rd-Rb) = (Rc-Ra)$. So, planar data entails cross-A rate difference equality (which is different from cross-A rate equality). It also entails cross-C difference rate equality: $Rd-Rc = Rb-Ra$,

D2d. $b_0=Ra$, $b_1=Rc-Ra$, $b_2=Rb-Ra$ for planar data.

For planar data, the corners of the surface are the rates:

D2e. $E(0,0)=Ra$, $E(0,1)=Rb$, $E(1,0)=Rc$, $E(1,1)=Rd$.

(2) Another form of the coefficients involves the ratios derived from the face values in Figure 20.

D3. $D2 = 1 - \{[P(A|C)-P(A|\sim C)][P(C|A)-P(C|\sim A)]\}$
$D2 = 1 - [DP(A:C)\ DP(C:A)]$

D3a. $b_0 = P(E) - \{[P(E|A)-P(E|\sim A)]P(A|\sim C) + [P(E|C)-P(E|\sim C)]P(C|\sim A)\}/D2$
$b_0 = P(E) - [DP(E:A)\ P(A|\sim C) + DP(E:C)\ P(C|\sim A)]/D2$

D3b. $b_1 = \{[P(E|A) - P(E|\sim A)] - [P(C|A) - P(C|\sim A)][P(E|C) - P(E|\sim C)]\}/D2$
$b_1 = \{DP(E:A) - [DP(C:A)\ DP(E:C)]\}/D2$

D3c. $b_2 = \{[P(E|C) - P(E|\sim C)] - [P(E|A) - P(E|\sim A)][P(A|C) - P(A|\sim C)]\}/D2$
$b_2 = \{DP(E:C) - [DP(E:A)\ DP(A:C)]\}/D2$[63,64]

The following can be derived from these equations:

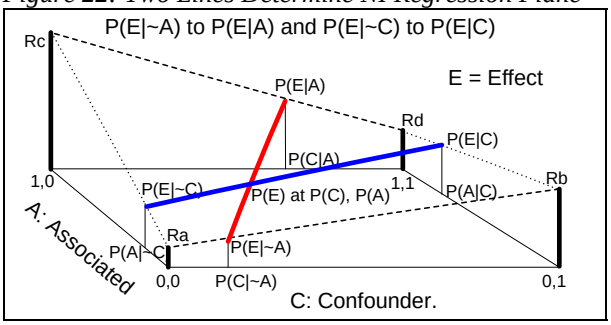
D4a. $P(E|A) = E[A=1, C=P(C|A)]$
$P(E|\sim A) = E[A=0, C=P(C|\sim A)]$

D4b. $P(E|C) = E[A=P(A|C), C=1]$
$P(E|\sim C) = E[A=P(A|\sim C), C=0]$

D4c. $P(E) = E[A=P(A), C = P(C)]$

Thus the regression plane contains the lines connecting *P(E|A)* with *P(E|~A)* and *P(E|C)* with *P(E|~C)* *as shown in* Figure 22. The point at which these two lines intersect has $P(E|A,C) = P(E)$. Not all ratios in categorical space lie on the surface of a given model: $Rd = P(E|A,C) \neq E(A=1,C=1)$. All ratios in categorical space will lie on this regression plane under restrictive conditions such as when the confounder has no association with the predictor, RP(A|C) = 1 and RP(C|A) = 1.[65]

Figure 22: Two Lines Determine NI Regression Plane



The four corners of the planar surface are:

D5a. $E(0,0) = P(E) - \{P(A|\sim C)[P(E|A)-P(E|\sim A)] + P(C|\sim A)[P(E|C)-P(E|\sim C)]\}/D2$
$E(0,0) = P(E) - \{P(A|\sim C)\ DP(E:A) + P(C|\sim A)\ DP(E:C)\}/D2$

D5b. $E(0,1) = P(E) - \{P(A|C)[P(E|A)-P(E|\sim A)] - [1-P(C|\sim A)][(P(E|C)-P(E|\sim C))]\}/D2$
$E(0,1) = P(E) - \{P(A|C)\ DP(E:A) - [1-P(C|\sim A)][DP(E:C)]\}/D2$

D5c. $E(1,0) = P(E) + \{[1-P(A|\sim C)][P(E|A)-P(E|\sim A)] - P(C|A)[P(E|C)-P(E|\sim C)]\}/D2$,
$E(1,0) = P(E) + \{[1-P(A|\sim C)]DP(E:A) - P(C|A)\ DP(E:C)\}/D2$

D5d. $E(1,1) = P(E) + \{[1-P(A|C)][P(E|A)-P(E|\sim A)] + [1-P(C|A)][P(E|C)-P(E|\sim C)]\}/D2$

---

[63] $b_2(E|A,C)$ is obtained from $b_1(E|A,C)$ by exchanging *A* with *C*, *P(A)* with *P(C)*, *P(E|A)* with *P(E|C)* and *P(C|A)* with *P(A|C)*. See D3b & D3c.

[64] If $b_1(E|A,C) = 0$, then DP(E:A)=DP(E:C) DP(C:A), so from D3c, $b_2(E|A,C)$ = [DP(E:C)-DP(E:C) DP(C:A) $DP(A:C)]/D2$ = DP(E:C)[1-DP(C:A) DP(A:C)]/*D2* = DP(E:C) = $b_2(E|C)$.

[65] If $RP(C:A)=1$ and $RP(A:C)=1$, then $b_0=P(E|\sim A)+P(E|\sim C)-P(E)$, $b_1=DP(E:A)$ and $b_2=DP(E:C)$ *so* $E(0,0)=P(E|\sim A)+P(E|\sim C)-P(E)$, $E(0,1)=P(E|C)+P(E|\sim A)-P(E)$, $E(1,0)=P(E|A)+P(E|\sim C)-P(E)$ and $E(1,1)=P(E|A)+P(E|C)-P(E)$. If $b_3 = 0$ then $Ra = E(0,0)$, $Rb = E(0,1)$, $Rc = E(1,0)$ and $Rd = E(1,1)$. If $P(E|A)=P(E|\sim A)$ the association is trivial and reversal is meaningless.

$$E(1,1) = P(E) + \{[1\text{-}P(A|C)]DP(E:A) + [1\text{-}P(C|A)]DP(E:C)\}/D2$$

D6a $P(E|C) = E(0,1) + P(A|C)[E(1,1) - E(0,1)]$.
D6b. $P(E|\sim C) = E(0,0) + P(A|\sim C)[E(1,0) - E(0,0)]$.
A partial test of the validity of these formulae is to nullify the association between the predictor and the confounder: RP(C:A) = 1 and RP(A:C) = 1.[66]

Appendix E: FORMS OF SLOPE: $b_1(E|A,C)$

There are many form of the slope, b_1(E|A,C). The first form involves differences of double ratios.

E1a. $$b_1 = \frac{DP(E:A) - [DP(C:A)DP(E:C)]}{1 - [DP(C:A)DP(A:C)]}.$$

E1b Let $K1 = 1/[1\text{-}DP(C:A)DP(A:C)]$.
E1c $b_1 = K1\{DP(E:A)\text{-}[DP(C:A)DP(E:C)]\}$
Note that K1 is always positive.

The second form involves the attributable fraction in the population which is closely related to ϕ.

E2a $$b_1 = \frac{P(E)[AFP(E:A) - AFP(C:A)AFP(E:C)]}{P(A)\{1 - [DP(C:A)DP(A:C)]\}}.$$

E2b Let $K2 = P(E)/\{P(A)[1\text{-}DP(C:A)DP(A:C)]\}$
E2c $b_1(E|A,C) = K2[AFP(E:A)\text{-}AFP(C:A)AFP(E:C)]$

Note that K2 = P(E)K1/P(A) so K2 > 0.

The third and fourth forms involve double ratios.

E3a. Double-ratio form with P(C|~A) in numerator:

$$b_1 = \frac{P(E)\{XRP(E:A)[P(C|\sim A)XRP(E:C)+1] - P(C|\sim A)XRP(C:A)XRP(E:C)\}}{[1 - DP(C:A)DP(A:C)][P(A)XRP(E:A)+1][P(C)XRP(E:C)+1]}$$

E3b. Let $K3 = P(A)\,K2\,/ \{[P(A)XRP(E:A)+1][P(C)XRP(E:C)+1]\}$

E3c. $$b_1(E|A,C) = K3\{XRP(E:A)[P(C|\sim A)\,XRP(E:C)+1] - [P(C|\sim A)\,XRP(C:A)\,XRP(E:C)]\}$$

E4a. Double-ratio form, $P(A)$ and $P(C)$ in numerator:

$$b_1 = \frac{P(E)\{XRP(E:A)[P(A)XRP(C:A) + P(C)XRP(E:C)+1] - P(C)XRP(E:C)XRP(C:A)\}}{[1 - DP(C:A)DP(A:C)][P(A)XRP(C:A)+1][P(A)XRP(E:A)+1][P(C)XRP(E:C)+1]}$$

E4b. Let K4 = P(E)/Denominator

E4c. b_1(E|A,C) = -K4{[P(C) XRP(E:C) XRP(C:A)] - XRP(E:A)[P(C) XRP(E:C)+1+P(A) XRP(C:A)]}

Cases with zero denominators are ignored. Non-zero denominators are always positive when *XRP(C:A)*, *XRP(E:C) and XRP(E:A)* are positive.

Equations E1 through E4 are the basis for F1 through F4. Cases with zero denominators are ignored. Non-zero denominators are always positive when *RP(C:A)*, *RP(E:C)* and *RP(E:A)* are greater than one.

Appendix F: NON-INTERACTIVE SPURIOSITY

I. THREE DOUBLE RATIOS PER EQUATION[67]

F1. $DP(E:A) = DP(C:A)\,DP(E:C)$

F2. $AFP(E:A) = AFP(C:A)\,AFP(E:C)$ [68]

F3a. $$XRP(E:A) = \frac{XRP(C:A)P(C|\sim A)XRP(E:C)}{1 + [P(C|\sim A)XRP(E:C)]}$$

F3b. $$XRP(E:C) = \frac{XRP(E:A)}{P(C|\sim A)[XRP(C:A) - XRP(E:A)]}$$

F3c. $$XRP(C:A) = XRP(E:A)\{1 + 1/[P(C|\sim A)XRP(E:C)]\}$$

F3d. Subtracting F3a from F3b eliminates XRP(C:A):

$$XRP(E:C) - XRP(E:A) = \frac{\{RP(E:C)P(C|\sim A) + [1 - P(C|A)]\}XRP(E:C)}{[P(C|\sim A)XRP(E:C)] + 1}$$

F4a. $$XRP(E:A) = \frac{P(C)XRP(C:A)XRP(E:C)}{P(A)XRP(C:A) + P(C)XRP(E:C) + 1}$$

F4b $$P(C)XRP(E:C) = \frac{XRP(E:A)\{1 + [P(A)XRP(C:A)]\}}{[XRP(C:A) - XRP(E:A)]}$$

F4c. $$XRP(C:A) = \frac{XRP(E:A)\{1 + [P(C)XRP(E:C)]\}}{[P(C)XRP(E:C)] - [P(A)XRP(E:A)]}$$

F4d.
$$\frac{P(A)XRP(E:A)}{P(A)XRP(E:A)+1} = \frac{P(C)XRP(E:C)}{P(C)XRP(E:C)+1}\,\frac{P(A)XRP(C:A)}{P(A)XRP(C:A)+1}$$

II. TWO DOUBLE RATIOS PER EQUATION

F5. $XRP(C:A) = XRP(E:A)P(E|\sim A)/[P(E|\sim A) - P(E|C)]$

F6a. $$RP(E:A) = \frac{[P(C|A)XRP(E:C)] + 1}{[P(C|\sim A)XRP(E:C)] + 1}$$

F6b. $$XRP(E:C) = \frac{XRP(E:A)}{P(C|A) - P(C|\sim A)RP(E:A)}$$

F7. $$XRP(E:C) = \frac{DP(E:A)}{P(C|A)P(E|\sim A) - P(C|\sim A)P(E|A)}$$

F8. $$P(A) = \frac{P(E) - P(E|\sim A)}{DP(C:A)DP(E:C)} = \frac{DP(E:C)[P(C) - P(C|\sim A)]}{DP(E:A)}$$

III. EQUAL SLOPES

F9a. $$\frac{\Delta Y}{\Delta X} = \frac{DP(E:A)}{DP(C:A)} = \frac{DP(E:C)}{(1-0)} = \frac{P(E|A) - P(E)}{P(C|A) - P(C)} = \frac{P(E|C) - P(E)}{1 - P(C)}$$

F9b. $$\frac{\Delta Y}{\Delta X} = \frac{P(E|\sim A) - P(E|\sim C)}{P(C|\sim A)} = \frac{[P(E|A) - P(E|\sim C)]}{P(C|A)}$$

F9c. $$\frac{\Delta Y}{\Delta X} = \frac{P(E|C) - P(E|\sim A)}{1 - P(C|\sim A)} = \frac{P(E|C) - P(E|A)}{1 - P(C|A)}$$

F9d. $$\frac{\Delta Y}{\Delta X} = \frac{P(E) - P(E|\sim A)}{P(C) - P(C|\sim A)} = \frac{P(E) - P(E|\sim C)}{P(C)}$$

IV. OTHER CONDITIONS (NOT SHOWN ABOVE)

F10. $P(E|A) = P(E|\sim C) + P(C|A)DP(E:C)$

F11. $P(E|\sim A) = P(E|\sim C) + P(C|\sim A)DP(E:C)$

F12. $RP(C:A) = [P(E|A) - P(E|\sim C)]/[P(E|\sim A) - P(E|\sim C)]$

---

[66] If *RP(C:A)=1* and *RP(A:C)=1,*
E(0,0) = *P(E) - {P(A)DP(E:A) + P(C)DP(E:C)},*
E(0,1) = *P(E) - {P(A)DP(E:A)-[1-P(C)]DP(E:C)}.*
E(1,0) = *P(E)+{[1-P(A)]DP(E:A)-P(C)DP(E:C)},*
E(1,1) = *P(E)+{[1-P(A)]DP(E:A)+[1-P(C)]DP(E:C)}.*
The form in footnote 65 is equivalent but simpler.

[67] F1, F5, F8 and F12 have two non-A ratios. All others have more.

[68] DP(X:Y)=P(X) XPR(X:Y)/[P(Y) XRP(X:Y)+1]=AFP(X:Y) P(X)/P(Y), where X & Y are any of E, A and C. Application of these identities to F1 produces F2 and F3a.

F13. $P(E)/P(E|\sim A) = \dfrac{[P(E)/P(E|\sim C)][P(C)/P(C|\sim A)]}{[P(E)/P(E|\sim C)]+[P(C)/P(C|\sim A)]-1}$

F14. This equates the whole with the indirect effect.

$$DP(E:A) = P(E|C)DP(C:A) + P(E|A)P(C|\sim A) - P(E|\sim A)P(C|A)$$

F15. $\dfrac{P(E|A)}{P(E)} - 1 = [\dfrac{P(E|C)}{P(E)} - 1][\dfrac{P(C|A)}{P(C)} - 1][\dfrac{P(C)}{1-P(C)}]$

APPENDIX G: EXPECTED RATIOS

In this Appendix, relationships are worked out two ways. First, by treating **the standard value** as a variable, C. There is no rule saying that standardization must be done using the common prevalence, P(C). Second by standardizing to the common prevalence, P(C). This gives both subgroups the same mixture as the combined group so P(C|A) = P(C|~A) = P(C). EXP{Ratio|S} indicates this is the expected value of the ratio when standardized: P(C|A) = P(C|~A) = P(C).

G1. E(A,C) = P(E) + b1[A-P(A)] + b2[C-P(C)][69]

G2a. P(E|A) = E[A=1, C=P(C|A)]
G2b. P(E|A) = P(E) +b1[1-P(A)] + b2[P(C|A)-P(C)]
G2c. EXP{P(E|A)|S} = P(E) + b1[1-P(A)]

G3a. P(E|~A) = E[A=0, C=P(C|~A)]
G3b. P(E|~A) = P(E) + b1[0-P(A)] +b2[P(C|~A)-P(C)]
G3c. EXP{P(E|~A)|S} = P(E) - b1 P(A)

G4a. EXP{P(E)|S} = EXP{P(E|A)|S}P(A)
+ EXP{P(E|~A)|S}[1-P(A)]
G4b. EXP{P(E)|S} = {P(E) +b1 [1-P(A)]} P(A)
+ {P(E) - b1 P(A)} [1-P(A)]
G4c. EXP{P(E)|S} = P(E)

G5a. P(E|A)-P(E|~A) =
{P(E) + b1[1-P(A)] + b2[P(C|A)-P(C)]}
- {P(E) + b1[0-P(A)] + b2[P(C|~A)-P(C)]}
G5b. P(E|A)-P(E|~A) = b1 + b2[P(C|A)-P(C|~A)]
G5c. EXP{[P(E|A) – P(E|~A)] | S} = b1

G6a. EXP{RP(E:A)|S} = EXP{P(E|A)|S}
/ EXP{P(E|~A)|S}

G6b. EXP{RP(E:A)|S}
= {P(E) + b1[1-P(A)]} / {P(E) - b1 P(A)}

G6c. EXP{XRP(E:A) | S} = EXP{[RP(E:A)-1] | S}
= {P(E) + b1[1-P(A)]} / [P(E) - b1 P(A)]-1
G6d. EXP{XRP(E:A)|S} = b1/{P(E) -b1 P(A)}

G7a. EXP{[P(E|A)/P(E)]|S}
= EXP{P(E|A)|S} / EXP{P(E)|S}
= {P(E) + b1[1-P(A)]}/P(E)

G8a. EXP{[P(E) / P(E|~A)]|S}
= EXP{P(E)|S}/ EXP{P(E|~A)|S}
G8b. EXP{[P(E)/P(E|~A)]|S} = P(E)/[P(E) - b1 P(A)]

G9a. EXP{AFP(E:A)|S}
= EXP{[[P(E)-P(E|~A)]/P(E)]|S}
G9b. EXP{AFP(E:A)|S} = b1 P(A)/P(E)

APPENDIX H: EXP{[P(E)/P(E|~A)]|S}

Expanding Eq. G6b using b1 from Eq. E4a gives:

H1a. T0 = 1 + T1(T2 + T3 T4 - T3 - T4) where
H1b. T0 = 1/{P(A) EXP{RP(E:A)|S} + [1-P(A)]}
H1c. T1 = $1/[1-\phi(C,A)^2]$
H1d. T2 = 1/{P(A) RP(E:A) + [1-P(A)]}
H1e. T3 = 1/{P(C) RP(E:C) + [1-P(C)]}
H1f. T4 = 1/{P(A) RP(C:A) + [1-P(A)]}

H1g. $\phi(C,A)^2 = \dfrac{P(A)[1-P(A)]P(C)[XRP(C:A)]^2}{[1-P(C)][P(A)XRP(C:A)+1]^2}$

H2a. P(E)/P(E|~A) = P(A) RP(E:A) + [1-P(A)][70]
H2b. P(E)/P(E|~C) = P(C) RP(E:C) + [1-P(C)]
H2c. P(C)/P(C|~A) = P(A) RP(C:A) + [1 - P(A)]
H2d. EXP{[P(E)/P(E|~A)] | S}
= P(A) EXP{RP(E:A)|S}+[1-P(A)]

H3a. T0 = EXP{P(E|~A)|S}/P(E)
H3b. T2 = P(E|~A)/P(E)
H3c. T3 = P(E|~C)/P(E)
H3d. T4 = P(C|~A)/P(C)

Only T2 depends on P(E|A). Define K1.
H4a. K1 = - [1 + T1(T3 T4 - T3 - T4)][71]
H4b. T0 = (T1 P(E|~A)/P(E)) - K1
H4c. 1/T0 = 1/[(T1 P(E|~A)/P(E)) - K1]
H4d. 1/T0 = [P(E)/P(E|~A)]/{T1– [K1 P(E)/P(E|~A)]}
H4e. EXP{[P(E)/P(E|~A)]|S}
= [P(E)/P(E|~A)]/{T1 – [K1 P(E)/P(E|~A)]}

Special Cases:
If K1 = 0, then
H5a. EXP{[P(E)/P(E|~A)]|S} = (P(E)/P(E|~A)) / T1
H5b. P(A) EXP{RP(E:A)|S}
= [[P(A) RP(E:A)+[1-P(A)]]/T1] - [1-P(A)]
H5c. EXP{RP(E:A)|S} = – [1-P(A)]/P(A)
+ [[RP(E:A)+[1-P(A)]/P(A)]/T1]

If RP(E:A)=1,
H7a. P(A) EXP{RP(E:A)|S} + [1-P(A)] = 1/(T1 - K1)
H7b. EXP{RP(E:A)|S}= {[1/(T1-K1)]-[1-P(A)]}/P(A)

If RP(E:A) = 1, P(A)=1/2 and ϕ(C,A)=0 so T1=1, then
H8a. EXP{RP(E:A)|S} = (1+K1)/(1 - K1)
EXP{RP(E:A)|S} = 0 if K1= -1, 1 if K1=0 and ∞ if K1=1.

If EXP{RP(E:A)|S} = RP(E:A), then C has no influence on the RP(E:A) evaluated at C = P(C). Thus,

---

[69] Eq. D1: E(A,C) = b0 + b1 A + b2 C.
Eq. D4c: P(E) = b0 + b1 P(A) + b2 P(C)

[70] P(A)[P(E|A)/P(E|~A)]+1-P(A) = {P(A)P(E|A) + P(E|~A) [1-P(A)]} / P(E|~A) = P(E)/P(E|~A)

[71] K1 = -{1 + [P(E|C)/P(E) + P(C|A)/P(C) - 1] / [(1-Phi(A,C)^2)(P(E|C)P(C|A))/(P(E) P(C))]}

H9a T1 – [K1 P(E)/P(E|~A)] = 1 From H4e
H9b (T1-1)/K1 = P(E)/P(E|~A)

APPENDIX I: EXP{RP(E:A)|S}

Let S be the size of the confounder where P(C) = P(A).
I1a Let U = 1/(P(A) S + [1-P(A)])
I1b. Let V = 1/(P(A)/S + [1-P(A)])
I1c. ϕ^2(C,A)=[P(A)XRP(C:A)]^2/[P(A)XRP(C:A)+1]^2

Lower Limit: Let RP(E:C) = S and RP(C:A) = S.
I2a. T3 = U = T4.
Note: I2b from Eq. H4a; I2c from H4e.
I2b. K1= - [1 + T1(U^2-U-U)] = - [1 + (T1 U)(U-2)]
I2c. EXP{[P(E)/P(E|~A)]|S} = [P(E)/P(E|~A)]
/ {T1 + [1+ (T1 U)(U-2)][P(E)/P(E|~A)]}
I2d. P(A) EXP{RP(E:A)|S}+1-P(A) =
[P(A) RP(E:A)+1-P(A)] /
{T1+[1+(T1 U)(U-2)][P(A) RP(E:A)+1-P(A)]}

Upper Limit: Let RP(E:C) = S and RP(C:A) = 1/S.
I3a. T3 = U, and T4 = V
I3b. K1 = - [1 + T1(U V – U – V)]
I3c. EXP{[P(E)/P(E|~A)]|S} = [P(E)/P(E|~A)]
/ {T1 + [1+T1(U V - U - V)][P(E)/P(E|~A)]}
I3d. P(A) EXP{RP(E:A)|S} + [1-P(A)]
= {P(A) RP(E:A) + [1-P(A)]} /{T1+
[1+T1(U V-U-V)][P(A) RP(E:A)+1-P(A)]}

Special Cases:
Let P(A) = 0.5:
I4a. ϕ^2(C,A) = [RP(C:A)-1]^2 / [RP(C:A)+1]^2
I4b. 1-ϕ^2(C,A) = 4 RP(C:A)/[RP(C:A)+1]^2
I4c. T1 = 1/[1-ϕ^2(C,A)] = [RP(C:A)+1]^2/[4 RP(C:A)]
Lower Limit:
I5a. T1Low = 1/[1-ϕ^2(C,A)] = [(S+1)^2]/ 4S
I5b. EXP{RP(E:A)|S} + 1 = {(RP(E:A)+1)
/{T1Low + [1+T1Low U(U-2)][(RP(E:A)+1)/2]}}
Upper Limit:
I6a. T1High = 1//[1-ϕ^2(C,A)] = [(1/S+1)^2]/ (4/S)
I6b. EXP{RP(E:A)|S} +1 = (RP(E:A)+1)
/{T1High + [1 + T1High(U V-U-V)][(RP(E:A)+1)/2]}

Let P(A) = 0.5 and S = 2 so U = 2/3 and V = 4/3.
Lower Limit:
I7a. T1Low = [(S+1)^2]/ 4S = 9/8
I7b. EXP{RP(E:A)|S} + 1 = [RP(E:A)+1]
/ {(9/8) + [1+(9/8)(2/3)(-4/3)][(RP(E:A)+1)/2]}
I7c. EXP{RP(E:A)|S} + 1 = [RP(E:A)+1] (8/9)
So when RP(E:A) = 2,
I7d. EXP{RP(E:A)|S} = {3/(9/8)} -1 = (8/3)-1 = 1.67
Upper Limit:
I8a. T1High = [(1/S+1)^2]/ (4/S) = (9/4)/(4/2) = 9/8
I8b. EXP{RP(E:A)|S} +1 = [RP(E:A)+1]
/ {(9/8) + [1 + (9/8)(8/9-2/3-4/3)][(RP(E:A)+1)/2]}
I8c. EXP{RP(E:A)|S} +1 = [RP(E:A)+1]
/ {(9/8) - (1/4)[(RP(E:A)+1)/2]}
So when RP(E:A) = 2,
I8d. EXP{RP(E:A)|S} = {3/[(9/8)-(3/8)]} -1 = 3

For a size S=2 confounder, the confounder interval for RP(E:A) = 2 with P(A) = 0.5 is [1.67, 3.0].