ALGEBRAIC RELATIONSHIPS BETWEEN RELATIVE RISK, PHI AND MEASURES OF NECESSITY AND SUFFICIENCY IN A 2×2 TABLE

Milo Schield, Augsburg College and Thomas V.V. Burnham, Cognitive Consulting.
Dept. of Business Administration. 2211 Riverside Drive. Minneapolis, MN 55454

Abstract: Epidemiologists have long regarded relative risk (*RR*) as a key measure of association between two binary variables. Yet even when the sample is representative of the population, associations having a *RR* > 9 may have a relatively small value for Phi (ϕ < .3). This paper provides additional reasons for using *RR* instead of ϕ. (1) Formulas relating ϕ and *RR* are derived. A relative ϕ is constructed; the absolute value is shown to equal that of the attributable fraction in the population (*AFP*). *RR* is shown to increase monotonically with this relative ϕ for a given exposure factor prevalence (*H*). (2) Constructs are created to measure degrees of necessity (*N*) and sufficiency (*S*). Related formulas are derived. *RR* is shown to increase monotonically with *N* for a given exposure factor prevalence (*H*). (3) Coordinates that will always generate admissible results are discussed. (4) *RR* is a good measure of association between two binary variables because it increases monotonically with *AFP*, with a relative ϕ^2 and with *N* for a given exposure factor prevalence (*H*). In Appendix B, auxiliary identities are presented including a Bayes' rule comparison, the over-involvement rule and the non-response bias effect size.
Keywords: Epidemiology, Constructs

INTRODUCTION

Statistics and epidemiology share a common respect for confounding, a common interest in modeling, and a common basis in empirical data. Yet statistics and epidemiology are tellingly different.

Statistics generally deals with a random sample that is representative of the population; epidemiology often deals with two random samples that are separately representative of the exposure and control groups, but not of the population. Statistics focuses more on experiments and manipulative control. Epidemiology focuses more on observational studies and associations. But epidemiology focuses more on identifying a necessary condition whose removal would reduce undesirable outcomes than on identifying sufficient conditions whose presence would produce undesirable outcomes.

1.1 EPIDEMIOLOGISTS AND CORRELATION

Epidemiologists use *RR* and *OR*, but seldom use any Pearson-based correlations (e.g., ϕ). 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 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.

1.2 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* ≥ 9) as strong evidence of a non-spurious association. See Cornfield (1959).

Suppose that Table 1 is a random sample of deaths.[1] 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	Other Causes	Lung Cancer	Total
Non-smokers	875	25	900
Smokers	75	25	100
Total	950	50	1000

What is the algebraic relation between *RR* and ϕ?

1.3 ALGEBRAIC IDENTITIES

The left column in Appendix A summarizes the algebraic identities between many of the common measures of association between two binary variables.[2] The

[1] This hypothetical data is not totally irrelevant. 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. Source: Statistical Abstract of the United States: 2001, tables 105 and 190.

[2] *Cases* are not individual subjects in a list; they are subjects having the outcome of interest. Subjects are classified in the exposure or

algebraic notation involves F and H. F is the prevalence of cases in the population; H is the prevalence of subjects exposed to the factor of interest.

The relation between RR and Phi (ϕ) is presented in equation G1.

$$\phi^2(RR,F,H) = \left[\frac{F}{H}\frac{(1-H)}{(1-F)}\right]\left[\frac{H(RR-1)}{H(RR-1)+1}\right]^2 \qquad 1$$

In Table 1, $\phi^2 = (9/19)(0.8/1.8)^2 = 0.094$; $\phi = 0.306$. 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 ideal standard.

1.4 RELATIVE PHI

What is the maximum value of ϕ for an exposure factor having a certain prevalence (H)? Equation 1 shown previously specifies the relationship between the binary correlation coefficient (ϕ), the relative risk (RR), the prevalence of the exposure factor (H) and the prevalence of the outcome of interest (F). In the notation of Appendix A, RR is P/Q. When RR is infinite (when Q is zero), equation 1 gives the maximum value of ϕ as $|[F/H][(1-H)/(1-F)]|$.

Suppose we compare the observed ϕ with the maximum ϕ possible *given the observed prevalence of the exposure* (H). This would compare the observed factor with the factor *in its prevalence class* having the maximum relative risk*:* $RR = \infty$ when $Q = 0$ ($b = 0$).

$$|\phi(RR,F,H)/\phi(RR=\infty,F,H)| = |H(RR\text{-}1)/[H(RR\text{-}1)+1]| \qquad 2$$

Using equation f in Appendix A we see that the expression on the right side of 2 is the attributable fraction in the population (AFP): 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 (Q) and if that exposure factor were eliminated. [3]

$$|\phi(RR,F,H) \,/\, \phi(RR=\infty,F,H)| = |AFP| \qquad 3$$

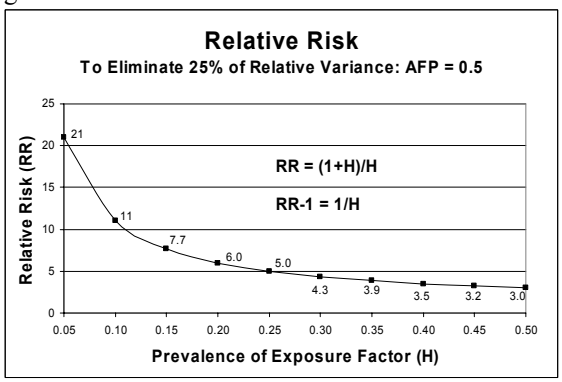
Thus the attributable fraction in the population (AFP) with an exposure prevalence (H) is the same as this **relative correlation**: the observed ϕ relative to the ϕ of a genuinely necessary factor (no exceptions; $b = 0$ which implies RR = ∞) that has the same exposure prevalence (H). Using equation f in Appendix A, we can see that RR increases monotonically with AFP:

$$RR - 1 = AFP \,/\, [H(1\text{-}AFP)] \qquad 4$$

1.5 AFP ≥ 0.5

If we require a relative correlation of 0.5 (which explains 25% of the **relative variance** per equation 3), then AFP is 0.5 and $RR\text{-}1 \geq 1/H$. See Figure 1.

Figure 1: Relative Risk versus H for $AFP = 0.5$



Reducing the relative variance in a 2x2 table by 25% may not be appropriate in a broader perspective, but it does offer a formal criterion that may be useful.

1.6 COMMENTS[4]

The attributable fraction of cases among the exposed, *AFE,* has been misrepresented as the chance that a case among those exposed was caused by their exposure.[5] Suppose that $RR = 3$ and $AFE = 67\%$. 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, the claim would be true. But that begs the question.

non-exposure groups, and in the case or non-case groups. In this discussion of AFP, F and H, 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).

In Appendix A, equations a-h define common measures of association such as relative risk (RR), the odds ratio (OR), the attributable fraction in the exposure group (AFE), the attributable fraction in the population (AFP) and the Pearson correlation coefficient (ϕ) in a 2×2 table of binary data. Equations C through G relate ϕ with these other measures of association.

The *attributable fraction in the exposure group* (AFE) is simply $(P\text{-}Q)/P$ where P is the prevalence of cases in the exposure group and Q is the prevalence of cases in the non-exposure group. See Eq. e. The *attributable fraction in the population* (AFP) is simply $(F\text{-}Q)/F$ where F is the prevalence of cases in the population. Note that this statistic (which has population in its name) can be calculated for a sample as well as for an entire population. See Abramson (1994) for a discussion of these measures.

Typically, risk is said of unwanted outcomes while prevalence is said of exposure factors (including genetic factors and assignment to a treatment or control group). If each exposure factor is a row and if each case outcome is a column, then in this paper relative risk (RR) compares row ratios and relative prevalence (RP) compares column ratios. In this context, these terms (RR and RP) provide short abbreviations that distinguish row and column ratios.

RR, AFE and OR are independent of the relative size of the exposure group (H) assuming P and Q are constant. Similarly, RP and OR are independent of the relative size of the cases (F). AFP and ϕ are dependent on the prevalence of the exposed subjects and that of cases.

[3] $\phi^2 = AFP$ times its *diagonal exchange* partner. Appendix A, Eq. P7.

[4] The expression *ad-bc* is the numerator in OR-1, RR-1, RP-1, AFE, AFP, S/F-1, N/H-1, ϕ and $(d - f{\cdot}h/n)$. These expressions differ only in their denominators. Perhaps the numerator of the difference between any two parallel ratios in a 2x2 table is *ad-bc*: e.g., *P-Q, S-F, N-H.*

[5] Source: www.toxictorts.com/relrisk.htm. "*Relative Risk: Proving Causation by the Numbers*" by Raphael Metzger, Esq.

2.1 NECESSITY AND SUFFICIENCY

In logic, necessity and sufficiency are both binary: yes or no, true or false. But in reality many factors may be required to cause an outcome and the presence of some factors vary, so necessity and sufficiency are sometimes more usefully viewed as matters of degree.

Quantitative measures for sufficiency (*S*) and necessity (*N*) are introduced in the right column of Appendix A in the context of a 2×2 table. **Quantitative sufficiency** (*S*) is defined as the fraction of the exposure group that are cases. **Quantitative necessity** (*N*) is defined as the fraction of the cases that are in the exposure group. The formulas presented in the right column of Appendix A relate various measures of association to these two constructs.[6,7]

2.2 NECESSITY VS. SUFFICIENCY

Why do epidemiologists give so much attention to exposure factors that are obviously insufficient? Most smokers do not die of lung cancer, yet public health officials try to reduce the prevalence of smoking. The 1854 cholera death rate was only 71 per 10,000 houses supplied by a particular London pump, yet John Snow recommended removing that pump handle. Hill (1987).

Epidemiologists may focus more on necessity than sufficiency. Epidemiologists may want to *reduce* disease incidence more than they want to *predict* disease. Focusing on necessity may be more important for them than focusing on sufficiency since eliminating a necessary condition is sufficient to prevent the outcome.

Unless an effect can be produced by a single sufficient cause (RARE!), producing the effect requires supplying ALL of its necessary conditions, while preventing it requires removing or eliminating only ONE of those necessary conditions. Therefore, research into prevention need identify only a SINGLE removable necessary condition, while research on production must identify ALL of the necessary conditions.

Although *RR* can be viewed as a relative sufficiency (*RR=P/Q*), *RR* can also be viewed as increasing monotonically with this measure of necessity (*N*) for a given exposure prevalence (*H*). *AFE* and *AFP* increase monotonically with *N* (for *N>H*) for a given exposure prevalence (*H*). See equations J1, K and L in Appendix A.[8]

$$RR = [N/(1-N)] / [H/(1-H)] \quad 5a$$

$$AFE = (N-H)/[N(1-H)]^{9} \quad 5b$$

$$AFP = (N-H)/(1-H) \quad 5c$$

If a 2×2 table is organized so $RR > 1$ ($S > F$) and $RP > 1$ ($N > H$), then $AFP \leq N$. For small prevalences of the exposure ($H << N$), *AFP* approaches *N*.

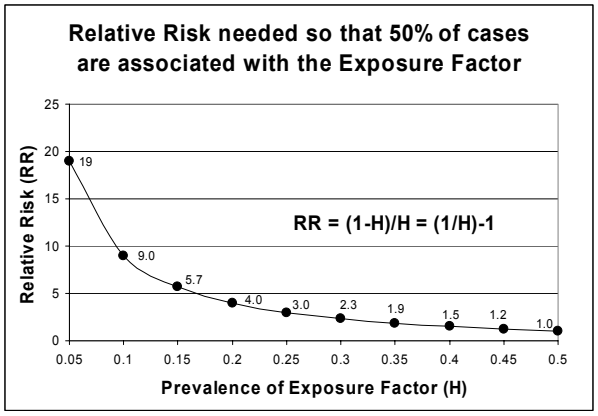
2.3 "NECESSITY" AS A POTENTIAL CAUSE

A fully established causal condition is one which is seen as necessary. An association that measures the *degree of empirical necessity* can indicate the strength of evidence for viewing a factor as a potential cause.

Epidemiologists must often recommend decisions using observational associations without having the time for experimental confirmation. Thus epidemiologists may view an exposure as "a potential cause of the cases", if "cases are more likely than not to be exposure subjects." (This is not the same as saying P > Q.)

Operationally this "more likely than not" criteria means N > 0.5. So if at least 50% of the cases are associated with the exposure, then the exposure may be viewed as a potential causal factor for Y. Using equation J1 we can relate *RR* to the exposure factor prevalence (*H*) for $N = 0.5$ as shown in Figure 2.

Figure 2: Relative Risk versus *H* for $N = 0.5$



This '*more likely than not to be associated with the exposure factor*' criteria provides a relative-risk criterion for viewing an exposure factor as a potential cause that is independent of any causal model. Consider Table 1 where 10% of adults are long-term smokers ($H = 10\%$), and the relative risk of lung cancer for long-term smokers is 9 or greater. Since $N \geq 0.5$, this data would give epidemiologists some evidence for saying, "Smoking is a potential cause of lung cancer."

Consider London cholera deaths. Most of those who drank from the well in question did not die ($S << .5$), but most of those who died drank from that well ($N > .5$).

Viewing "necessity" as a sign of a potential cause may make sense socially when the benefits of eliminating cases (e.g., lung cancer deaths) is much greater than the costs of eliminating the exposure (e.g., smoking).

The relative risk for $N = .5$ is always 2 less than the *RR* for $AFP = 0.5$. Proof: $[(1+H)/H] - [(1-H)/H] = 2$. Requiring $AFP = .5$ sets a higher standard for *RR* than $N = 0.5$. *N* is related to *RR* and *H* via equation J1 or 5a.

$$N = H \cdot RR / [H(RR-1) + 1] \quad 6$$

[6] Equations M-Q are not used herein but are shown for completeness.
[7] $RR/RP = (1-S)(1-H)/[(1-N)(1-F)]$ is elegant but over-specified.
[8] $RR = P/Q$ so $N/(1-N) = (P/Q)[H/(1-H)]$.
[9] This relation is better seen in $AFE = [1/(1-H)] - \{(1/N)[H/(1-H)]\}$.

3 ADMISSIBLE COORDINATES

We would like to know whether there are sets of algebraic variables from which we can calculate all other variables without needing to check the range of the outputs. Such sets we call sets of admissible coordinates.[10]

One set of admissible coordinates is the set of body cell counts. In a 2×2 table (Appendix A) the 4 body cells (*a*, *b*, *c* and *d*) are coordinates that always generate admissible counts and ratios. Note that the body cell (*d*) and its three adjacent margin values (*f*, *h* and *n*) are not admissible coordinates since they can generate inadmissible results (e.g., when $d > f$ or $d > h$).

Another set of admissible coordinates involves a series of linked perpendicular single ratios. In a 2×2 table for a given total count (*n*), the margin ratio (*H*), and the two ratios perpendicular to that margin (*P* and *Q*) form a set of admissible coordinates. Using these coordinates the resulting body-cell counts can never be negative.

Many combinations of ratios do not form a set of admissible coordinates. E.g., *F*, *H* and either *P* or *Q*;[11] *F, P* and *Q*;[12] *RR* and either *P* or *Q*;[13] *RR*, *F* and *H*.[14]

Unfortunately we may want to use formulas involving a set of non-admissible coordinates (e.g., *RR*, *F* and *H*) as independent variables. We want to know how to limit their values so that we do not generate any inadmissible values in the four body cells. This is done by generating formulas for the body cell values in terms of the coordinates.[15] The conditions (necessary and sufficient) for these values to be non-negative are identified. If all the body cell values (as ratios over n) are non-negative and if they sum to 1, then none of these ratios can be greater than 1. Unfortunately conditions that are both necessary and sufficient can be fairly complex.[16]

Sufficient conditions for all cell values to be admissible can be much simpler. When $RR > 1$, admissible results will always be given by these conditions:

- For *RR*, *F* and *H*, require $F < H$ or $RR \le (1\text{-}H)/(F\text{-}H)$
- For *P*, *Q* and *F*, require $P > F > Q$ or $P < F < Q$.

4 CONCLUSION

Aside from the independence of *RR* from the size of the exposure group (for a given *P* and *Q*), there are two additional reasons for using relative risk to measure the strength of an association between two binary variables. (1) *RR* increases monotonically with *AFP,* and *AFP* equals a relative correlation $|\phi/\phi_{max}|$: the observed correlation relative to that of a genuinely necessary factor having the same exposure prevalence (*H*). (2) *RR* increases monotonically with the degree of necessity (*N*) for a given exposure prevalence (*H*). By selecting minimum values for *AFP* or *N*, these relationships may provide minimum conditions for *RR* such that the associated exposure can be viewed as a potential cause.

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., 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, 173-203.

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.

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

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

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 discipline." The authors thank Dr. Shoukri for equations E1 and E2 since these equations gave us a starting point. In gratitude, we describe these equations as *Shoukri equations* and describe equal margin values as *Shoukri conditions*. Thanks to Romney Schield (Olomouc University, Czech Republic) for his validation of Shoukri equations and to Jan Hajek, Netherlands, for noting an error in our verbal description of necessity in Appendix A.[17] Dr. Schield is at Schield@augsburg.edu. This paper is posted at www.augsburg.edu/ppages/~schield.

[10] 'Coordinate' is used to designate an independent algebraic variable.
[11] If $P{\cdot}H > F$ or if $Q(1\text{-}H) > F$.
[12] If *P* and *Q* are both more – or less – than *F*.
[13] If $RR > 1/Q$, or if $RR < P$.
[14] If $F > H(RR\text{-}1)+1$ since $Q = F/[H(RR\text{-}1)+1]$.
[15] $RR = (d/h)/(b/g)$ where $g = n\text{-}h$. So, $b(h,RR) = d(n\text{-}h)/(h{\cdot}RR)$.
$f = d+b$ so that $d(RR, h, f) = (f{\cdot}h{\cdot}RR)/(h{\cdot}RR + n\text{-}h)$.
$c = h\text{-}d$ so that $c(RR, h, f) = \{h[RR(h - f) + n - h]\}/(h{\cdot}RR + n\text{-}h)$.
$b = f\text{-}d$ so that $b(RR, h, f) = [f(n - h)]/(h{\cdot}RR + n\text{-}h)$.
$a = n\text{-}h\text{-}b$, so that $a(RR, f, h) = [(n\text{-}h)(h{\cdot}RR + n\text{-}h\text{-}f)]/(h{\cdot}RR + n\text{-}h)$.
[16] Denominators in the preceding footnote are always non-negative, so numerators must never be negative. $c \ge 0$ iff $[RR(h - f) + n - h] \ge 0$. $a \ge 0$ iff $(h{\cdot}RR + n\text{-}h\text{-}f) \ge 0$.

[17] In Appendix A, Romney Schield validated equations E1 and E2 (Shoukri's equations) and created equations J1 and J2. Milo Schield designed the constructs N and S, derived equation F2 in Schield (1999) and derived equations M3 and P2. Equation I is Bayes rule. Thomas Burnham created all the other equations from A1 through Q, and validated all the A-Q equations using the Derive software.

Appendix A. 2×2 COUNT TABLE IDENTITIES[18]

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

Definitions and Basic Relations:[19]

a. $E = e/n$, $F = f/n$, $G = g/n$, $H = h/n$. $P = d/h$, $Q = b/g$.
b. Relative Risk: $RR = (d/h)/(b/g) = (d\cdot g)/(b\cdot h) = P/Q$
c. Relative Prevalence: $RP = (d/f)/(c/e) = (e\cdot d)/(c\cdot f)$
Note: *RR* compares rows; *RP* compares columns.
d. Odds Ratio: $OR = a\cdot d/(b\cdot c) = (a/b)/(c/d) = (a/c)/(b/d)$
e. Attributable fraction in exposure group (*AFE*):
$AFE = (P - Q)/P = (RR-1)/RR = 1-(1/RR)$
f. Attributable fraction in population (*AFP*):
$AFP = (F - Q)/F = H(RR-1)/[H(RR-1)+1]$
g. Phi: $\phi = (a\cdot d - b\cdot c) / w = r$ where $w^2 = e\cdot f\cdot g\cdot h$[20]
h. Case Prevalence: $F = P\cdot H + Q(1-H)$

Identities Using Existing Factors

A1. $F=P-(P-Q)(1-H) = P[1-(1-H)(RR-1)/RR]$
A2. $H = (F-Q)/(P-Q)$. $F = Q[H(RR-1)+ 1]$.
B1. $AFP=H\cdot AFE/[1-AFE(1-H)]$. $AFE=AFP/[H+AFP(1-H)]$
B2. $OR = RR[H(RR-1)+(1-F)] / [(RR-1)(H-F)+(1-F)]$
B3. $[(RR-1)/RR] / [(RP-1)/RP] = (1-F)/(1-H)$
C. $\phi^2(d,f,h,n) = (d\cdot n - f\cdot h)^2/[f\,(n-f)\,h\,(n-h)]$
D. $\phi = (P-Q)\sqrt{G\cdot H/(E\cdot F)}$ Proportions test[21]
E1. $a = (e\cdot g + \varphi\cdot w)/n$ $\quad b = (f\cdot g - \varphi\cdot w)/n$
$c = (e\cdot h - \varphi\cdot w)/n$ $\quad d = (f\cdot h + \varphi\cdot w)/n$
E2. When $e = f = g = h$: See Shoukri (2000)
$\phi = (OR-1)/(\sqrt{OR}+1)^2 = (\sqrt{OR}-1)/(\sqrt{OR}+1)$

F1. $\phi^2(P,F,Q) = [(P-F)/(1-F)][(F-Q)/F]$[22] $\quad P=d/h$

F2. $\phi^2(P,Q,H) = \dfrac{H(P-Q)^2(1-H)}{[(1-Q-H(P-Q)][H(P-Q)+Q]}$

G1. $\phi^2(RR,F,H) = \left[\dfrac{F}{H}\dfrac{(1-H)}{(1-F)}\right]\left[\dfrac{H(RR-1)}{H(RR-1)+1}\right]^2$

G2. $\phi^2(AFP,F,H) = [F/(1-F)][(1-H)/H]AFP^2$

G3. $\phi^2(RR,RP,H) = \dfrac{H(RR-1)[H(RR-1)+1-(RR/RP)]}{[H(RR-1)+1]^2}$

G4. $\phi^2(OR,RR,RP) = \dfrac{(RR-1)(RP-1)(OR-RR)(OR-RP)}{RR(OR-1)^2\,RP}$

G5. $\phi^2(OR,RR,H) = \dfrac{H(1-H)(RR-1)(OR-RR)}{[OR-H(OR-RR)][H(RR-1)+1]}$

G6. $\phi^2 = \dfrac{[(OR-1)(1-F)-(RR-1)][RR(OR-1)F-OR+RR]}{F(1-F)RR(OR-1)^2}$

$\phi^2(OR, F, H)$ is intractable with a large radical.
$\phi^2(P, F, H) = \phi^2(S, F, H)$. See equation Q.

Identities Involving 'Necessity' or 'Sufficiency'[18]

Let $S = d/h$. S = 'sufficiency' of *exposure* for *case.*
Let $N = d/f$. N = 'necessity' of *exposure* for *case.*

H1. $(RR-1)/(1-S) = (RP-1)/(1-N) = OR - 1$
H2. $RR/RP = (1-S)(1-H)/[(1-N)(1-F)]$
I. $S/N = (d/h)/(d/f) = f/h = F/H$ Bayes' Rule
J1. $RR = [N/(1-N)] / [H/(1-H)] = S(1 - H)/(F - S\cdot H)$
J2. $RR-1 = [(1-F)/(1-N)]\{(RP-1)/[F(RP-1)+1]\}$
K. $AFE = (N-H)/[N(1-H)] = [(S-F)/S]/(1-H)$
L. $AFP = (N-H)/(1-H) = [(S-F)/F][H/(1-H)]$

Four factor ϕ with S or N: [These are over-specified.][18]

M1. $\phi^2(RR,S,N,RP) = S\cdot N[(RR-1)/RR][(RP-1)/RP]$

M2. $\phi^2(S,N,H,F) = [(S-F)(N-H)]/[(1-F)(1-H)]$

M3. $\Phi^2(RR,S,F,H) = [S(RR-1)/RR]^2\{H(1-H)/[F(1-F)]\}$

Three-factor ϕ with S or N, and with no margin value:

N1. $\phi^2(RR,S,N) = \dfrac{S\cdot N(1-N)(RR-1)^2}{RR[RR(1-N)+N-S]}$

N2. $\phi^2(RR,RP,S) = \dfrac{S[RR-RP(1-S)-S]}{RR}\dfrac{(RP-1)}{RP}$

N3. $\phi^2(OR,RR,N) = \dfrac{N(1-N)(RR-1)(OR-RR)}{RR[OR(1-N)+N]}$

N4. $\phi^2(OR,S,N) = \dfrac{N(1-N)S(1-S)(OR-1)^2}{[(OR-1)(1-N)+1][(OR-1)(1-S)+1]}$

Underdetermined: $\phi^2(OR,RR,S)$ and $\phi^2(OR,RP,N)$

Three-factor ϕ with S or N, and with one margin value:

P1. $\phi^2(RR,S,H) = \dfrac{H\cdot S(RR-1)^2(1-H)}{[RR(1-H\cdot S)-S(1-H)][H(RR-1)+1]}$

P2. $\phi^2(RR,S,F) = \dfrac{[(S/F)-1][RR-(S/F)]}{RR}\dfrac{F}{(1-F)}$

P3. $\phi^2(RR,N,F) = [F/(1-F)]N(1-N)(RR-1)^2/RR$

P4. $\phi^2(N,S,F) = N[(S-F)/(1-F)][(S-F)/(S-F\cdot N)]$

P5. $\phi^2(OR,S,H) = \dfrac{H(1-H)S(1-S)(OR-1)^2}{[(OR-1)H(1-S)+1][(OR-1)(1-S\cdot H)+1]}$

P6. $\phi^2(OR,S,F) = \dfrac{(S-F)[F(OR-1)(1-S)-S+F]}{F(1-F)[(1-S)(OR-1)+1]}$

P7. $\phi^2(AFP,S,F) = AFP(S-F)/(1-F)$

Underdetermined: $\phi^2(RR,N,H)$ and $\phi^2(RP,S,F)$.

Three-factor ϕ with S or N, and with two margin values:

Q. $\phi^2(S,F,H) = [H(S-F)]^2/[F(1-F)H(1-H)]$

Relations Between Admissible Values:
Some equations generate inadmissible results for some input values.

- If $P > Q$, then $RR > 1$, $P > F > Q$, $RR > F/Q$ and $RR > P/F$
- If $H < F$, then $RR \le (1-H)/(F-H)$ since $P \le 1$.

Over-specified equations D and M1-M3 allow inconsistent inputs.
E.g., $\phi^2(F, H, N, Q) = (N-H)(F-Q)/[H(1-F)]$.

Diagonal Exchange: If cells *b* and *c* exchange then *OR* and ϕ are unchanged while these exchange: *RR* & *RP*, *F* & *H*, *E* & *G* and *S* & *N*. If a valid equation involves only these variables and they are exchanged the new equation is valid. e.g. J1:$\rightarrow RP=[S/(1-S)]/[F/(1-F)]$

[18] Equations may be over-specified: allow inadmissible inputs.
[19] Lower case indicates counts; upper case indicates ratios.
[20] $X^2 = \Sigma[(\text{actual value} - \text{expected value})^2 / \text{expected value}] = n\cdot\phi^2$
[21] When squared and multiplied by *n*, this is the test for independence.
[22] Note that the right-hand term $[(F-Q)/F]$ is *AFP*.

APPENDIX B: AUXILIARY IDENTITIES

BAYES' COMPARISONS

We define *Bayes' comparisons* as *S/F* or *N/H*. A Bayes comparison is a ratio comparison of two ratios. Bayes rule says $P(F|H)\cdot P(H) = P(H|F)\cdot P(F)$. Let $S = P(F|H)$, $N = P(H|F)$, $F = P(F)$ and $H = P(H)$. Bayes' rule equates two Bayes' comparisons: $S/F = N/H$. If $S/F = k$ then $N/H = k$.

> If "Those having been in prison are **4** times as likely *to have low IQ* as are those in the entire population", then "Low IQ adults are **4** times as likely *to have been in prison* as are those in the entire population."[23]

The difference between these Bayes' comparisons is so subtle that some may not realize what is changed.

The benefit of this comparison is that even if one inadvertently reverses the part and test whole, the number in the comparison remains unchanged. But the need for caution is not diminished. Bayes' comparison look-alikes can confound the unwary reader.

Bayes rule means: If "Cases are *k* times as likely *to be exposed subjects* as are those in the general population," then "Exposed subjects are *k* times as likely *to be cases* as are those in the general population."[24]

Bayes' rule is one of many statements that reflect a diagonal exchange. (See Diagonal Exchange in Appendix A where *a*, *d* and *n* exchange with themselves.) Under this diagonal exchange, equation 7 is valid.

$$N/H = (d/f)/(h/n) = d\cdot n/(f\cdot h) = (d/h)/(f/n) = S/F \qquad 7$$

This exchange may appear to work only with the *d* cell in the lower right corner of the four body cells. But we can make a row or column exchange of the index values so that any of the four body cells can be the *d* cell.

GETTING *RR* FROM A BAYES' COMPARISON

A relative risk (which is row based) can be obtained from the *N* and *H* in a column-based Bayes comparison.

$$RR = S/Q = (S/F)/(Q/F) = (N/H)/[(1-N)/(1-H)] \qquad 8$$

Suppose 40% of prisoners are black (*N*) and 10% of the population are black (*H*). Blacks are 4 times (N/H = 40/10) as likely to be in prison as are the general population. From *N* and *H*, we see that 60% of prisoners are non-black (*1-N*) and 90% of the population are non-black (*1-H*). So non-blacks are two-thirds (60/90) as likely to be in prison as are the general population: $Q/F = (1-N)/(1-H)$.[25] Thus, blacks are 6 times as likely to be in prison *as are non-blacks*: $RR = 4/(2/3) = 6$.

NON-RESPONSE BIAS EFFECT SIZE

Even though we cannot typically measure non-response bias, we can determine the associated effect size necessary to generate a particular outcome. Suppose the prevalence of responders among those polled is *H*. Among the responders, the fraction who say "Yes" is *P*. Typically the fraction who would have said "Yes" (*Q*) among the non-responders is unobserved. Thus the fraction who would have said "Yes" in the population (*F*) is unobserved as is the non-response bias: *P-Q*. But we can determine the non-response bias (*P-Q*) needed to generate a specified value of *F*.

$$P-Q = (P-F)/(1-H) \qquad 9a$$
$$Q = P - (P-F)/(1-H) = [F - (P\cdot H)]/(1-H) \qquad 9b$$

Consider a two-way election where $P > ½$. The non-response bias needed to give a tie ($F = ½$) for $H < ½$ is:

$$P-Q = (P- ½) / (1-H) \qquad 9c$$
$$Q = P - (P- ½)/(1-H) = [½-(P\cdot H)]/(1-H) \qquad 9d$$

As the fraction of non-respondents (1-H) increases, *H* decreases so that the size of the non-response bias (P-Q) necessary to yield a tie decreases and Q increases.

Stating the non-response bias or the value of *Q* necessary to give a specific value in the population (*F*) gives readers a better context for the uncertainty involved. For example, if 55% of the respondents say "Yes" but 25% of population do not respond, then a prediction of a majority of "Yes" in the population would fail if more than the 65% of the non-respondents would say "No" (if non-response bias exceeds 20 percentage points).

IDENTITY CONNECTING *RR* AND *RP*

An identity involving relative risk (*RR*) and relative prevalence (*RP*) is given by equation B3 (Appendix A).

$$[(RR-1)/RR] / [(RP-1)/RP] = (1-F)/(1-H) \qquad 10$$

The terms *RR*-1 and *RP*-1 can be described using "involvement." Over-involvement occurs if *RR*>1 or *RP*>1. Under-involvement occurs if *RR*<1 or *RP*<1.

In equation 10, the right-hand terms are never negative. So *RR*-1 and *RP*-1 must have the same sign. This justifies the over-involvement rule:[26] over-involvement by row (*RR*>1) implies over-involvement by column (*RP*>1), and vice versa.

To see how this works, consider deaths classified by smoking and lung cancer. Suppose that among the deceased, smokers were more prevalent among those who had lung cancer than among those who didn't (*RP*>1). Even if we didn't know the data in Table 1, we could say, "Dying of lung cancer was more likely for smokers than non-smokers" (*RR* > 1), or "Smoking had a positive 'influence' on dying of lung cancer."

---

[23] In *prevalent* language: If "having a low IQ was four times as *prevalent* among those having been in prison as among [those in] the general population" then "having been in prison was 4 times as prevalent among those with low-IQ as in the general population."

[24] OR form: If "The OR of cases among exposed versus non-exposed is *k*" then "The OR of exposed among cases versus non-cases is *k*."

[25] *Q=b/g. F=f/n. 1-N =b/f.* Q/F = (b/g)/(f/n)=(b/f)/(g/*n*)=*(1-N)/(1-H)*

[26] Carlson, William L. and Betty Thorn. *Applied Statistical Methods,* Prentice Hall. p. 151