In: The Handbook of Research Synthesis and Meta Analysis, 2nd Ed. Cooper, Hedges & Valentine (eds.). Sage.

13

EFFECT SIZES FOR DICHOTOMOUS DATA

JOSEPH L. FLEISS
Columbia University

JESSE A. BERLIN
Johnson & Johnson Pharmaceutical Research and Development

CONTENTS

13.1 Introduction 238
13.2 Difference Between Two Probabilities 239
13.2.1 A Critique of the Difference 239
13.2.2 Inference About the Difference 239
13.2.3 An Example 239
13.3 Ratio of Two Probabilities 239
13.3.1 Rate Ratio 239
13.2 Inferences About the Rate Ratio 240
13.3 Problems with the Rate Ratio 240
13.4 Phi Coefficient 241
13.4.1 Inference About φ 241
13.4.2 Variance Estimation for φ 242
13.4.3 Problems with the Phi Coefficient 242
13.5 Odds Ratio 243
13.5.1 A Single Fourfold Table 244
13.5.2 An Alternative Analysis: $(O-E)/V$ 244
13.5.3 Inference in the Presence of Covariates 245
13.5.3.1 Regression Analysis 245
13.5.3.2 Mantel-Haenszel Estimator 246
13.5.3.3 Combining Log Odds Ratios 246
13.5.3.4 Exact Stratified Method 247
13.5.3.5 Comparison of Methods 247
13.5.3.6 Control by Matching 247
13.5.4 Reasons for Analyzing the Odds Ratio 248
13.5.5 Conversion to Other Measures 250

13.6 Acknowledgments 251
13.7 References 251

13.1 INTRODUCTION

In many studies measurements are made on binary (dichotomous) rather than numerical scales. Examples include studies of attitudes or opinions (the two categories for the response variable being agree or disagree with some statement), case-control studies in epidemiology (the two categories being exposed or not exposed to some hypothesized risk factor), and intervention studies (the two categories being improved or unimproved or, in studies of medical interventions, experiencing a negative event or not). In this chapter we present and analyze four popular measures of association or effect appropriate for categorical data: the difference between two probabilities, the ratio of two probabilities, the phi coefficient, and the odds ratio. Considerations in choosing among the various measures are presented, with an emphasis on the context of research synthesis. The odds ratio is shown to be the measure of choice according to several statistical criteria, and the major portion of the discussion is devoted to methods for making inferences about this measure under a variety of study designs. The chapter wraps up by discussing converting the odds ratio to other measures that have more straightforward substantive interpretations.

Consider a study in which two groups are compared with respect to the frequency of a binary characteristic. Let π_1 and π_2 denote the probabilities in the two underlying populations of a subject being classified into one of the two categories, and let P_1 and P_2 denote the two sample proportions based on samples of sizes n_1 and n_2. Three of the four parameters to be studied in this chapter are the simple difference between the two probabilities,

$$\Delta = \pi_1 - \pi_2; \tag{13.1}$$

the ratio of the two probabilities, or the rate ratio (referred to as the risk ratio or relative risk in the health sciences),

$$RR = \pi_1 / \pi_2; \tag{13.2}$$

and the odds ratio,

$$\omega = (\pi_1/(1-\pi_1))/(\pi_2/(1-\pi_2)). \tag{13.3}$$

The fourth measure, φ (the phi coefficient), is defined later.

The parameters Δ and φ are such that the value 0 indicates no association or no difference. For the other two parameters, RR and ω, the logarithm of the measure is typically analyzed to overcome the awkward features that the value 1 indicates no association or no difference, and that the finite interval from 0 to 1 is available for indexing negative association, but the infinite interval from 1 on up is available for indexing positive association. With the parameters transformed so that the value 0 separates negative from positive association, we present for each the formula for its maximum likelihood estimator, say L, and for its non-null standard error, say SE. (A non-null standard error is one in which no restrictions are imposed on the parameters, in particular, no restrictions that are associated with the null hypothesis.)

We assume throughout that the sample sizes within each individual study are large enough for classical large-sample methods assuming normality to be valid. The quantities L, SE and

$$w = 1/SE^2 \tag{13.4}$$

are then adequate for making the following important inferences about the parameter being analyzed within each individual study, and for pooling a given study's results with the results of others. Let C_α denote the value cutting off the fraction α in the upper tail of the standard normal distribution. The interval $L \pm C_\alpha SE$ is, approximately, a $100(1-2\alpha)$ percent confidence interval for the parameter that L is estimating. As detailed in chapter 14 of this volume (see, in particular, formula 14.14), the given study's contribution to the numerator of the pooled (meta-analytic) estimate of the parameter of interest is, under the model of fixed effects, wL, and its contribution to the denominator is w. (In a so-called fixed-effects model for k studies, the assumption is made that there is interest in only these studies. In a random effects model, the assumption is made that these k studies are a sample from a larger population of studies. We restrict our attention to fixed-effects analyses.)

13.2 THE DIFFERENCE BETWEEN TWO PROBABILITIES

13.2.1 A Critique of the Difference

The straightforward difference, $\Delta = \pi_1 - \pi_2$, is the simplest parameter for estimating the effect on a binary variable of whatever characteristic or intervention distinguishes group 1 from group 2. Its simplicity and its interpretability with respect to numbers of events avoided (or caused) are perhaps its principal virtues, especially in the context of research syntheses. A technical difficulty with Δ is that its range of variation is limited by the magnitudes of π_1 and π_2: the possible values of Δ when π_1 and π_2 are close to 0.5 are greater than when π_1 and π_2 are close to 0 or to 1. If the values of π_1 and π_2 vary across the k studies whose results are being synthesized, the associated values of Δ may also vary. There might then be the appearance of heterogeneity (that is, nonconstancy in the measure of effect) in this scale of measurement due to the mathematical constraints imposed on probabilities rather than to substantive reasons. Empirically, two separate investigations have found that across a wide range of meta-analyses, heterogeneity among studies is most pronounced when using the difference as the measure of effect (Deeks 2002; Engels et al. 2000).

The example in table 13.1 illustrates this phenomenon. The values there are mortality rates from lung cancer in four groups of workers: exposed or not to asbestos in the work place cross-classified by cigarette smoker or not (Hammond, Selikoff, and Seidman 1979). The asbestos–no asbestos difference in lung cancer mortality rates for cigarette smokers is 601.6 to 122.6, or approximately 500 deaths per 100,000 person-years. Given the mortality rate of 58.4 per 100,000 person-years for nonsmokers exposed to asbestos, it is obviously impossible for the asbestos-no asbestos difference between lung cancer mortality rates for nonsmokers to come anywhere near 500 per 100,000 person-years. Heterogeneity—a difference between smokers and nonsmokers in the effect of exposure to asbestos, with effect measured as the difference between two rates—exists, but it is difficult to assign biological meaning to it.

13.2.2 Inference about the Difference

In spite of the possible inappropriateness for a meta-analysis of the difference between two probabilities, the investigator might nevertheless choose to use Δ as the measure of effect. The estimate of the difference is simply

$$D = P_1 - P_2$$

The factor w by which a study's estimated difference is to be weighted in a fixed-effects meta-analysis is equal to the reciprocal of its estimated squared standard error,

$$w = 1/SE^2 = \left(\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}\right)^{-1}. \qquad (13.5)$$

Table 13.1 Mortality Rates from Lung Cancer (Per 100,000 Person-Years)

	Exposed to Asbestos	
Smoker	Yes	No
Yes	601.6	122.6
No	58.4	11.3

SOURCE: Authors' compilation.

13.2.3 An Example

Consider as an example a comparative study in which one treatment was applied to a randomly constituted sample of $n_1 = 80$ subjects and the other was applied to a randomly constituted sample of $n_2 = 70$ subjects. If the proportions unimproved are $P_1 = 0.60$ and $P_2 = 0.80$, the value of D is -0.20 and the standard error of D is

$$SE = \left(\frac{0.60 \times 0.40}{80} + \frac{0.80 \times 0.20}{70}\right)^{1/2} = 0.0727.$$

The resulting weighting factor is $w = 1/SE^2 = 189.2$. An approximate 95 percent confidence interval for Δ is $D \pm 1.96SE$. In this example, the interval is $-0.20 \pm 1.96 \times 0.0727$, or the interval from -0.34 to -0.06.

13.3 RATIO OF TWO PROBABILITIES

13.3.1 Rate Ratio

The ratio of two probabilities, $RR = \pi_1/\pi_2$, is another popular measure of the effect of an intervention. Its use requires that one be able to distinguish between the two outcomes so that one is in some sense the undesirable one and the other is in some sense the preferred one.

The names given to this measure in the health sciences, the *risk* ratio or relative *risk*, reflect the fact that the measure is not symmetric (that is, $\pi_1/\pi_2 \neq (1-\pi_1)/(1-\pi_2)$), but is instead the ratio of the first group's probability of an undesirable outcome to the second's probability. (In fact, in the health sciences, the term *rate* is often reserved for quantities involving the use of person-time, that is, the product of the number of individuals and their length of follow-up. For the remainder of this chapter, the term will be used in its more colloquial sense, or simply a number of events divided by a number of individuals.) *RR* is a natural measure of effect for those investigators accustomed to comparing two probabilities in terms of their proportionate (relative) difference, that is, their ratio. If π_1 is the probability of an untoward outcome in the experimental group and if π_2 is the same in the control or comparison group, the proportionate reduction in the likelihood of such an outcome is

$$\frac{\pi_2 - \pi_1}{\pi_2} = 1 - RR \tag{13.6}$$

Inferences about a proportionate reduction may therefore be based on inferences about the associated relative risk.

13.3.2 Inferences About the Rate Ratio

RR is estimated simply as

$$rr = P_1/P_2.$$

The range of variation of *rr* is an inconvenient one for drawing inferences. Only the finite interval from 0 to 1 is available for indexing a lower risk in population 1, but the infinite interval from 1 up is, theoretically, available for indexing a higher risk in population 1. A corollary is that the value 1 (rather than the more familiar and convenient value 0) indexes no differential risk. It is standard practice to undo these inconveniences by carrying out one's statistical analysis on the logarithms of the rate ratios, and by transforming point and interval estimates back to the original units by taking antilogarithms of the results. An important by-product is that the sampling distribution of the logarithm of *rr* is more nearly normal than the sampling distribution of *rr*.

The large sample standard error of $\ln(rr)$, the natural logarithm of *rr*, may be estimated as

$$SE = \left(\frac{1-P_1}{n_1 P_1} + \frac{1-P_2}{n_2 P_2}\right)^{1/2} \tag{13.7}$$

An approximate two-sided $100(1-\alpha)$ percent confidence interval for *RR* is obtained by taking the antilogarithms of the limits

$$\ln(rr) - C_{\alpha/2} SE \tag{13.8}$$

and

$$\ln(rr) + C_{\alpha/2} SE. \tag{13.9}$$

Consider again the numerical example in section 13.2.3 $n_1 = 80$, $P_1 = 0.60$; $n_2 = 70$ and $P_2 = 0.80$. The estimated rate ratio is $rr = 0.60/0.80 = 0.75$, so group 1 is estimated to be at a risk that is 25 percent less than group 2's risk. Suppose that a confidence interval for *RR* is desired. One first obtains the value $\ln(rr) = -0.2877$, and then obtains the value of its estimated standard error,

$$SE = \left(\frac{0.40}{80 \times 0.60} + \frac{0.20}{70 \times 0.80}\right)^{1/2} = 0.0119^{1/2} = 0.1091$$

(the associated weighting factor is $w = 1/SE^2 = 84.0$). An approximate 95 percent confidence interval for $\ln(RR)$ has as its lower limit

$$\ln(RR_L) = -0.2877 - 1.96 \times 0.1091 = -0.5015$$

and as its upper limit

$$\ln(RR_U) = -0.2877 + 1.96 \times 0.1091 = -0.0739.$$

The resulting interval for *RR* extends from $RR_L = \exp(-0.5015) = 0.61$ to $RR_U = \exp(-0.0739) = 0.93$. The corresponding interval for the proportionate reduction in risk, finally, extends from $1 - RR_U = 0.07$ to $1 - RR_L = 0.39$, or from 7 percent to 39 percent. Notice that the confidence interval is not symmetric about the point estimate of 25 percent.

13.3.3 Problems with the Rate Ratio

Thanks to its connection with the intuitively appealing proportionate reduction in the probability of an undesirable response, the rate ratio continues to be a popular measure of the effect of an intervention in controlled studies. It is also a popular and understandable measure of association in nonexperimental studies. There are at least two technical problems with this measure, however. If the chances are high in the two groups being compared (experimental treatment versus control or exposed versus not exposed) of a subject's experiencing the outcome

under study, many values of the rate ratio are mathematically impossible. For example, if the probability π_2 in the comparison group is equal to 0.40, only values for RR in the interval $0 \leq RR \leq 2.5$ are possible. The possibility therefore exists of the appearance emerging of study-to-study heterogeneity in the value of RR only because the studies differ in their values of π_2. As we will see in section 13.5, this kind of constraint does not characterize the odds ratio, a measure related to the rate ratio.

A second difficulty with RR, one that is especially important in epidemiological studies, is that it is not estimable from data collected in retrospective studies. Let E and $\bar{E}$ denote exposure or not to the risk factor under study, and let D and $\bar{D}$ denote the development or not of the disease under study. The rate ratio may then be expressed as the ratio of the two conditional probabilities of developing disease,

$$RR = \frac{P(D \mid E)}{P(D \mid \bar{E})}.$$

Studies that rely on random, cross-sectional sampling from the entire population and those that rely on separate random samples from the exposed and unexposed populations permit $P(D|E)$ and $P(D|\bar{E})$, and thus RR, to be estimated. In a retrospective study, subjects who have developed the disease under investigation, as well as subjects who have not, are evaluated with respect to whether or not they had been exposed to the putative risk factor. Such a study permits $P(E|D)$ and $P(E|\bar{D}R)$ to be estimated, but these two probabilities are not sufficient to estimate RR. We will see in Section 5 that, unlike RR, the odds ratio is estimable from all three kinds of studies. Jonathan Deeks presented an in-depth discussion of the various measures of effect described here, including consideration of rate ratio for a harmful, versus a beneficial, outcome from the same study (2002). He noted, for example that the apparent heterogeneity of the rate ratio can be very sensitive to the choice of benefit versus harm as the outcome measure.

13.4 PHI COEFFICIENT

13.4.1 Inference About φ

Consider a series of cross-sectional studies in each of which the correlation coefficient between a pair of binary random variables, X and Y, is the measure of interest. Table 13.2 presents notation for the underlying parameters and table 13.3 presents notation for the observed

Table 13.2 Underlying Probabilities Associated with Two Binary Characteristics

	Y		
X	Positive	Negative	Total
Positive	π_{11}	π_{12}	$\pi_{1\cdot}$
Negative	π_{21}	π_{22}	$\pi_{2\cdot}$
Total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	1

SOURCE: Authors' compilation.

Table 13.3 Observed Frequencies in a Study Cross-Classifying Subjects

	Y		
X	Positive	Negative	Total
Positive	n_{11}	n_{12}	$n_{1\cdot}$
Negative	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

SOURCE: Authors' compilation.

frequencies in the 2×2 table cross-classifying subjects' categorizations on the two variables. Let the two levels of X be coded numerically as 0 or 1, and let the same be done for Y. The product-moment correlation coefficient in the population between the two numerically coded variables is equal to

$$\varphi = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\sqrt{\pi_{1\cdot}\pi_{2\cdot}\pi_{\cdot 1}\pi_{\cdot 2}}} \tag{13.10}$$

and its maximum likelihood estimator (assuming random, cross-sectional sampling) is equal to

$$\hat{\varphi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}. \tag{13.11}$$

Note that $\hat{\varphi}$ is closely related to the classical chi-square statistic for testing for association in a fourfold table: $\chi^2 = n \cdot\cdot \hat{\varphi}^2$. If numerical values other than 0 or 1 are assigned to the categories of X or Y, the sign of $\hat{\varphi}$ may change but not its absolute value.

The value of $\hat{\varphi}$ for the frequencies in table 13.4 is

$$\hat{\varphi} = \frac{135 \times 10 - 15 \times 40}{\sqrt{150 \times 50 \times 175 \times 25}} = 0.13,$$

a modest association.

Table 13.4 Hypothetical Frequencies in a Fourfold Table

	Y		
X	Positive	Negative	Total
Positive	135	15	150
Negative	40	10	50
Total	175	25	200

SOURCE: Authors' compilation.

Yvonne Bishop, Stephen Fienberg, and Paul Holland ascribe the following formula for the large-sample standard error of $\hat{\varphi}$ to Yule (1975, 381–82):

$$\frac{1}{\sqrt{n_{..}}}\left(1-\hat{\varphi}^2+\hat{\varphi}\left(1+\frac{\hat{\varphi}^2}{2}\right)\frac{(p_{1\cdot}-p_{2\cdot})(p_{\cdot 1}-p_{\cdot 2})}{\sqrt{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}}}-\frac{3}{4}\hat{\varphi}^2\left[\frac{(p_{1\cdot}-p_{2\cdot})^2}{p_{1\cdot}p_{2\cdot}}+\frac{(p_{\cdot 1}-p_{\cdot 2})^2}{p_{\cdot 1}p_{\cdot 2}}\right]\right)^{1/2}. \quad (13.12)$$

where $p_{1\cdot}$ and $p_{2\cdot}$ are the overall (marginal) row probabilities and $p_{\cdot 1}$ and $p_{\cdot 2}$ are the overall (marginal) column probabilities.

For the data under analysis,

$$SE=\frac{1}{\sqrt{200}}(1.245388)^{1/2}=0.079 \quad (13.13)$$

13.4.2 Variance Estimation for φ

A straightforward large sample procedure known as the jackknife (Quenouille 1956; Tukey 1958) was used extensively in the past and provided an excellent approximation to the standard error of $\hat{\varphi}$ (as well as the standard errors of other functions of frequencies (Fleiss and Davies 1982) without the necessity for memorizing or programming a formula as complicated as the one in expression 13.12. The jackknife involves reestimating the φ coefficient (or other function) with one unit deleted from the sample and repeating this process for each of the units in the sample, then calculating a mean and variance using these estimates.

With the existence of faster and more powerful computers, methods for variance estimation that are computationally intensive have become possible. In particular, the so-called bootstrap approach has achieved some popularity (Efron and Tibshirani 1993; Davison and Hinkley 1997). Essentially, the bootstrap method involves resampling from the population giving rise to the observed sample and taking an empirical variance estimate, or directly estimating a 95 percent confidence interval from the distribution of resampled estimates. Technically, the steps are as follows: first, sample an observation (with replacement) from the observed data (that is, put the sampled observation into the new dataset, record its value, then put it back so it can be sampled again); second, repeat this n times for a sample of n observations; third, when n observations have been sampled, calculate the statistic of interest (in this case, $\hat{\varphi}$); and, fourth, repeat this entire process a large number of times, say 1,000, generating 1,000 estimates of the value of interest. The bootstrap estimate of φ is just the sample mean of the 1,000 estimates. A variance estimate can then be obtained empirically from the usual sample variance of the 1,000 estimates, or a 95 percent confidence interval can be obtained directly from the distribution of the 1,000 estimates.

Although the bootstrap is technically described in terms of the notion of sampling with replacement, it may be easier to think of the process as generating an infinitely large population of individual subjects, that exactly reflects the distribution of observations in the observed sample. Each bootstrap iteration can then be considered a sample from this population.

13.4.3 Problems with the Phi Coefficient

A number of theoretical problems with the phi coefficient limit its usefulness as a measure of association between two random variables. If the binary random variables X and Y result from dichotomizing one or both of a pair of continuous random variables, the value of φ depends strongly on where the cut points are set (Carroll 1961; Hunter and Schmidt 1990). A second problem is that two studies with populations having different marginal distributions (for example, $\pi_{1\cdot}$ for one study different from $\pi_{1\cdot}$ for another) but otherwise identical conditional probability structures (for example, equal values of $\pi_{11}/\pi_{1\cdot}$ in the two study populations) may have strongly unequal phi coefficients. An untoward consequence of this phenomenon is that there will be the appearance of study to-study heterogeneity (that is, unequal correlations across studies), even though the associations defined in terms of the conditional probabilities are identical.

Finally, the phi coefficient shares with other correlation coefficients the characteristic that it is an invalid measure of association when a method of sampling other

Table 13.5 Hypothetical Fourfold Tables, Problems with Phi Coefficient

	Y		
X	Positive	Negative	Total
Second Study			
Positive	45	5	50
Negative	120	30	150
Total	165	35	200
Third Study			
Positive	90	10	100
Negative	80	20	100
Total	170	30	200

SOURCE: Authors' compilation.
NOTE: Data for the original study are in table 13.4.

than cross-sectional (multinomial and simple random sampling are synonyms) is used. The two fourfold tables in table 13.5, plus the original one in table 13.4, illustrate this as well as the previous problem. The associations are identical in the three tables in the sense that the estimated conditional probabilities that Y is positive, given that X is positive, are all equal to 0.90, and the estimated conditional probabilities that Y is positive, given that X is negative, are all equal to 0.80. If these tables summarized the results of three studies in which the numbers of subjects positive on X and negative on X were sampled in the ratio 150:50 in the first study, 50:150 in the second and 100:100 in the third, the problem becomes apparent when the three phi coefficients are calculated. Unlike the other three measures of association studied in this chapter, the values of φ vary across the three studies, from 0.13 to 0.11 to 0.14. The important conclusion is not that the values of φ are nearly equal one to another, but that the values are not identical (for further discussion of the phi coefficient, see Bishop, Fienberg, and Holland 1975, 380–83).

13.5 ODDS RATIO

We have seen that the phi coefficient is estimable only from data collected in a cross-sectional study and that the simple difference and the rate ratio are estimable only from data collected using a cross-sectional or prospective study. The odds ratio, however, is estimable from data collected using any of the three major study designs: cross-sectional, prospective or retrospective. Consider a bivariate population with underlying multinomial probabilities given in table 13.2. The odds ratio, sometimes referred to as the cross-product ratio, associating X and Y is equal to

$$\omega = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{13.17}$$

If the observed multinomial frequencies are as displayed in table 13.3, the maximum likelihood estimator of ω is

$$o = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{13.18}$$

Suppose that the study design calls for $n_{1\cdot}$ units to be sampled from the population positive on X, and for $n_{2\cdot}$ units to be sampled from the population negative on X. $P(Y+|X+)$, the conditional probability that Y is positive given that X is positive, is equal to $\pi_{11}/\pi_{1\cdot}$, and the odds for Y being positive, conditional on X being positive, are equal to

$$\begin{aligned}\text{Odds}\,(Y+|X+) &= P(Y+|X+)\,/\,P(Y-|X+)\\ &= (\pi_{11}/\pi_{1\cdot})/(\pi_{12}/\pi_{1\cdot}) = \pi_{11}/\pi_{12}.\end{aligned}$$

Similarly,

$$\begin{aligned}\text{Odds}\,(Y+|X-) &= P(Y+|X-)\,/\,P(Y-|X-)\\ &= (\pi_{21}/\pi_{2\cdot})/(\pi_{22}/\pi_{2\cdot}) = \pi_{21}/\pi_{22}.\end{aligned}$$

The odds ratio is simply the ratio of these two odds values (see equation 13.3),

$$\omega = \frac{Odds(Y+|X+)}{Odds(Y+|X-)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{13.19}$$

The odds ratio is therefore estimable from a study in which prespecified numbers of units positive on X and negative on X are selected for a determination of their status on Y.

A similar analysis shows that the same is true for the odds ratio associated with a study in which prespecified numbers of units positive on Y and negative on Y are selected for a determination of their status on X. Because the latter two study designs correspond to prospective and retrospective sampling, it is clear that the odds ratio is estimable using data from either of these two designs,

as well as from a cross-sectional study. Other properties of the odds ratio are discussed in the final subsection of this section.

Unlike the meanings of the measures Δ and RR, the meaning of the odds ratio is not intuitively clear, perhaps because the odds value itself is not. Consider, for example, the quantity

$$\text{Odds}\,(Y+\,|X+) = \pi_{11}/\pi_{12}.$$

The odds value is thus the ratio, in the population of individuals who are positive on X, of the proportion who are positive on Y to the proportion who are negative. Given the identities

$$\text{P}(Y+\,|X+) = \text{Odds}\,(Y+\,|X+)\,/\,(1+\text{Odds}\,(Y+\,|X+))$$

and the complementary

$$\text{Odds}\,(Y+\,|X+) = P(Y+\,|X+)\,/\,(1-P(Y+\,|X+)),$$

it is clear that the information present in Odds $(Y+|X+)$ is identical to the information present in $P(Y+|X+)$. Nevertheless, the impact of "for every 100 subjects negative on Y (all of whom are positive on X), 100 × Odds $(Y+|X+)$ is the number expected to be positive on Y" may be very different from that of "out of every 100 subjects (all of whom are positive on X), 100 × $P\,(Y+|X+)$ is the number expected to be positive on Y." Once one develops an intuitive understanding of an odds value, it quickly becomes obvious that two odds values are naturally contrasted by means of their ratio—the odds ratio.

The functional form of the maximum likelihood estimator of $\hat{\omega}$ is the same for each study design (see equation 13.18). Remarkably, the functional form of the estimated standard error is also the same for each study design. For reasons similar to those given in section 13.3.2, it is customary to perform statistical analyses on $\ln(o)$, the natural logarithm of the sample odds ratio, rather than on o directly. Tosiya Sato, however, obtained a valid confidence interval for the population odds ratio without transforming to logarithms (1990). The large sample standard error of $\ln(o)$, due to Barnet Woolf (1955), is given by the equation

$$SE = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2} \qquad (13.20)$$

for each of the study designs considered here.

13.5.1 A Single Fourfold Table

Thus, for data arrayed as in table 13.3, equations 13.1 and 13.20 provide the point estimate of the odds ratio an the estimated standard error of its logarithm. To reduc the bias caused by one or more small cell frequencie (note that neither $\ln(o)$ nor SE is defined when a cell fre quency is equal to zero), authors in the past have recom mended that one add 0.5 to each cell frequency befor proceeding with the analysis (Gart and Zweifel 1967; For the frequencies in table 13.4, the values of the severa statistics calculated with (versus without) the adjustmen factor are $o = 2.25$ (versus 2.27), $\ln\,(o) = 0.81$ (versu 0.82), and $SE = 0.4462$ (versus 0.4380). The effect o adjustment is obviously minor in this example. More re cently, other authors have called into question the prac tice of addition of 0.5 to cell counts, suggesting insteac that other values may be preferred based on improved sta tistical properties (Sweeting, Sutton, and Lambert 2004). For example, Michael Sweeting and his colleagues sug gest using a function of the reciprocal of the sample size of the opposite treatment arm. That approach reduces the bias that may be associated with the use of a constant cor rection factor of 0.5. These authors also suggest Bayesian analyses as yet another alternative. In the context of a meta-analysis, the need to add any constant to cells with a zero cell count has also been called into question, as there are alternative methods that yield more accurate estimates with improved confidence intervals (Bradburn et al. 2007; Sweeting, Sutton, and Lambert 2004).

13.5.2 An Alternative Analysis: $(O-E)/V$

In their meta-analysis of randomized clinical trials of a class of drugs known as beta blockers, Salim Yusuf and his colleagues used an approach that has become popular in the health sciences. It may be applied to a single fourfold table as well as to several tables, such as in a study incorporating stratification (1985). Consider the generic table displayed in table 13.3, and select for analysis any of its four cells (the upper left hand cell is traditionally the one analyzed). The observed frequency in the selected cell is

$$O = n_{11}, \qquad (13.21)$$

the expected frequency in that cell, under the null hypothesis that X and Y are independent and conditional on the

marginal frequencies being held fixed, is

$$E = \frac{n_{1\cdot} n_{\cdot 1}}{n_{\cdot\cdot}}, \tag{13.22}$$

and the exact hypergeometric variance of n_{11}, under the same restrictions as above, is

$$V = \frac{n_{1\cdot} n_{2\cdot} n_{\cdot 1} n_{\cdot 2}}{n_{\cdot\cdot}^2 (n_{\cdot\cdot} - 1)}. \tag{13.23}$$

When ω is close to unity, a good estimator of ln (ω) is

$$L = \frac{O - E}{V}, \tag{13.24}$$

with an estimated standard error of

$$SE = 1/\sqrt{V}. \tag{13.25}$$

For the frequencies in table 13.4,

$$L = \frac{135 - 131.25}{4.1222} = 0.9097$$

[which corresponds to an estimated odds ratio of exp(0.9097) = 2.48], with an estimated standard error of

$$SE = 1/\sqrt{4.1222} = 0.49.$$

In this example, what has become known in the health sciences as the Peto method, or sometimes as the Yusuf-Peto method, or as the $(O - E)/V$ method, produces overestimates of the point estimate and its standard error.

The method is capable of producing underestimates as well as overestimates of the underlying parameter. For the frequencies from the two studies summarized in table 13.5, the respective values of L are 0.6892 (yielding 1.99 as the estimated odds ratio) and 0.7804 (yielding 2.18 as the estimated odds ratio). The latter estimate, coming from a study in which one of the two sets of marginal frequencies was uniform, comes closest of all (for the frequencies in tables 13.4 and 13.6) to the correct value of 2.25. Sander Greenland and Alberto Salvan showed that the $(O - E)/V$ method may yield biased results when there is serious imbalance on both margins (1990). Furthermore, they showed that the method may yield biased results when the underlying odds ratio is large, even when there is balance on both margins. These findings have been confirmed in subsequent work (Sweeting et al. 2004; Bradburn et al. 2006). Earlier, Nathan Mantel, Charles Brown, and David Byar (1977) and Joseph Fleiss, Bruce Levin, and Myunghee C. Paik (2003) pointed out flaws in a measure of association identical to $(O - E)/V$, the standardized difference.

Under many circumstances, given the simplicity and theoretical superiority of the more traditional methods described in the preceding and following subsections of section 5, there might be no compelling reason for the $(O - E)/V$ method to be used. However, more recent simulation studies have demonstrated that, when the margins of the table are balanced, and the odds ratio is not large, the $(O - E)/V$ method performs well, especially relative to use of a constant as a correction factor (Bradburn et al. 2006).

13.5.3 Inference in the Presence of Covariates

Covariates (that is, variables predictive of the outcome measure) are frequently incorporated into the analysis of categorical data. In randomized intervention studies, they are sometimes adjusted for to improve the precision with which key parameters are estimated and to increase the power of significance tests. In nonrandomized studies, they are adjusted for to eliminate the bias caused by confounding, that is, the presence of characteristics associated with both the exposure variable and the outcome variable. There are three major classes of methods for controlling for the effects of covariates: regression adjustment, stratification and matching. Each is considered.

13.5.3.1 Regression Analysis Let Y represent the binary response variable under study (Y=1 when the response is positive, $Y = 0$ when it is negative), let X represent the binary exposure variable ($X = 1$ when the study unit was exposed, $X = 0$ when it was not), and let $Z_1, \ldots, Z_p$ represent the values of p covariates. A popular statistical model for representing the conditional probability of a positive response as a function of X and the Z's is the *linear logistic regression model*,

$$\Pi = P(Y = 1 | X, Z_1, \ldots, Z_p) = \frac{1}{1 + e^{-(\alpha + \beta X + \gamma_1 Z_1 + \cdots + \gamma_p Z_p)}}. \tag{13.26}$$

This model is such that the logit or log odds of π,

$$\text{logit}(\pi) = \ln \frac{\pi}{1 - \pi}, \tag{13.27}$$

is linear in the independent variables:

$$\text{logit}(\pi) = \alpha + \beta X + \gamma_1 Z_1 + \cdots + \gamma_p Z_p \tag{13.28}$$

(for a thorough examination of logistic regression analysis, see Hosmer and Lemeshow 2000).

The coefficient β is the difference between the log odds for the population coded $X = 1$ and the log odds for the population coded $X = 0$, adjusting for the effects of the p covariates (that is, β is the logarithm of the adjusted odds ratio). The antilogarithm, $\exp(\beta)$, is therefore the odds ratio associating X with Y. Every major package of computer programs has a program for carrying out a linear logistic regression analysis using maximum likelihood methods. The program produces $\hat{\beta}$, the maximum likelihood estimate of the adjusted log odds ratio, and SE, the standard error of $\hat{\beta}$. Inferences about the odds ratio itself may be drawn in the usual ways.

13.5.3.2 Mantel-Haenszel Estimator When the p covariates are categorical (for example, sex, social class, and ethnicity), or are numerical but may be partitioned into a set of convenient and familiar categories (for example, age twenty to thirty-four, thirty-five to forty-nine, and fifty to sixty-four), their effects may be controlled for by stratification. Each subject is assigned to the stratum defined by his or her combination of categories on the p prespecified covariates, and, within each stratum, a fourfold table is constructed cross-classifying subjects by their values on X and Y. Within a study, the strata defined for an analysis will typically be based on the kinds of covariates just described (for example, sex, social class, and ethnicity). Of note, though, in performing a meta-analysis, the covariate of interest is the study, that is, the stratification is made by study and a fourfold table is constructed within each study. Because the detailed formulae for this stratified approach are presented in chapter 14 of this volume (formulas 14.14 and 14.15), only the principles are presented here.

The frequencies in table 13.6 are adapted from a comparative study reported by Peter Armitage (1971, 266). The associated relative frequencies are tabulated in table 13.7. Within each of the three strata, the success rate for the subjects in the experimental group is greater than the success rate for those in the control group. The formula for the summary odds ratio is attributable to Nathan Mantel and William Haenszel (1959). In effect, the stratified estimate is a weighted average of the within-stratum odds ratios, with weights proportional to the inverse of the variance of the stratum-specific odds ratio. (For a meta-analysis, again, the strata are defined by the studies, but the mathematical procedures are identical.) By the formula presented in chapter 14 (see 14.14 or 14.15), the value of the Mantel-Haenszel odds ratio is $o_{MH} = 7.31$.

James Robins, Norman Breslow, and Sander Greenland derived the non-null standard error of $\ln(o_{MH})$, which is also presented in chapter 14 as formula 14.16 (1986; see also Robins, Greenland, and Breslow 1986). For the data under analysis (in table 13.6), $o_{MH} = 7.3115$, as above, and $SE = 0.8365$. An approximate 95 percent confidence interval for $\ln(\omega)$ is: $\ln(7.3115) \pm 1.96 \times 0.8365$, or the interval from 0.35 to 3.63. The corresponding interval for ω extends from 1.42 to 37.7.

Table 13.6 Stratified Comparison of Two Treatments

		Outcome		
	Treatment	Success	Failure	Total
Stratum 1	Experimental	4	0	4
	Control	0	1	1
	Total	4	1	5
Stratum 2	Experimental	7	4	11
	Control	3	8	11
	Total	10	12	22
Stratum 3	Experimental	1	0	1
	Control	4	9	13
	Total	5	9	14

SOURCE: Authors' compilation.

Table 13.7 Quantities to Calculate the Pooled Log Odds Ratio

Stratum	L_s	SE_s	L_s/SE^2_s	$1/SE^2_s$
1	3.2958	2.2111	0.6741	0.2045
2	1.3981	0.8712	1.8421	1.3175
3	1.8458	1.7304	0.6164	0.3340
Total			3.1326	1.8560

SOURCE: Authors' compilation.

13.5.3.3 Combining Log Odds Ratios The pooling of log odds ratios across the S strata is a widely used alternative to the Mantel-Haenszel method of pooling. The pooled log odds ratio is explicitly a weighted average of the stratum-specific log odds ratios, with weights equal to the inverse of the variance of the stratum-specific estimate of the log odds ratio.

The reader will note, for the frequencies in table 13.6, that the additive adjustment factor of 0.5, or one of the other options defined by Michael Sweeting, Alex Sutton, and Paul Lambert (2004), must be applied for the log

odds ratios and their standard errors to be calculable in strata 1 and 3. Summary values appear in table 13.7. They yield a point estimate for the log odds ratio of $\bar{L}_{\cdot} = 1.6878$, and a point estimate for the odds ratio of $o_L = \exp(1.6878) = 5.41$. This estimate is appreciably less than the Mantel-Haenszel estimate of $o_{MH} = 7.31$. As all sample sizes increase, the two estimates converge to the same value. The estimated standard error of $\bar{L}_{\cdot}$ is equal to $SE = 0.7340$. The (grossly) approximate 95 percent confidence interval for ω derived from these values extends from 1.28 to 22.8.

13.5.3.4 Exact Stratified Method The exact method estimates the odds ratio by using permutation methods under the assumption of a conditional hypergeometric response within each stratum (Mehta, Patel, and Gray 1985). It is the stratified extension of Fisher's exact test in which the success probabilities are allowed to vary among strata. The theoretical basis is provided by John Gart (1970), but the usual implementation is provided by network algorithms such as those developed by Cyrus Mehta for the computer program StatXact, as described in the manual for that program (Cytel Software 2003).

13.5.3.5 Comparison of Methods When the cell frequencies within the S strata are all large, the Mantel-Haenszel estimate will be close in value to the estimate obtained by pooling the log odds ratios, and either method may be used (see Fleiss, Levin, and Paik 2003). As seen in the preceding subsection, close agreement may not be the case when some of the cell frequencies are small. In such a circumstance, the Mantel-Haenszel estimator is superior to the log odds ratio estimator (Hauck 1989). Note that there is no need to add 0.5 (or any other constant) to a cell frequency of zero in order for a stratum's frequencies to contribute to the Mantel-Haenszel estimator. In fact, it would be incorrect to add any constant to the cell frequencies when applying the Mantel-Haenszel method. The method must be applied to the cell count values directly.

Others have compared competing estimators of the odds ratio in the case of a stratified study (Agresti 1990, 235–37; Gart 1970; Kleinbaum, Kupper, and Morgenstern 1982, 350–51; McKinlay 1975; Sweeting et al. 2004; Bradburn et al. 2006). There is some consensus that the Mantel-Haenszel procedure is the method of choice, though logistic regression may also be applied. In situations where data are sparse, simulation studies have demonstrated reasonable performance by the logistic regression and Mantel-Haenszel methods (Sweeting et al. 2004; Bradburn et al. 2006). When there are approximately equal numbers of subjects in both treatment arms, and the intervention (or exposure) effects are not large, the stratified version of the one-step (Peto or Yusuf-Peto) method also performs surprisingly well, actually outperforming other methods, including exact methods, in situations in which the data are very sparse (Sweeting, Sutton, and Lambert 2004; Bradburn et al. 2006).

Table 13.8 Notation for Observed Frequencies Within Typical Matched Set

	Outcome Characteristic		
Group	Positive	Negative	Total
1	a_m	b_m	t_{m1}
2	c_m	d_m	t_{m2}
Total	$a_m + c_m$	$b_m + d_m$	$t_{m\cdot}$

SOURCE: Authors' compilation.

13.5.3.6 Control by Matching Matching, the third and final technique considered for controlling for the effects of covariates, calls for the creation of sets of subjects similar one to another on the covariates. Unlike stratification, in which the total number of strata, S, is specifiable in advance of collecting one's sample, with matching one generally cannot anticipate the number of matched sets one will end up with. Consider a study with $p = 2$ covariates, sex and age, and suppose that the investigators desire close matching on age - say ages no more than one year apart. If the first subject enrolled in the study is, for example, a twelve-year-old male, then the next male between eleven and thirteen, whether from group 1 or group 2, will be matched to the first subject. If one or more later subjects are also males within this age span, they should be added to the matched set that was already constituted, and should not form the bases for new sets. If M represents the total number of matched sets, and if the frequencies within set m are represented as in table 13.8, all that is required for the inclusion of this set in the analysis is that $t_{m1} \geq 1$ and $t_{m2} \geq 1$. The values $t_{m1} = t_{m2} = 1$ for all m correspond to pairwise matching, and the values $t_{m1} = 1$ and $t_{m2} = r$ for all m correspond to r:1 matching. Here, such convenient balance is not necessarily assumed. The reader is referred to David Kleinbaum, Lawrence Kupper, and Hal Morgenstern (1982, chap. 18) or to Kenneth Rothman and Sander Greenland (1998, chap. 10) for a discussion of matching as a strategy in study design.

Table 13.9 Results of Study of Association Between Use of Estrogens and Endometrial Cancer

a_m	b_m	c_m	d_m	t_{m*}	Matched Sets with Given Pattern
1	0	0	3	4	1
1	0	1	2	4	3
1	0	0	4	5	4
1	0	1	3	5	17
1	0	2	2	5	11
1	0	3	1	5	9
1	0	4	0	5	2
0	1	0	4	5	1
0	1	1	3	5	6
0	1	2	2	5	3
0	1	3	1	5	1
0	1	4	0	5	1

SOURCE: Authors' compilation.

The Mantel-Haenszel estimator of the odds ratio is given by the adaptation of equation 14.15 to the notation in table 13.8:

$$o_{\text{MH, matched}} = \frac{\sum a_m d_m / t_m}{\sum b_m c_m / t_m}. \tag{13.29}$$

Several studies have derived estimators of the standard error of $\ln(o_{\text{MH, matched}})$ in the case of matched sets (Breslow 1981; Connett et al. 1982; Fleiss 1984; Robins, Breslow, and Greenland 1986; Robins, Greenland, and Breslow 1986). Remarkably, formula 14.16, with the appropriate changes in notation, applies to matched sets as well as to stratified samples. The estimated standard error of $\ln(o_{\text{MH, matched}})$ is identical to formula 14.16 for meta-analyses (Robins, Breslow, and Greenland 1986).

The frequencies in table 13.9 have been analyzed by several scholars (Mack et al. 1976; Breslow and Day 1980, 176–82; Fleiss 1984). There were, for example, nine matched sets consisting of one exposed case, no unexposed cases, three exposed controls and one unexposed control. For the frequencies in the table, $o_{\text{MH, matched}} = 5.75$, $\ln(o_{\text{MH, matched}}) = 1.75$, and the estimated standard error of $\ln(o_{\text{MH, matched}})$ is $SE = 0.3780$. The resulting approximate 95 percent confidence interval for the value in the population of the adjusted odds ratio extends from $\exp(1.75 - 1.96 \times 0.3780) = \exp(1.0091) = 2.74$ to $\exp(1.75 + 1.96 \times 0.3780) = \exp(2.4909) = 12.07$.

An important special (and simple) case is that of matched pairs. This analysis simplifies to the following. Let D represent the number of matched pairs that are discordant in the direction of the case having been exposed and the control not, and let E represent the number of matched pairs that are discordant in the other direction. The Mantel-Haenszel estimator of the underlying odds ratio is simply $o_{\text{MH, matched}} = D/E$, and the estimated standard error of $\ln(o_{\text{MH, matched}})$ simplifies to

$$SE = \left(\frac{D+E}{DE}\right)^{1/2}. \tag{13.30}$$

13.5.4 Reasons for Analyzing the Odds Ratio

The odds ratio is the least intuitively understandable measure of association of those considered in this chapter, but it has a great many practical and theoretical advantages over its competitors. One important practical advantage, that it is estimable from data collected according to a number of different study designs, is demonstrated in section 13.5. A second, illustrated in section 13.5.3.1, is that the logarithm of the odds ratio is the key parameter in a linear logistic regression model, a widely used model for describing the effects of predictor variables on a binary response variable. Other important features of the odds ratio will now be pointed out.

If X represents group membership (coded 0 or 1), and if Y is a binary random variable obtained by dichotomizing a continuous random variable Y^* at the point y, then the odds ratio will be independent of y if Y^* has the cumulative logistic distribution,

$$P(Y^* \leq y^* \mid X) = \frac{1}{1 + e^{-(\alpha + \beta X + \gamma y^*)}}, \tag{13.31}$$

for all y^*. If, instead, a model specified by two cumulative normal distributions with the same variance is assumed, the odds ratio will be nearly independent of y (Edwards 1966; Fleiss 1970).

Many investigators automatically compare two proportions by taking their ratio, rr. When the proportions are small, the odds ratio provides an excellent approximation to the rate ratio. Assume, for example, that $P_1 = 0.09$ and $P_2 = 0.07$. The estimated rate ratio is

$$rr = P_1/P_2 = 0.09/0.07 = 1.286,$$

and the estimated odds ratio is nearly equal to it,

$$o = \frac{P_1(1-P_2)}{P_2(1-P_1)} = \frac{0.09 \times 0.93}{0.07 \times 0.91} = 1.314.$$

A practical consequence is that the odds ratio estimated from a retrospective study—the kind of study from which the rate ratio is not directly estimable—will, in the case of relatively rare events and in the absence of bias, provide an excellent approximation to the rate ratio.

Two theoretically important properties of the odds ratio pertain to underlying probability models. Consider, first, the sampling distribution of the frequencies in table 13.3, conditional on all marginal frequencies being held fixed at their observed values. When X and Y are independent, the sampling distribution of the obtained frequencies is given by the familiar hypergeometric distribution. When X and Y are not independent, however, the sampling distribution, referred to as the *noncentral hypergeometric distribution*, is more complicated. This exact conditional distribution depends on the underlying probabilities only through the odds ratio:

$$\Pr(n_{11}, n_{12}, n_{21}, n_{22} \mid n_{1\cdot}, n_{2\cdot}, n_{\cdot 1}, n_{\cdot 2}, \omega) = \frac{\frac{n_{1\cdot}!}{n_{11}!n_{12}!} \cdot \frac{n_{2\cdot}!}{n_{21}!n_{22}!}\omega^{n_{11}}}{\sum_x \frac{n_{1\cdot}!}{x!(n_{1\cdot}-x)!} \cdot \frac{n_{2\cdot}!}{(n_{\cdot 1}-x)!(n_{\cdot\cdot}-n_{1\cdot}-n_{\cdot 1}+x)!}\omega^x}, \quad (13.32)$$

where the summation is over all permissible integers in the upper left hand cell (Agresti 1990, 66–67).

The odds ratio also plays a key role in *loglinear models*. The simplest such model is for the fourfold table. When the row and column classifications are independent, the linear model for the natural logarithm of the expected frequency in row i and column j is

$$\ln\left(E(n_{ij})\right) = \mu + \alpha_i + \beta_j, \quad (13.33)$$

with the parameters subject to the constraints $\alpha_1 + \alpha_2 = \beta_1 + \beta_2 = 0$. In the general case, dependence is modeled by adding additional terms to the model:

$$\ln\left(E(n_{ij})\right) = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad (13.34)$$

with $\Sigma\gamma_{1j} = \Sigma\gamma_{2j} = \Sigma\gamma_{i1} = \Sigma\gamma_{i2} = 0$. Let $\gamma = \gamma_{11}$. Thanks to these constraints, $\gamma = -\gamma_{12} = -\gamma_{21} = \gamma_{22}$, and the association parameter γ is related to the odds ratio by the simple equation

$$\gamma = \tfrac{1}{4}\ln(\omega). \quad (13.35)$$

Odds ratios and their logarithms represent descriptors of association in more complicated loglinear models as well (Agresti 1990, chap. 5).

A final property of the odds ratio with great practical importance is that it can assume any value between 0 and ∞, no matter what the values of the marginal frequencies and no matter what the value of one of the two probabilities being compared. As we learned earlier, neither the simple difference nor the rate ratio nor the phi coefficient is generally capable of assuming values throughout its possible range. We saw for the rate ratio, for example, that the parameter value was restricted to the interval $0 \leq RR \leq 2.5$ when the probability π_2 was equal to 0.40. The odds ratio can assume any nonnegative value, however. Given that

$$\omega = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} = 1.5\frac{\pi_1}{(1-\pi_1)}$$

when $\pi_2 = 0.40$, $\omega = 0$ when $\pi_1 = 0$ and $\omega \to \infty$ when $\pi_1 \to 1$.

Consider, as another example, the frequencies in table 13.4. With the marginal frequencies held fixed, the simple difference may assume values only in the interval $-0.17 \leq D \leq 0.50$ (and not $-1 \leq D \leq 1$); the rate ratio with Positive as the untoward outcome may assume values only in the interval $0.83 \leq RR \leq 2.0$ (and not $0 \leq RR \leq \infty$); and the phi coefficient may assume values only in the interval $-0.22 \leq \varphi \leq 0.65$ (and not $-1 \leq \varphi \leq 1$); but the odds ratio may assume values throughout the interval $0 \leq \omega \leq \infty$. (To be sure, the rate ratio may assume values throughout the interval $0 \leq RR \leq \infty$ when Negative represents the untoward outcome.)

The final example brings us back to table 13.1. We saw earlier that there was heterogeneity in the association between exposure to asbestos and mortality from lung cancer, when association was measured by the simple difference between two rates:

$$D_{smokers} = 601.6 - 122.6 = 479.0 \text{ deaths per } 100{,}000 \text{ person years}$$

and

$$D_{nonsmokers} = 58.4 - 11.3 = 47.1 \text{ deaths per } 100{,}000 \text{ person years.}$$

This appearance of heterogeneous association essentially vanishes when association is measured by the odds ratio:

$$o_{\text{smokers}} = \frac{601.6 \times (100{,}000 - 122.6)}{122.6 \times (100{,}000 - 601.6)} = 4.93$$

and

$$o_{\text{nonsmokers}} = \frac{58.4 \times (100{,}000 - 11.3)}{11.3 \times (100{,}000 - 58.4)} = 5.17.$$

Whether one is considering smokers or nonsmokers, the odds that a person exposed to asbestos will die of lung cancer are approximately five times those that a person not so exposed will. In fairness, the rate ratio in this example also has a relatively constant value for smokers ($rr = 601.6/122.6 = 4.91$) and for nonsmokers ($rr = 58.4/11.3 = 5.17$), also illustrating the similarity of the values of the odds ratio and the rate ratio for this rare outcome.

As a result, we see that the odds ratio is not prone to the artifactual appearance of interaction across studies due to the influence on other measures of association or effect of varying marginal frequencies or to constraints on one or the other sample proportion. On the basis of this and all of its other positive features, the odds ratio is recommended as the measure of choice for measuring effect or association when the studies contributing to the research synthesis are summarized by fourfold tables.

13.5.5 Conversion to Other Measures

The discussion of the choice of an effect measure is complicated by the mentioned difficulty in the interpretation of the odds ratio. Despite its favorable mathematical properties, the odds ratio remains less than intuitive. In addition, though the odds ratio nicely reproduces the rate ratio (a more intuitive quantity) when the event of interest is uncommon (for example, for the data in table 13.1), it is well known to overestimate the rate ratio when the event of interest is more common (Deeks 1998; Altman, Deeks, and Sackett 1998). The report on the effect of race and sex on physician referrals by Kevin Schulman and his colleagues, illustrates this distortion, which generated some controversy in the medical literature (Schulman et al. 1999; see also, for example, Davidoff 1999). Presented with a series of clinical scenarios portrayed on video recordings involving trained actors, physicians were asked to make a decision as to whether to refer the given patient for further cardiac testing. Whites were referred by 91 percent of physicians, compared with African Americans by 85 percent. The odds ratio of 0.6, meaning the odds of referral for blacks are 60 percent of the odds of referral for whites, is a mathematically valid measure of the effect of race on referrals, but the temptation to think in terms of blacks being 60 percent as likely to be referred clearly needs to be resisted. There was a similar pattern for referral of women (by 85 percent of physicians) versus men (by 91 percent of physicians). On closer examination of the data, black women were the least likely to be referred, with all others being about as likely as each other to be referred.

It is possible, however, and potentially useful, to perform analyses on the odds ratio scale and then convert the results back into terms of either risk ratios or risk differences (Localio, Margolis, and Berlin 2007). Such an approach would be particularly useful when adjustment for multiple covariates in a logistic regression model is called for in an analysis. In the simplest case, the relationship between the two measures is captured by:

$$rr = \frac{\omega}{(1-\pi_1)(\omega\pi_1)}. \tag{13.36}$$

One recommended approach, using this relationship, takes the estimate of the proportion in the unexposed or untreated group (π_1) from the data, as the raw risk in the reference group, and takes the estimated odds ratio from a logistic regression model. The upper and lower confidence limits for the rate ratio are calculated using the same relationship, substituting the respective values from the confidence interval for the odds ratio. That approach, known as the method of substitution, has subsequently been criticized for its failure to take into account the variability in the baseline risk of π_1 (McNutt, Hafner, and Xue 1999; McNutt et al. 2003).

It is also possible to estimate adjusted relative risks using alternative models, for example, log-binomial models or Poisson regression models. Variations on these approaches, however, have been demonstrated to suffer from theoretical, as well as practical, limitations (Localio, Margolis, and Berlin 2007). As an alternative, Russell Localio and his colleagues proposed the use of logistic regression models, with adjustment for all relevant covariates, followed by conversion to the *RR* using the following relationship, which follows directly from equation 13.26 (2007):

$$rr = \frac{1+e^{-\alpha}}{1+e^{-\alpha-\beta}}. \tag{13.37}$$

Two challenges present themselves in such conversions. One is that equation 13.37 appears to ignore the values of the covariates. In fact, this is a deliberate simplification. From 13.26, it is clear that the value of the risk ratio will depend on both the value of the odds ratio and on the value of the baseline risk. In particular, it is clear that in the logistic model, the risk is specified as

depending on the values of the covariates. Thus, even though the value of the odds ratio is constant in the logistic model, the value of the relative rate should depend on the values of the covariates. In 13.37, it is assumed that the covariates have been centered, that is, that the mean values have been subtracted. The value of the relative rate calculated in 13.37 is then the value calculated for the average member of the sample under study. One can extend equation 13.37 to include specified values of the covariates. This would allow, for example, calculating relative risks separately for males and females from a model in which sex is specified as being associated with risk of the outcome.

The second challenge in the conversion from the odds ratio to the relative rate (or risk difference) is variance estimation. As noted, the bootstrap approach is a convenient one in situations where variance estimation is otherwise complicated (Carpenter and Bithell 2000; Efron and Tibshirani 1993; Davison and Hinkley 1997) and this method is suggested by Localio and his colleagues (2007). They also provide a method that doesn't depend on the iterative bootstrap approach and can be used when fitting a single logistic regression model.

The availability of the conversion approach described here makes it all the more compelling to fit models using odds ratios and then to convert the results onto the appropriate scale to facilitate interpretation by substantive or policy experts.

In some meta-analyses, some of the component studies may have reported continuous versions, and others dichotomized versions of the outcome variables. When the majority of meta-analytic studies have maintained the continuous nature of the outcome variable, and only a few others have dichotomized it, a practical strategy would be to transform the odds ratio to the *d* index (described in detail in chapter 12, this volume) with one of the conversion formulae proposed and discussed in a number of papers (for example, Chinn 2000; Haddock, Rindskopf, and Shadish 1998; Hasselblad and Hedges 1995; Sánchez-Meca, Marín-Martínez, and Salvador Chacón-Moscoso 2003; Ulrich and Wirtz 2004). The reverse strategy could be applied when the majority of the studies present odds ratios and only a few include *d* indices.

13.6 ACKNOWLEDGMENTS

Drs. Melissa Begg and Bruce Levin provided valuable criticisms of a version of this chapter written for the previous edition. The original work was supported in part by grant MH45763 from the National Institute of Mental Health. Arthur Fleiss kindly provided permission for Dr. Berlin to modify this chapter for the current edition.

13.7 REFERENCES

Agresti, Alan. 1990. *Categorical Data Analysis*. New York: John Wiley & Sons.

Altman, Douglas G., Jonathan J. Deeks, and David L. Sackett. 1998. "Odds Ratios Should Be Avoided when Events are Common." Letter. *British Medical Journal* 317(7168): 1318.

Armitage, Peter. 1971. *Statistical Methods in Medical Research*. New York: John Wiley & Sons.

Bishop, Yvonne M.M., Stephen E. Fienberg, and Paul W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass.: MIT Press.

Bradburn, Michael J., Jonathan J. Deeks, A. Russell Localio, Jesse A. Berlin. 2007. "Much Ado About Nothing: A Comparison of the Performance of Meta-Analytical Methods with Rare Events." *Statistics in Medicine* 26(1): 53–77.

Breslow, Norman E. 1981. "Odds Ratio Estimators When the Data are Sparse." *Biometrika* 68(1): 73–84.

Breslow, Norman E., and Nicholas E. Day. 1980. *Statistical Methods in Cancer Research*, vol. 1, *The Analysis of Case-Control Studies*. Lyons: International Agency for Research on Cancer.

Carpenter, James, and John Bithell. 2000. "Bootstrap Confidence Intervals: When, Which, What? A Practice Guide for Medical Statisticians." *Statistics in Medicine* 19(9): 1141–64.

Chinn, Susan. 2000. "A Simple Method for Converting an Odds Ratio to Effect Size for Use in Meta-Analysis." *Statistics in Medicine* 19(22): 3127–31.

Carroll, John B. 1961. "The Nature of the Data, or How to Choose a Correlation Coefficient." *Psychometrika* 26(4): 347–72.

Connett, John, Ejigou, Ayenew, McHugh, Richard, and Breslow, Norman. 1982. "The Precision of the Mantel-Haenszel Estimator in Case-Control Studies with Multiple Matching." *American Journal of Epidemiology* 116(6): 875–77.

Cytel Software Corporation. 2003. *StatXact*: Cambridge, Mass.

Davidoff, Frank. 1999. "Race. Sex, and Physicians' Referral for Cardiac Catheterization." Letter. *New England Journal of Medicine* 341(4): 285–86.

Davison, Anthony C., and David V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Deeks, Jonathan J. 1998. "When Can Odds Ratio Mislead?" Letter. *British Medical Journal* 317(7166): 1155.

———. 2002. "Issues in the Selection of a Summary Statistic for Meta-Analysis of Clinical Trials with Binary Outcomes." *Statistics in Medicine* 21(11): 1575–1600.

Edwards, John H. 1966. "Some Taxonomic Implications of a Curious Feature of the Bivariate Normal Surface." *British Journal of Preventive and Social Medicine* 20(1): 42–43.

Efron, Bradley, and Rob J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Engels, Eric A., Christopher H. Schmid, Norma Terrin, Ingram Olkin, and Joseph Lau. 2000. "Heterogeneity and Statistical Significance in Meta-Analysis: An Empirical Study of 125 Meta-Analyses." *Statistics in Medicine* 19(13): 1707–28.

Fleiss, Joseph L. 1970. "On the Asserted Invariance of the Odds Ratio." *British Journal of Preventive and Social Medicine* 24(1): 45–46.

———. 1984. "The Mantel-Haenszel Estimator in Case-Control Studies with Varying Numbers of Controls Matched to Each Case." *American Journal of Epidemiology* 120(1): 1–3.

Fleiss, Joseph L., and Mark Davies. 1982. "Jackknifing Functions of Multinomial Frequencies, with an Application to a Measure of Concordance." *American Journal of Epidemiology* 115(6): 841–45.

Fleiss, Joseph L., Bruce Levin, and Myunghee C. Paik. 2003. *Statistical Methods for Rates and Proportions*, 3rd ed. New York: John Wiley & Sons.

Gart, John J., and James R. Zweifel. 1967. "On the Bias of Various Estimators of the Logit and its Variance, with Application to Quantal Bioassay." *Biometrika* 54(1): 181–87.

Gart, John J. 1970. "Point and Interval Estimation of the Common Odds Ratio in the Combination of 2×2 Tables with Fixed Marginals." *Biometrika* 57(3): 471–75.

Greenland, Sander, and Alberto Salvan. 1990. "Bias in the One-Step Method for Pooling Study Results." *Statistics in Medicine* 9(3): 247–52.

Haddock, C. Keith, David Rindskopf, and William R. Shadish. 1998. "Using Odds Ratios as Effect Sizes for Meta-Analysis of Dichotomous Data: A Primer on Methods and Issues." *Psychological Methods* 3(3): 339–53.

Hammond, E. Cuyler, Irving J. Selikoff, and Herbert Seidman. 1979. "Asbestos Exposure, Cigarette Smoking and Death Rates." *Annals of the New York Academy of Sciences* 330: 473–90.

Hasselblad, Victor, and Larry V. Hedges. 1995. "Meta-Analysis of Screening and Diagnostic Tests." *Psychological Bulletin* 117(10: 167–78.

Hauck, Walter W. 1989. "Odds Ratio Inference from Stratified Samples." *Communications in Statistics* 18A(2): 767–800.

Hosmer, David W., and Stanley Lemeshow. 2004. *Applied Logistic Regression*, 2nd ed. New York: John Wiley & Sons.

Hunter, John E., and Frank L. Schmidt. 1990. "Dichotomization of Continuous Variables: The Implications for Meta-Analysis." *Journal of Applied Psychology* 75: 334–49.

Kleinbaum, David G., Lawrence L. Kupper, and Hal Morgenstern. 1982. *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, Calif.: Lifetime Learning Publications.

Localio, A. Russell, D. J. Margolis, Jesse A. Berlin. 2007. "Relative Risks and Confidence Intervals Were Easily Compared Indirectly from Multivariable Logistic Regression." *Journal of Clinical Epidemiology* 60(9): 74–82.

Mack, Thomas M., Malcolm C. Pike, Brian E. Henderson, R. I. Pfeffer, V. R. Gerkins, B. S. Arthur, and S. E. Brown. 1976. "Estrogens and Endometrial Cancer in a Retirement Community." *New England Journal of Medicine* 294(23): 1262–67.

Mantel, Nathan, Charles Brown, and David P. Byar. 1977. "Tests for Homogeneity of Effect in an Epidemiologic Investigation." *American Journal of Epidemiology* 106(2): 125–29.

Mantel, Nathan, and William Haenszel. 1959. "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease." *Journal of the National Cancer Institute* 22(4): 719–48.

McKinlay, Sonja M. 1975. "The Effect of Bias on Estimators of Relative Risk for Pair-Matched and Stratified Samples." *Journal of the American Statistical Association* 70(352): 859–64.

McNutt Louise A., Jean P. Hafner, and Xue Xiaonan. 1999. "Correcting the Odds Ratio in Cohort Studies of Common Outcomes." Letter. *Journal of the American Medical Association* 282(6): 529.

McNutt, Louise A., Chuntao Wu, Xiaonan Xue, and Jean P. Hafner. 2003. "Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes." *American Journal of Epidemiology* 157(10): 940–43.

Mehta, Cyrus, Nitin R. Patel, and Robert Gray. 1985. "On Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2×2 Contingency Tables." *Journal of the American Statistical Association* 80(392): 969–73.

Quenouille, M. H. 1956. "Notes on Bias in Estimation." *Biometrika* 43(3–4): 353–60.

Robins, James, Norman Breslow, and Sander Greenland. 1986. "Estimators of the Mantel-Haenszel Variance Consistent in

Data and Large-Strata Limiting Models." *Bio-*
): 11–23.
s., Sander Greenland, and Norman E. Breslow.
neral Estimator for the Variance of the Mantel-
lds Ratio." *American Journal of Epidemiology*
-23.
nneth J., and Sander Greenland. 1998. *Modern*
y, 2nd ed. Philadelphia, Pa.: Lippincott Williams
.
a, Julio, Fungencio Marín-Martínez, and Sal-
in-Moscoso. 2003. "Effect-Size Indices for Di-
Outcomes in Meta-Analysis." *Psychological*
): 448–67.
1990. "Confidence Limits for the Common
Based on the Asymptotic Distribution of the
nszel Estimator." *Biometrics* 46: 71–80.
vin A., Jesse A. Berlin, William Harless, Jon F.
rl Sistrunk, Benard J. Gersh et al. 1999. "The
ıce and Sex on Physicians' Recommendations
for Cardiac Catheterization." *New England Journal of Medicine* 340(8): 618–26.
Sweeting, Michael J., Alex J. Sutton, and Paul C. Lambert. 2004. "What to Add to Nothing? Use and Avoidance of Continuity Corrections in Meta-Analysis of Sparse Data." *Statistics in Medicine* 23(9): 1351–75.
Tukey, John W. 1958. "Bias and Confidence in Not-Quite-Large Samples." *Annals of Mathematical Statistics* 29: 614.
Ulrich, Rolf, and Markus Wirtz. 2004. "On the Correlation of a Naturally and an Artificially Dichotomized Variable." *British Journal of Mathematical and Statistical Psychology* 57(2): 235–51.
Woolf, Barnet. 1955. "On Estimating the Relation Between Blood Group and Disease." *Annals of Human Genetics* 19(4): 251–53.
Yusuf, Salim, Richard Peto, John Lewis, Rory Collins, and Peter Sleight. 1985. "Beta Blockade During and After Myocardial Infarction: An Overview of the Randomized Trials." *Progress in Cardiovascular Diseases* 27(5): 335–71.