Shore, R. E., Pasternack, B. S., and Curnen, M. G: Relating influenza epidemics to childhood leukemia in tumor registries without a defined population base: A critique with suggestions for improved methods. *Am. J. Epidemiol.* 1976;103:527–535.

Woolf, B. On estimating the relation between blood group and disease. *Ann. Hum. Genet.* 1955;19:251–253.

# 12. STRATIFIED ANALYSIS

Two different analytic concerns motivate the division of data into strata: one is the need to evaluate and remove *confounding;* the other is to evaluate and describe *effect modification.* Because stratification is the preferred means of dealing with both of these analytic issues, the beginning student is apt to become bewildered in the attempt to distinguish between the aims and procedures involved in considering these two aspects of epidemiologic data analysis.

Effect modification refers to a change in the magnitude of an effect measure according to the value of some third variable (after exposure and disease), which is called an *effect modifier.* Effect modification differs from confounding in several ways. The most central difference is that, whereas confounding is a bias that the investigator hopes to prevent or, if necessary, to remove from the data, effect modification is an elaborated description of the effect itself. Effect modification is thus a finding to be reported rather than a bias to be avoided. Epidemiologic analysis is generally aimed at eliminating confounding and discovering and describing effect modification.

It is a useful contrast to think of confounding as a nuisance that may or may not be present depending on the study design. Of course, confounding originates from the interrelation of the confounding factors and study variables in the source population from which the study subjects are selected. Nevertheless, restriction in subject selection, for example, can prevent a variable from becoming a confounding factor in a situation in which it otherwise would be confounding. Effect modification, on the other hand, rather than being a nuisance the presence of which depends on the specifics of the study design, is a natural phenomenon that exists independently of the study. It is a phenomenon that the study is intended to divulge and describe if at all possible. Whereas the existence of confounding with respect to a given factor depends on the design of a study, effect modification has a conceptual constancy that transcends the study design.

Although effect modification is a constant of nature, in its most general sense it cannot correspond to any biologic property because there is one aspect of the concept that is not absolute: Effect modification in its most general context includes modification of an effect without specifying which effect measure is modified. Since there are two effect measures, the difference and ratio measures, that are commonly used in epidemiology as well as others that are used less often, the concept of effect modification without further specification is too ambiguous to be useful as a description of nature.

In Figure 12-1, age can be considered a modifier of the effect of exposure, since the incidence rate difference between exposed and unexposed increases with increasing age. On the other hand, the ratio of incidence among exposed to incidence among unexposed is constant over age. Thus, age modifies the effect of exposure with regard to the difference measure

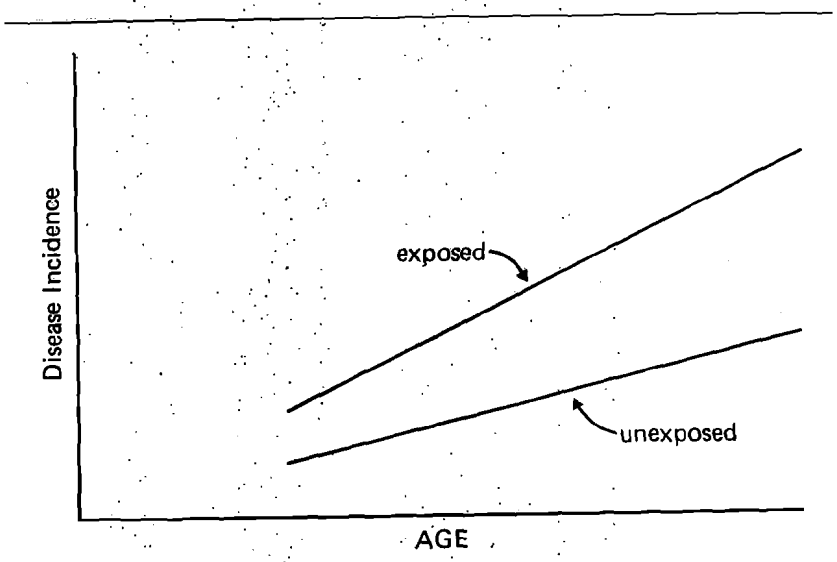


*Fig. 12-1. Disease incidence by exposure and age indicating a constant ratio of incidence with age.*

of effect but not with regard to the ratio measure. The opposite situation is described in Figure 12-2: The difference in incidence rate between exposed and unexposed is constant over age, but the ratio of incidence among exposed to incidence among unexposed declines with age. These diagrams illustrate why effect modification should be described only in relation to a specific effect measure. If effect modification is absent with regard to either the difference measure or the ratio measure, it will be present with regard to the other measure unless the disease rate among the unexposed is unassociated with the potential effect modifier.

This chapter presents the fundamental analytic strategies for dealing with confounding and effect modification in a stratified analysis. The biologic and public health interpretations of effect modification are considered in Chapter 15.

## EVALUATION AND CONTROL OF CONFOUNDING

Confounding is a distortion in an effect measure that results from the effect of another variable that is associated with the exposure under study. In Chapter 7, confounding was defined, and the general characteristics of confounding factors were discussed. To review, a confounding factor must

1. Be a risk factor for the disease among the nonexposed.
2. Be associated with the exposure variable in the population from which the cases derive.



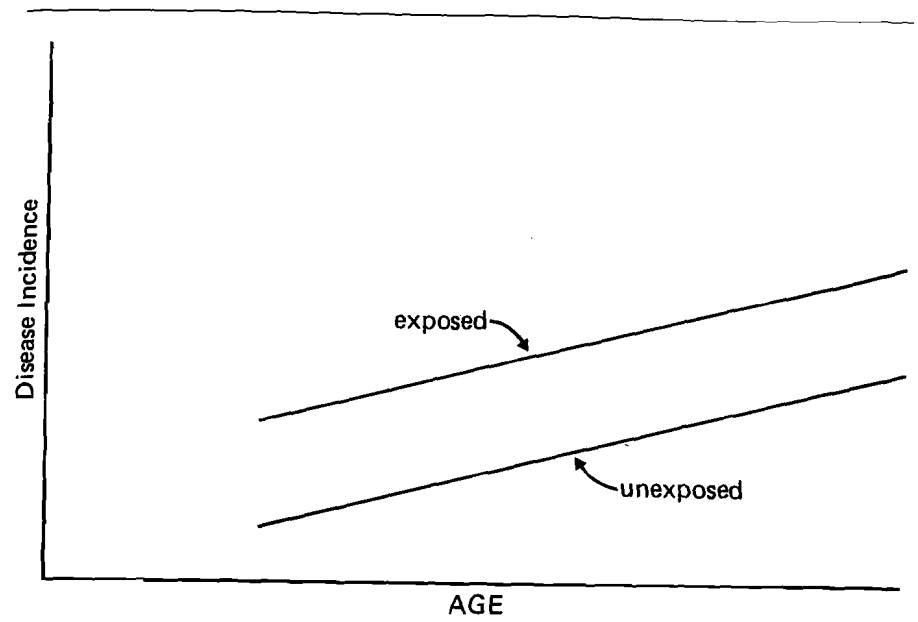


*Fig. 12-2. Disease incidence by exposure and age indicating a constant difference of incidence with age.*

3. Not be an intermediate step in the causal path between the exposure and the disease.

The case-control data in Example 12-1 demonstrate confounding by age. If the effect of oral contraceptives on the risk of myocardial infarction is estimated from the crude data, the odds ratio estimate is 2.2. If the data are divided, however, into two age categories, the odds ratio estimate in each category is 2.8, which corresponds to a 50 percent greater effect than the estimate of 2.2 ([2.8 − 1]/[2.2 − 1] = 1.5).

It is clear that the variable "age" in Example 12-1 meets the criteria for a confounding factor. First, age is a risk factor for myocardial infarction among the nonexposed, that is, nonusers of oral contraceptives. We know in general that age is a strong risk factor for myocardial infarction; more directly, we can see that among the subjects in this particular study who are nonusers the proportion of subjects who are classified as cases is greater in the age category 40 to 44 (88/183 = 0.48) than in the age category < 40 (26/85 = 0.31). These proportions do not represent any meaningful epidemiologic measure; because these are case-control data, these proportions reflect the overall case-control ratio arbitrarily chosen by the investigators. The proportions might be described as the "prevalence of disease among nonexposed study subjects," which, given the case-control

*Example 12-1: Case-control data describing the effect of oral contraceptive use on risk of myocardial infarction, with confounding by age [Mann et al., 1968]*

| | Age < 40 | | Age 40–44 | | Totals | |
|---|---|---|---|---|---|---|
| | User | Nonuser | User | Nonuser | User | Nonuser |
| Myocardial infarction cases | 21 | 26 | 18 | 88 | 39 | 114 |
| Controls | 17 | 59 | 7 | 95 | 24 | 154 |
| Odds ratio estimate | 2.8 | | 2.8 | | 2.2 | |

design, are not meaningful prevalences. Nevertheless, for age to be confounding, these proportions must vary by age.

In addition, for age to be confounding, it must be associated with oral contraceptive use among the source population that gave rise to the cases. Looking among the controls, who are sampled from that source population, we note that the proportion of oral contraceptive users is much greater (17/76 = 0.22) among younger controls than among older controls (7/102 = 0.07), indicating that this condition has been fulfilled.

Since age cannot be construed as a causal link between oral contraceptive use and myocardial infarction, it meets the criteria for a confounding factor in these data. There is a more direct method, however, by which confounding can be assessed. It is possible to evaluate the magnitude of confounding by comparing the estimate of effect derived from the crude data with the estimate derived from the stratified data (provided that the potential confounder is judged not to be a link in the causal path). Ignoring whatever residual age confounding there might be within these two age categories, we can say that the estimate of the incidence rate ratio of oral contraceptive use on the risk of myocardial infarction unconfounded by age is 2.8, since the estimate is 2.8 in each of the two age strata. The estimate based on the crude data, however, is 2.2. If these estimates were identical, the data would indicate no confounding. The magnitude of confounding in the data is estimated by the degree of discrepancy between the crude and unconfounded estimates.

Some investigators have attempted to assess confounding through statistical tests of significance. For example, in a clinical trial, the age distribution in the treatment and comparison groups may be compared by a t test; if the test statistic is "significant," then age would be judged potentially confounding, whereas lack of "significance" would imply that age is not confounding. There is probably no more grievous routine misuse of statistical testing than in this common circumstance. Since confounding is a bias that depends on the magnitude of two component associations, confounder with exposure and confounder with disease, proper assessment



of confounding must be based on the magnitude of those associations. Statistical "significance" testing reflects a mixture of both magnitude of association and number of observations and therefore does not correspond to an assessment of magnitude of association alone. A large number of observations will produce statistical "significance" in situations in which the magnitude of one of the component associations of a potentially confounding factor is puny and would preclude any substantial confounding. Conversely, strong associations that produce serious confounding might be judged "not significant" if the number of observations is sparse. Confounding should therefore never be assessed by statistical tests.

Although it is possible to obtain a general appreciation for the presence or absence of confounding in data by examining whether the potentially confounding factor is associated with disease among nonexposed and with exposure among nondiseased, the magnitude of the confounding in the data is difficult to assess in this way because the confounding represents a function of both of these component associations. Furthermore, when several factors are simultaneously confounding, the component associations should ideally be examined conditional on the other confounding factors, thereby complicating the problem. The preferred method of assessing confounding is direct comparison of the crude and unconfounded estimates of effect. (An exception would be the unusual situation in which prior knowledge outweighs the evidence in the data about confounding, as discussed in Chapter 7, or when the potential confounder is judged to be a link in the causal pathway.) This comparison clearly and unambiguously reveals the magnitude of the confounding, which the investigator can then take into account in further analyses or reporting of results. Furthermore, this comparison can be made while controlling for other factors if necessary.

### *Point Estimation of a Uniform Effect*

In Example 12-1, the point estimate of the incidence rate ratio was 2.8 in each of the two age strata, so there is no difficulty in inferring that an overall estimate of effect unconfounded by age should be 2.8. Even if the parameter value of the effect is identical across strata, however, it is reasonable to expect that estimates of the effect will vary among strata because of random error. Typically, then, the investigator must derive an overall estimate of effect from stratified data by taking a weighted average of the stratum-specific effect estimates. If the parameter value of the effect is assumed to be uniform—that is, constant over the range of the confounding variable—then each stratum provides a separate estimate of the same parameter value, the stratum-specific estimates varying only randomly. In weighting the estimates to get an average, it is desirable to assign greater weight to those stratum-specific estimates with smaller random variability and vice versa. Theoretically, the optimum procedure for reduc-

ing the variance in the overall weighted average is to assign weights to the stratum-specific values that are inversely proportional to the variance of each stratum-specific estimate:

$$\text{Overall effect estimate} = \frac{\sum_i [w_i \cdot (\text{effect estimate in stratum } i)]}{\sum_i w_i}$$

in which

$$w_i = \frac{1}{(\text{variance of effect estimate in stratum } i)}$$

This method of point estimation, in which the individual strata are weighted to enhance the precision of the overall estimate, is known as *pooling*. (The reader should note that the term *pooled* is sometimes used by statisticians to mean "crude.")

Pooling can be performed by calculating the weights for averaging the stratum-specific effect estimates directly from the estimated variance of the effect calculated from the data in each stratum separately; this method requires enough information within each stratum to get reasonable variance estimates. Another approach, the method of maximum likelihood, involves the solution of a set of equations and produces the pooled estimate without explicitly determining stratum-specific weights. The maximum likelihood approach can be thought of as a weighting process in which the weights are implicit in the equations that yield the point estimate. This description is not literally correct, since, for example, no weighting scheme would work if one of the stratum-specific estimates were infinity, whereas the maximum likelihood approach produces an appropriate finite result in this situation; indeed, the ability to average erratic stratum-specific estimates efficiently when data are relatively sparse is one of the main advantages of the maximum likelihood approach. Another set of pooled estimators, the Mantel-Haenszel estimators, have explicit weights that are built into the formulas; the Mantel-Haenszel estimators are the easiest to calculate, and, considering that their statistical properties are nearly as good as the difficult-to-calculate maximum likelihood estimators, they are often the method of choice.

In the following sections, the above three approaches to pooling are presented: the direct approach, using explicit weights inversely proportional to stratum-specific variance estimates; the maximum likelihood approach; and the Mantel-Haenszel approach. The specific formulas used for determining the pooled estimators depend on the type of data supplied and the effect measure being estimated.



## POOLING WITH INVERSE VARIANCES (DIRECT POOLING)

*Directly Pooled Point Estimation of a Uniform Effect with Person-Time Data.* INCIDENCE RATE DIFFERENCE. Using the notation in Table 12-1, the variance for the estimate of incidence rate difference (IRD) from a single stratum in a stratified analysis is approximately

$$\text{Var}(\hat{\text{IRD}}_i) \doteq \frac{a_i}{N_{1i}^2} + \frac{b_i}{N_{0i}^2} \qquad [12\text{-}1]$$

(see formula 11-15). Therefore, a pooled estimator for an IRD that is constant over strata can be obtained from

$$\hat{\text{IRD}} = \frac{\sum_i w_i \hat{\text{IRD}}_i}{\sum_i w_i} \qquad [12\text{-}2]$$

in which

$$w_i = \frac{1}{\text{Var}(\hat{\text{IRD}}_i)} = \frac{N_{0i}^2 \cdot N_{1i}^2}{a_i N_{0i}^2 + b_i N_{1i}^2} \qquad [12\text{-}3]$$

and

$$\hat{\text{IRD}}_i = \frac{a_i}{N_{1i}} - \frac{b_i}{N_{0i}}$$

as with crude data.

For an example of pooling used to estimate incidence rate difference, consider the data in Example 12-2. Stratum-specific estimates of incidence rate difference for these data, the corresponding stratum-specific variances, and the weights for pooling are given in Table 12-2, based on formulas 12-1 through 12-3. The pooled estimate of rate difference is obtained by taking the sum of the product of each stratum-specific estimate with the weight and dividing by the sum of the weights. The result is $5.95 \times 10^{-4}\text{yr}^{-1}$, which is expectedly close to the estimated incidence rate

*Table 12-1. Notation for incidence rate data with person-time denominators in stratum i of a stratified analysis*

| | Exposed | Unexposed | Total |
|---|---|---|---|
| Cases | $a_i$ | $b_i$ | $M_{1i}$ |
| Person-time | $N_{1i}$ | $N_{0i}$ | $T_i$ |

*Example 12-2. Age-specific coronary disease deaths among British male doctors by cigarette smoking [Doll and Hill, 1966]*

| | Smokers | | Nonsmokers | |
|---|---|---|---|---|
| Age | Deaths | Person-years | Deaths | Person-years |
| 35–44 | 32 | 52,407 | 2 | 18,790 |
| 45–54 | 104 | 43,248 | 12 | 10,673 |
| 55–64 | 206 | 28,612 | 28 | 5,710 |
| 65–74 | 186 | 12,663 | 28 | 2,585 |
| 75–84 | 102 | 5,317 | 31 | 1,462 |
| Total | 630 | 142,247 | 101 | 39,220 |

*Table 12-2. Stratum-specific estimates of incidence rate difference, with variances and weights for pooling for the data in example 12-2*

| Age | Estimate of incidence rate difference ($\times 10^4$yr) | Variance ($\times 10^8$yr$^2$) | Weight ($\times 10^{-6}$yr$^{-2}$) |
|---|---|---|---|
| 35–44 | 5.04 | 1.73 | 57.8 |
| 45–54 | 12.8 | 16.1 | 6.21 |
| 55–64 | 23.0 | 111 | 0.90 |
| 65–74 | 38.6 | 535 | 0.19 |
| 75–84 | −20.2 | 1810 | 0.06 |

difference for stratum 1, the stratum with the smallest variance and the largest weight. The crude incidence rate difference is

$$\frac{630}{142{,}247\ \text{yr}} - \frac{101}{39{,}220\ \text{yr}} = 18.5 \times 10^{-4}\text{yr}^{-1}$$

which differs considerably from the pooled estimate of $5.95 \times 10^{-4}\text{yr}^{-1}$, indicating a substantial amount of age confounding in these data.

INCIDENCE RATE RATIO. For ratio estimators, pooling is performed after a logarithmic transformation of the estimates, which stabilizes the variances. The weights are the inverses of the variances of the logarithmically transformed stratum-specific estimates of incidence rate ratio (IRR). An approximate formula for this variance is

$$\text{Var}[\ln(\widehat{\text{IRR}}_i)] \doteq 1/a_i + 1/b_i \qquad [12\text{-}4]$$

and therefore the weight for pooling is

$$w_i = \frac{a_i b_i}{a_i + b_i}$$



*Example 12-3. Mortality by sex and age for patients with trigeminal neuralgia [Rothman and Monson, 1973]*

| | Age < 65 | | Age 65+ | | Totals | |
|---|---|---|---|---|---|---|
| | Males | Females | Males | Females | Males | Females |
| Deaths | 14 | 10 | 76 | 121 | 90 | 131 |
| Person-years | 1516 | 1701 | 949 | 2245 | 2465 | 3946 |

*Table 12-3. Stratum-specific estimates of incidence rate ratio, with logarithmic transformations, variances, and weights for pooling for the data in example 12-3*

| Age | Estimate of incidence rate ratio | Logarithm of $\widehat{\text{IRR}}$ | Variance of $\ln(\widehat{\text{IRR}})$ | Weight |
|---|---|---|---|---|
| < 65 | 1.57 | 0.45 | 0.17 | 5.83 |
| 65+ | 1.49 | 0.40 | 0.021 | 46.7 |

and the pooled estimator, after reversing the logarithmic transformation, is

$$\widehat{\text{IRR}} = \exp\left[\frac{\sum_i w_i \ln(\widehat{\text{IRR}}_i)}{\sum_i w_i}\right] \qquad [12\text{-}5]$$

where

$$\widehat{\text{IRR}}_i = \frac{a_i/N_{1i}}{b_i/N_{0i}}$$

as with crude data.

The application of formula 12-5 is demonstrated using the data of Example 12-3 from a survival study of patients with trigeminal neuralgia. The male-to-female ratio of mortality rates from the crude data is (90/2465 yr)/(131/3946 yr) = 1.10. A pooled estimate, controlling for age using the two age categories in Example 12-3, is obtained using the calculations given in Table 12-3. The pooled estimate is obtained by taking the sum of the weight in each stratum multiplied by the logarithm of the stratum-specific point estimate, dividing that sum by the sum of the weights, and then taking the antilogarithm of the result to reverse the transformation, giving 1.50 for the data in Example 12-3. The large discrepancy between this unconfounded estimate and the crude estimate of 1.10 indicates that the crude result was substantially biased by age confounding.

*ctly Pooled Point Estimation of a Uniform Effect with Cumulative In-*
*nce Data.* CUMULATIVE INCIDENCE DIFFERENCE: The notation for stratified
2 tables is given in Table 12-4. The approximate variance for the risk
erence (RD) in stratum i is

$$\text{Var}(\hat{RD}_i) = \frac{a_i(N_{1i} - a_i)}{N_{1i}^3} + \frac{b_i(N_{0i} - b_i)}{N_{0i}^3} \qquad [12\text{-}6]$$

ere

$$\hat{RD}_i = \frac{a_i}{N_{1i}} - \frac{b_i}{N_{0i}}$$

weight for pooling stratum-specific estimates of risk difference is the
erse of the variance:

$$w_i = \frac{N_{1i}^3 N_{0i}^3}{N_{0i}^3 a_i(N_{1i} - a_i) + N_{1i}^3 b_i(N_{0i} - b_i)} \qquad [12\text{-}7]$$

ooled estimator for risk difference is

$$\hat{RD} = \frac{\sum_i w_i \hat{RD}_i}{\sum_i w_i} \qquad [12\text{-}8]$$

e cumulative incidence data in Example 12-4 indicate a crude risk dif-
ence of 30/204 − 21/205 = 0.045, but this is confounded by age, as is
own in Table 12-5, in which the age-specific risk differences are each
en to be in the vicinity of 0.035.
The unconfounded pooled estimate of cumulative incidence difference
obtained from formulas 12-7 and 12-8; as shown in Table 12-5, the older
e category gives an estimate of cumulative incidence difference that has
nuch greater variance than that from the younger category, and therefore
nuch greater weight is assigned to the younger age category. The pooled
timate is 0.034, which reflects the greater weight assigned to the younger
tegory.

*ble 12-4. Notation for 2 × 2 tables in stratum i of a stratified analysis*

| | Exposed | Unexposed | Total |
|---|---|---|---|
| ases | $a_i$ | $b_i$ | $M_{1i}$ |
| oncases | $c_i$ | $d_i$ | $M_{0i}$ |
| Total | $N_{1i}$ | $N_{0i}$ | $T_i$ |

*Example 12-4. Age-specific comparison of deaths from all causes for tolbutamide and placebo treatment groups, University Group Diabetes Program [1970]*

| | Age < 55 | | Age ⩾ 55 | | Totals | |
|---|---|---|---|---|---|---|
| | Tolbutamide | Placebo | Tolbutamide | Placebo | Tolbutamide | Placebo |
| Dead | 8 | 5 | 22 | 16 | 30 | 21 |
| Surviving | 98 | 115 | 76 | 69 | 174 | 184 |
| Totals | 106 | 120 | 98 | 85 | 204 | 205 |

*Table 12-5. Stratum-specific estimates of cumulative incidence difference, with variances and weights for pooling for the data in example 12-4*

| Age | Estimate of cumulative incidence difference | Variance of cumulative incidence difference | Weight |
|---|---|---|---|
| < 55 | 0.034 | 0.00099 | 1009 |
| ≥ 55 | 0.036 | 0.00357 | 280 |

CUMULATIVE INCIDENCE RATIO. The ratio estimator is obtained, as before, using a logarithmic transformation. The approximate variance of the logarithm of the stratum-specific cumulative incidence ratio (RR) is

$$\text{Var}[\ln(\hat{RR}_i)] \doteq \frac{c_i}{a_i N_{1i}} + \frac{d_i}{b_i N_{0i}} \qquad [12\text{-}9]$$

The weight for pooling is equal to the inverse of this variance:

$$w_i = \frac{a_i b_i N_{1i} N_{0i}}{a_i d_i N_{1i} + b_i c_i N_{0i}} \qquad [12\text{-}10]$$

and the pooled estimator is

$$\hat{RR} = \exp\left[\frac{\sum_i w_i \ln(\hat{RR}_i)}{\sum_i w_i}\right] \qquad [12\text{-}11]$$

where

$$\hat{RR}_i = \frac{a_i/N_{1i}}{b_i/N_{0i}}$$

Let us consider Example 12-4 again, this time for risk ratio estimation. The crude estimate is (30/204)/(21/205) = 1.44. From a visual inspection it is difficult to assess the extent to which confounding is present, since the two stratum-specific estimates of cumulative incidence ratio bracket the crude estimate as shown in Table 12-6. An estimate unconfounded by age is obtained by applying formula 12-11, using the weights shown in Table 12-6. The variance for the effect estimate is considerably larger in the younger age category, just the reverse of the result seen in Table 12-5 for risk difference estimation. Small values of risk lead to stable estimates of risk difference but unstable estimates of risk ratio. For risk ratio estimation, then, a relatively large weight is assigned to the older age category.



*Table 12-6. Stratum-specific estimates of cumulative incidence ratio, with logarithmic transformations, variances, and weights for pooling for the data in example 12-4*

| Age | Estimate of cumulative incidence ratio | Logarithm of $\hat{RR}_i$ | Variance of $\ln(\hat{RR}_i)$ | Weight |
|---|---|---|---|---|
| < 55 | 1.81 | 0.59 | 0.31 | 3.25 |
| ≥ 55 | 1.19 | 0.18 | 0.09 | 11.6 |

The antilogarithm of the weighted average of the logarithms of the stratum-specific risk ratio estimates gives the pooled estimate, which is 1.31 for the data in Example 12-4. The discrepancy between the crude estimate, 1.44, and the unconfounded estimate, 1.31, indicates the extent of confounding.

*Directly Pooled Point Estimation of a Uniform Effect with Case-Control (or Prevalence) Data.* The effect parameter of interest with case-control data is the odds ratio, which serves as an estimator of the incidence rate ratio. The odds ratio is also the measure of interest for cross-sectional prevalence data, which should generally be treated as case-control data for the reasons given in Chapter 6. As discussed in Chapter 11, the odds ratio may also be used as an approximate estimator of the risk ratio or prevalence ratio from 2 × 2 tables with cumulative incidence or prevalence data, in which case the same formulas for pooling as given below for case-control data would apply.

For the odds ratio, as for other ratio estimators, a logarithmic transformation is desirable before weighting the stratum-specific estimates. The approximate variance of the stratum-specific estimate of the logarithm of the odds ratio (OR) is [Woolf, 1954]

$$\text{Var}[\ln(\hat{OR}_i)] \doteq \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \qquad [12\text{-}12]$$

and therefore the weight is

$$w_i = \frac{a_i b_i c_i d_i}{a_i b_i c_i + a_i b_i d_i + a_i c_i d_i + b_i c_i d_i} = \frac{1}{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}} \qquad [12\text{-}13]$$

*Example 12-5. Infants with congenital heart disease and Down syndrome, and healthy controls, according to maternal spermicide use before conception and maternal age at delivery [Rothman, 1982]*

| | Maternal age < 35 | | | Maternal age 35+ | | | Totals | | |
|---|---|---|---|---|---|---|---|---|---|
| | Spermicide use | | | Spermicide use | | | Spermicide use | | |
| | + | − | Total | + | − | Total | + | − | Total |
| Down syndrome | 3 | 9 | 12 | 1 | 3 | 4 | 4 | 12 | 16 |
| Controls | 104 | 1059 | 1163 | 5 | 86 | 91 | 109 | 1145 | 1254 |
| Total | 107 | 1068 | 1175 | 6 | 89 | 95 | 113 | 1157 | 1270 |

*Table 12-7. Stratum-specific estimates of the odds ratio with logarithmic transformations, variances, and weights for pooling for the data in example 12-5*

| Maternal age | Odds ratio | Logarithm of $\hat{OR}_i$ | Variance of $\ln(\hat{OR}_i)$ | Weight |
|---|---|---|---|---|
| < 35 | 3.4 | 1.22 | 0.46 | 2.20 |
| 35+ | 5.7 | 1.75 | 1.54 | 0.65 |

and the pooled estimator is

$$\hat{OR} = \exp\left[\frac{\sum_i w_i \ln(\hat{OR}_i)}{\sum_i w_i}\right] \qquad [12\text{-}14]$$

The case-control data in Example 12-5 describe an association between spermicide use near the time of conception and the risk of Down syndrome. The crude estimate of effect is $(4 \cdot 1145)/(12 \cdot 109) = 3.5$. The application of formula 12-14, based on the calculations in Table 12-7, gives a result of $\hat{OR} = 3.8$, which indicates only a modest degree of confounding by maternal age at delivery.

### POOLING USING THE METHOD OF MAXIMUM LIKELIHOOD

A full discussion of the maximum likelihood approach to estimation is beyond the scope of this book; the method is described adequately in many statistics texts. Briefly, the approach involves specifying the likelihood equation for the data as a function of the parameter of interest; the maximum likelihood estimate of the parameter is the value of the parameter that makes the observations in hand most probable under the likelihood model. The maximization is usually accomplished by maximizing the logarithm of the likelihood rather than the likelihood itself because the two maxima occur at the same value for the parameter, and the maximum of the logarithm of the likelihood is usually easier to determine. By



setting the first derivative of the log-likelihood function equal to zero, an equation or set of equations is derived that yields the maximum likelihood estimate for the parameter.

For most applications, the maximum likelihood estimator requires the iterative solution of a high-order equation or system of high-order equations, clearly a task for a computer rather than pencil and paper. The complicated equations do not involve any direct set of weights by which stratum-specific effect estimates are averaged, but the solution is always within the range of the stratum-specific estimates and behaves as if it were a weighted average in the sense that appropriately large weight is given to the strata with small variances for the effect estimate. In compensation for the difficulty of computation, maximum likelihood estimators have the most desirable statistical properties of all estimators, being highly efficient and minimally biased asymptotically.

A major disadvantage of the directly pooled estimators is that the pooling weight for each stratum is taken as the inverse of the variance of the effect estimate for that stratum, as estimated from the data in that stratum alone. For data with small frequencies, the variance estimates and therefore the weights can be highly inaccurate. Indeed, for data containing one or more zero frequencies, some of the variance estimates given above are infinity, corresponding to a weight of zero. Consider, for example, formula 12-12, which estimates the variance of the logarithm of the odds ratio for a $2 \times 2$ table. If any of the four cells in the table is zero, this formula gives a value of infinity, and a weight of zero would be assigned to that table. (Furthermore, the logarithm of the odds ratio for that stratum would not be finite with a zero cell.) If the remaining cells in the table are large, there might be a considerable amount of information in the table that would be lost by assigning a weight of zero. Since the odds ratio for the stratum is either zero or infinity, which are the most extreme possibilities, it seems obviously incorrect to ignore such information. One proposed solution to this problem has been to modify formula 12-12 and, by extension, other formulas like it by adding a constant value (usually 0.5 or 1.0) to each observed frequency [Haldane, 1955] or to substitute a small constant for the zero frequencies when they occur. Although this solution mitigates the problem and avoids the difficulty of dividing by zero, it does not completely overcome the inaccuracy of the variance estimates for each stratum-specific estimate of effect when some of the observations are small. The maximum likelihood approach is preferable when some of the observed frequencies are small. Rather than treating each stratum in isolation, as does the directly weighted approach to pooling in the assignment of weights, the maximum likelihood approach automatically "adjusts" the observations in each stratum in a way that integrates the information among all strata.

In the following sections, the equations are presented for maximum likelihood estimation of a uniform effect measure. In each case, the result

can be obtained by writing the likelihood equation for the data as a function of a uniform effect measure, the observations, and whatever "nuisance" parameters may be involved, and then setting the derivative of the logarithm of the likelihood equal to zero.

*Maximum Likelihood Estimation of a Uniform Effect with Person-Time Data.* INCIDENCE RATE DIFFERENCE. The maximum likelihood estimation of incidence rate difference necessitates the solution of a set of equations that number one more than the number of strata. In addition to solving for the incidence rate difference (IRD), it is necessary to solve for the value of the incidence rate among the unexposed group in each stratum, satisfying the following likelihood equations:

$$\sum_i \frac{a_i}{\hat{R}_{0i} + I\hat{R}D} - \sum_i N_{1i} = 0 \qquad [12\text{-}15]$$

and, for each stratum i,

$$\frac{a_i}{\hat{R}_{0i} + I\hat{R}D} + \frac{b_i}{\hat{R}_{0i}} - T_i = 0 \qquad [12\text{-}16]$$

where $I\hat{R}D$ is the pooled estimate of incidence rate difference, $\hat{R}_{0i}$ is the estimate of incidence rate among unexposed in stratum i, and the general notation follows that of Table 12-1. The estimates $\{\hat{R}_{0i}\}$ are estimates of nuisance parameters that must be calculated to solve for the desired estimate, $I\hat{R}D$. It is convenient to begin the solution to the above equations using the observed rate for unexposed within each stratum as a starting value, but the value for $\hat{R}_{0i}$ that satisfies the equation can differ considerably from the observed value. The overall solution of equations 12-15 and 12-16 is best accomplished by starting with a trial value for $I\hat{R}D$, solving iteratively for each $\hat{R}_{0i}$, and then evaluating the left side of equation 12-15. Repeated trial values for $I\hat{R}D$ each require an iterative solution for equation 12-16 in each stratum, making the overall process tedious unless it is done by computer.

For the data in Example 12-2 the maximum likelihood solution for $I\hat{R}D$ is $5.91 \times 10^{-4}\text{yr}^{-1}$, which is in close agreement with the directly pooled result of $5.95 \times 10^{-4}\text{yr}^{-1}$ obtained previously.

INCIDENCE RATE RATIO. For the maximum likelihood estimation of incidence rate ratio (IRR), no nuisance parameters are involved, and the estimate is obtained by the iterative solution of a single equation:

$$\sum_i a_i - I\hat{R}R \sum_i \frac{M_{1i}}{I\hat{R}R + \frac{N_{0i}}{N_{1i}}} = 0 \qquad [12\text{-}17]$$



For the data in Example 12-3, the maximum likelihood estimate of IRR is 1.50, which is identical to the result obtained by direct pooling. One would expect good agreement between these two approaches when the observed frequencies are reasonably large, as they are in this example. In addition, the narrow spread between the stratum-specific estimates confines all pooled estimates to the same small range of possible values.

*Maximum Likelihood Estimation of a Uniform Effect with Cumulative Incidence Data.* CUMULATIVE INCIDENCE DIFFERENCE. The maximum likelihood estimation of cumulative incidence difference (RD) again involves the estimation of a set of nuisance parameters, the risk among the nonexposed group in each stratum. As with person-time data, the number of equations is one more than the number of strata:

$$\sum_i \frac{a_i}{\hat{R}_{0i} + \hat{RD}} - \sum_i \frac{N_{1i} - a_i}{1 - \hat{R}_{0i} - \hat{RD}} = 0 \qquad [12\text{-}18]$$

and, for each stratum i,

$$\frac{a_i}{\hat{R}_{0i} + \hat{RD}} + \frac{b_i}{\hat{R}_{0i}} - \frac{N_{1i} - a_i}{1 - \hat{R}_{0i} - \hat{RD}} - \frac{N_{0i} - b_i}{1 - \hat{R}_{0i}} = 0 \qquad [12\text{-}19]$$

where $\hat{R}_{0i}$ is the maximum likelihood estimate of cumulative incidence among unexposed in stratum i, and the notation follows that in Table 12-4.

Solving the above equations for $\hat{RD}$ using the data from Example 12-4 gives the maximum likelihood estimate of risk difference as 0.034, which is virtually identical to the value derived from direct pooling. The extremely narrow range separating the two stratum-specific point estimates ensures good agreement for any pooled estimators in this example.

CUMULATIVE INCIDENCE RATIO. For maximum likelihood estimation of cumulative incidence ratio (RR), again the risk among unexposed in each stratum is a nuisance parameter that must be estimated, but the maximum likelihood solution for each $\hat{R}_{0i}$ has a closed form solution conditional on $\hat{RR}$. The equations are

$$\sum_i \frac{a_i - (\hat{RR})\hat{R}_{0i}N_{1i}}{1 - (\hat{RR})\hat{R}_{0i}} = 0 \qquad [12\text{-}20]$$

and, for each stratum *i*,

$$\hat{R}_{0i} = \frac{a_i + N_{0i} + (\hat{RR})(b_i + N_{1i})}{2(\hat{RR})T_i} - \left(\left[\frac{a_i + N_{0i} + (\hat{RR})(b_i + N_{1i})}{2(\hat{RR})T_i}\right]^2 - \frac{M_{1i}}{(\hat{RR})T_i}\right)^{1/2} \qquad [12\text{-}21]$$

Solution of these equations for $\widehat{RR}$ using the data in Example 12-4 gives the maximum likelihood estimate of 1.31, again identical to the directly pooled estimate.

*Maximum Likelihood Estimation of a Uniform Effect with Case-Control (or Prevalence) Data.* The maximum likelihood estimation of a uniform odds ratio is the solution, $\widehat{OR}$, to the following equations:

$$\sum_i a_i - \sum_i \hat{a}_i = 0 \qquad [12\text{-}22]$$

The quantity $\hat{a}_i$ is the "expected" value for the a cell in each 2 × 2 table, calculated as a function of the odds ratio. For each 2 × 2 table $\hat{a}_i$ can be calculated from the formula

$$\hat{a}_i = \frac{\sum_{k=\max(0,M_{1i}-N_{0i})}^{\min(M_{1i},N_{1i})} k \binom{N_{1i}}{k}\binom{N_{0i}}{M_{1i}-k}\widehat{OR}^k}{\sum_{k=\max(0,M_{1i}-N_{0i})}^{\min(M_{1i},N_{1i})} \binom{N_{1i}}{k}\binom{N_{0i}}{M_{1i}-k}\widehat{OR}^k} \qquad [12\text{-}23]$$

where the notation follows that in Table 12-4. Equation 12-23 can be computationally tedious if the numbers within a stratum are large, but in such circumstances an excellent asymptotic approximation for $\hat{a}_i$ is obtained from the equation

$$\widehat{OR} = \frac{\hat{a}_i\hat{d}_i}{\hat{b}_i\hat{c}_i} \qquad [12\text{-}24]$$

in which $\hat{a}_i, \hat{b}_i, \hat{c}_i$, and $\hat{d}_i$ are the expected cell values that conform to equations 12-22 and 12-24 and to the marginal totals of the 2 × 2 table [Gart, 1970]. Solving equation 12-24 explicitly for $\hat{a}_i$ in terms of $T_i$, $N_{1i}$, $M_{1i}$ and $\widehat{OR}$ gives

$$\hat{a}_i = \mathrm{ABS}\left\{\mathrm{ABS}\left[\frac{1}{2}\left(\frac{T_i}{\widehat{OR}-1} + N_{1i} + M_{1i}\right)\right] - \sqrt{\left[\frac{1}{2}\left(\frac{T_i}{\widehat{OR}-1} + N_{1i} + M_{1i}\right)\right]^2 - \frac{M_{1i}N_{1i}\widehat{OR}}{\widehat{OR}-1}}\right\} \qquad [12\text{-}25]$$

which is computationally much simpler than equation 12-23. Equation 12-24 represents the maximum likelihood solution for a uniform odds ratio based on 2 × 2 tables with two independent binomials; this "unconditional" solution (it is conditional on one margin of the 2 × 2 table but not on both) generally gives nearly identical results to those obtained from the difficult-to-calculate conditional formula 12-23 except when the average number of subjects per stratum is small. In such instances, the unconditional approach can be substantially biased, and it is preferable to use the conditional approach or the Mantel-Haenszel estimator [Breslow, 1981; McKinlay, 1978; Lubin, 1981]. (Directly weighted pooled estimation is also unreliable if the number of subjects per stratum is small.)



Maximum likelihood estimation of the odds ratio in a set of 2 × 2 tables requires an iterative solution of equation 12-22 coupled with either equation 12-23 or equation 12-25, using trial values for the odds ratio until equation 12-22 is satisfied. For the data in Example 12-5, the conditional maximum likelihood estimate of the odds ratio (i.e., using equation 12-23) is 3.76; the unconditional maximum likelihood estimate (using equation 12-25) is 3.79. Despite the small cell frequencies for the cases and a moderate discrepancy between the stratum-specific estimates of the odds ratio (3.4 for younger mothers and 5.7 for older mothers), the two likelihood approaches give nearly identical results because the total number of subjects per stratum is large. Furthermore, these estimates agree closely with the directly pooled estimate for these data, which is 3.82.

### POOLING WITH MANTEL-HAENSZEL ESTIMATORS

Mantel and Haenszel [1959] have proposed a simple formula as an estimator of a uniform odds ratio in a set of 2 × 2 tables. The estimator is

$$\widehat{OR}_{MH} = \frac{\sum_i a_i d_i/T_i}{\sum_i b_i c_i/T_i} \qquad [12\text{-}26]$$

This formula represents a weighted average, without logarithmic transformation, of the stratum-specific estimates of the odds ratio, with the weight for each stratum equal to $b_ic_i/T_i$:

$$\frac{a_id_i}{b_ic_i} \cdot \frac{b_ic_i}{T_i} = \frac{a_id_i}{T_i}$$

These weights are inversely proportional to the variance of the logarithm of the odds ratio under the null condition. Consequently, the Mantel-Haenszel pooled estimator is optimally weighted for stratum-specific odds ratio estimates near 1.0. Theoretical statistical evaluation of the Mantel-Haenszel estimator with respect to bias and precision has shown that it compares favorably with the maximum likelihood estimator (formulas 12-22 through 12-25) under a variety of conditions [Breslow and Liang, 1982]. Whereas the directly pooled estimators require reasonably large frequencies within each stratum, the Mantel-Haenszel estimator, like the conditional maximum likelihood estimators, performs well even if the frequencies within

strata are small or if the data contain an occasional zero. Furthermore, it has the advantage of being extremely simple to calculate. For example, the Mantel-Haenszel estimator of a uniform odds ratio for the data in Example 12-5 is calculated as

$$\widehat{OR}_{MH} = \frac{(3)\,(1059)/(1175) + (1)\,(86)/(95)}{(104)\,(9)/(1175) + (5)\,(3)/(95)} = 3.8$$

This result is nearly identical to the maximum likelihood estimate (and the directly pooled estimate) but is extraordinarily simpler to produce. The combination of ease of computation and desirable statistical properties make this estimator the preferred choice for most situations in which an estimate of the odds ratio is desired for a set of 2 × 2 tables.

By analogy with the Mantel-Haenszel estimator, it is reasonable to construct estimators for the other ratio measures of effect weighted in a similar way. For incidence rate data, the analogous estimator is [Rothman and Boice, 1982]

$$\widehat{IRR}_{MH} = \frac{\sum_i a_i N_{0i}/T_i}{\sum_i b_i N_{1i}/T_i} \qquad [12\text{-}27]$$

This formula is a simple noniterative estimator for a uniform incidence rate ratio that is nearly as efficient as the maximum likelihood estimator [Walker, 1985]. Formula 12-27 may also be used for cumulative incidence to obtain $\widehat{RR}_{MH}$ [Nurminen, 1981; Tarone, 1981].

For the data in Example 12-3, formula 12-27 yields.

$$\widehat{IRR}_{MH} = \frac{(14)\,(1701)/(3217) + (76)\,(2245)/(3194)}{(10)\,(1516)/(3217) + (121)\,(949)/(3194)} = 1.50$$

which is identical for practical purposes with both the maximum likelihood and the directly pooled results.

Formula 12-27 applied to the data in Example 12-4 gives

$$\widehat{RR}_{MH} = \frac{(8)\,(120)/(226) + (22)\,(85)/(183)}{(5)\,(106)/(226) + (16)\,(98)/(183)} = 1.33$$

which is reasonably close to both the maximum likelihood and the directly pooled results.

Greenland and Robins [1985] have suggested extending the Mantel-Haenszel approach to difference measures. The statistical properties of the Mantel-Haenszel estimators for difference measures are better than any other approach for very sparse data within strata, but the variance of the



Mantel-Haenszel effect measures is much greater than that of either the directly weighted or maximum likelihood methods when the data are ample. Mantel-Haenszel difference measures are not covered here.

*Statistical Hypothesis Testing for Stratified Data*

Examples can be found in which a pooled estimate of rate difference shows a negative association whereas a pooled estimate of rate ratio shows a positive association for the same data, apparently indicating that there is not a perfect correspondence between ratio and difference measures with regard to the absence or direction of effect. These discrepancies stem from variation introduced by different weighting schemes. For the purposes of statistical hypothesis testing, there is a theoretical correspondence of different measures at the null point, and consequently only a single hypothesis test need be considered, whatever the parameter used to assess the effect. The tests in common use correspond to the conditional tests for simple data that assume either a fixed number of cases for incidence rate data or fixed marginal totals for 2 × 2 tables. Strictly speaking, these are tests of a departure from unity of the odds ratio or the incidence rate ratio, but the tests are valid as tests of the null hypothesis whatever the measure of interest.

With stratified data, it is possible that the effect may vary substantially from one stratum to another. Nevertheless, hypothesis testing is generally performed with respect to the overall departure of the data from the null value of no association. That is, even if the parameter value for the effect varies among strata, the hypothesis test represents a test of the departure of some single overall measure of effect from its null value; it is convenient to think of this process as testing the departure of a pooled estimate of effect from the null value. If the stratum-specific values of the odds ratio or incidence rate ratio are identical, the tests described later are extremely powerful; in fact, in the jargon of statistics they are "uniformly most powerful," which means that they are the best possible tests of the null hypothesis in those circumstances. If the values of the odds ratio or incidence rate ratio vary across strata, it is conceivable that specialized tests could be constructed that would be more powerful than the tests of overall departure from the null value described here; the specialized tests would have to be designed to detect a particular pattern of variation of the effect across strata. In general, however, the tests of the departure of the pooled estimate from the null value are still valid, even if they are not theoretically the most powerful tests that might be applied in a given situation. As a practical matter, usually no useful alternative exists.

In certain situations it is conceivable that estimates of an effect could be strongly positive in some strata and strongly negative in others. In such circumstances the pooled estimate of effect may be near the null value as a result of the balancing of the opposing effect estimates in individual

strata. A test of the overall departure of the data from the null condition would have little meaning in these circumstances as long as the opposing effect estimates reflect actual divergence of the parameter rather than simply random variability of the effect estimate around the null value.

Statistical hypothesis testing for stratified data represents a straightforward extension of the tests applied to crude data. The exact tests used are based on the probability calculations for a set of strata; the probability of observing a set of outcomes is the product of the probability for each outcome, so the probability of observing the set of observations in stratified data is calculated as the product of the probability of the outcome in each stratum. The latter probability is determined using the same probability model as that used for crude data. Although this extension of exact testing to stratified data is conceptually simple, in practice the large number of combinations of outcomes can make the computations tedious to enumerate and perform except by computer.

The approximate tests for stratified data retain the general form of expression 10-1 and merely extend the formulas for crude data given in Chapter 11 by deriving the components of the test statistics (the observed number of exposed cases, the number expected under the null hypothesis, and the variance) by summing the contributions to each of these three components over the set of strata.

### HYPOTHESIS TESTING WITH STRATIFIED PERSON-TIME DATA

*Exact Hypothesis Testing with Stratified Person-Time Data.* For stratified data, the overall probabilities needed for the calculation of exact *P*-values are the products of the probabilities obtained within each stratum. The total number of possible outcomes is usually large, especially in comparison with crude data, making exact *P*-value calculations for stratified data difficult. In practice, they are rarely done. Nevertheless, the principles for obtaining exact *P*-values with stratified data are straightforward extensions of the principles applicable to crude data, and the computations can be readily programmed into a computer.

The probability formula for the number of exposed cases in a single stratum is identical to the formula used for crude data:

$$\Pr(\text{number of exposed cases in stratum } i = a_i) = \binom{M_{1i}}{a_i}\left(\frac{N_{1i}}{T_i}\right)^{a_i}\left(\frac{N_{0i}}{T_i}\right)^{b_i}$$

The probability that a set of N strata will have exactly $a_i$ exposed cases in each stratum i, i = 1, 2, . . . N, is the product of the probability of finding exactly $a_i$ exposed cases in each of the component strata:

$$\Pr(\text{the set of observations } \{a_i\}) = \prod_{i=1}^{N}\binom{M_{1i}}{a_i}\left(\frac{N_{1i}}{T_i}\right)^{a_i}\left(\frac{N_{0i}}{T_i}\right)^{b_i} \qquad [12\text{-}28]$$



There is some complexity involved in determining which outcomes of the data are considered equally extreme and more extreme in relation to the actual observations. The problem calls for considering all possible combinations of values for the possible number of exposed cases in each stratum, the number in each stratum being subject to the constraint of the total number of cases, exposed or unexposed, that actually occurred in that stratum.

For example, consider the data presented in Example 12-6. There are a total of 16 cases in the three age strata, of which 9 are exposed. The most extreme outcome, conditional on the number of cases observed in each stratum, would be that all 16 cases were exposed, 2, 12, and 2, respectively, in each of the three age strata. Consider the possible outcomes for which 15 cases are exposed. There are three ways in which 15 of the 16 cases could be exposed: The unexposed case could fall into any of the three age strata. These three possibilities correspond to the distribution of the exposed cases being 1-12-2, 2-11-2, and 2-12-1. A complete enumeration of all the possible outcomes for at least 9 exposed cases is listed in Table 12-8. The 54 combinations constitute the outcomes in the upper tail of the probability distribution for testing the null hypothesis of no association between dose category and thyroid neoplasm. To obtain the exact upper-tail *P*-value, the probability of each of these 54 possible outcomes must be calculated according to formula 12-28.

For example, the probability of the actual observations 0-7-2 is calculated, according to Formula 12-28, to be

$$\binom{2}{0}\left(\frac{1054}{10{,}996}\right)^{0}\left(\frac{9942}{10{,}996}\right)^{2}\cdot\binom{12}{7}\left(\frac{2665}{18{,}075}\right)^{7}\left(\frac{15{,}410}{18{,}075}\right)^{5}$$
$$\cdot\binom{2}{2}\left(\frac{2217}{3747}\right)^{2}\left(\frac{1530}{3747}\right)^{0} = (0.8175)\times(0.00054036)\times(0.3501) = 0.000155$$

The sum of the probability of all the outcomes in Table 12-8 equals the upper-tail Fisher exact *P*-value testing the null hypothesis of no association

*Example 12-6. Incidence of thyroid neoplasms in females by age, for those exposed to less than 100 rad and those exposed to 300+ rad of radiation* [*Hempelmann et al., 1975*]

| | 0–14 Years | | 15–29 Years | | 30+ Years | |
|---|---|---|---|---|---|---|
| | 300+ rad | < 100 rad | 300+ rad | < 100 rad | 300+ rad | < 100 rad |
| Cases | 0 | 2 | 7 | 5 | 2 | 0 |
| Person-years | 1054 | 9942 | 2665 | 15,410 | 2217 | 1530 |

*Table 12-8. Enumeration of all possible combinations of exposed cases by age category, with at least nine exposed cases, for the data of example 12-6*

| Total no. exposed cases | Distribution of exposed cases by age category | Total no. exposed cases | Distribution of exposed cases by age category | Total no. exposed cases | Distribution of exposed cases by age category |
|---|---|---|---|---|---|
| 16 | 2–12–2 | 12 | 0–12–0 | 10 | 0–10–0 |
| 15 | 1–12–2 | 12 | 0–11–1 | 10 | 0–9–1 |
| 15 | 2–12–1 | 12 | 0–10–2 | 10 | 0–8–2 |
| 15 | 2–11–2 | 12 | 1–11–0 | 10 | 1–9–0 |
| 14 | 0–12–2 | 12 | 1–10–1 | 10 | 1–8–1 |
| 14 | 1–12–1 | 12 | 1–9–2 | 10 | 1–7–2 |
| 14 | 1–11–2 | 12 | 2–10–0 | 10 | 2–8–0 |
| 14 | 2–12–0 | 12 | 2–9–1 | 10 | 2–7–1 |
| 14 | 2–11–1 | 12 | 2–8–2 | 10 | 2–6–2 |
| 14 | 2–10–2 | 11 | 0–11–0 | 9 | 0–9–0 |
| 13 | 0–12–1 | 11 | 0–10–1 | 9 | 0–8–1 |
| 13 | 0–11–2 | 11 | 0–9–2 | 9 | 0–7–2 |
| 13 | 1–12–0 | 11 | 1–10–0 | 9 | 1–8–0 |
| 13 | 1–11–1 | 11 | 1–9–1 | 9 | 1–7–1 |
| 13 | 1–10–2 | 11 | 1–8–2 | 9 | 1–6–2 |
| 13 | 2–11–0 | 11 | 2–9–0 | 9 | 2–7–0 |
| 13 | 2–10–1 | 11 | 2–8–1 | 9 | 2–6–1 |
| 13 | 2–9–2 | 11 | 2–7–2 | 9 | 2–5–2 |

between level of radiation exposure and incidence of thyroid neoplasm. Algebraically, the tail probability is expressed as

$$\Pr(k \geq a) = \sum_{k=a}^{M_1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} \left(\frac{N_{1i}}{T_i}\right)^{k_i} \left(\frac{N_{0i}}{T_i}\right)^{M_{1i}-k_i} \qquad [12\text{-}29]$$

where $k_i$ represents the value for the number of exposed cases in stratum i, $k = \Sigma k_i$, $a = \Sigma a_i$, and $M_1 = \Sigma M_{1i}$. The sum of all the probabilities for the combinations listed in Table 12-8 is 0.000600. Interestingly, the sum of the probabilities for the nine combinations that are just as extreme as the actual observation, with exactly nine exposed cases, is 0.000522, nearly as great as the sum for all 54 outcomes listed in Table 12-8. If the nine combinations with 10 exposed cases are included, the sum increases to 0.000592, and by including the possibilities with 11 exposed cases it increases to 0.000599. Clearly it is not necessary to carry out all 54 computations to get an answer accurate enough for any scientific interpretation, since one digit of precision is usually adequate for the *P*-value.



For the lower-tail Fisher exact *P*-value, which would be calculated when the observed effect is less than the null value, the summation is

$$\Pr(k \leq a) = \sum_{k=0}^{a} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} \left(\frac{N_{1i}}{T_i}\right)^{k_i} \left(\frac{N_{0i}}{T_i}\right)^{M_{1i}-k_i} \qquad [12\text{-}30]$$

where again $k = \Sigma k_i$, and so on.

To obtain the exact mid-*P* value, only half the probability of all observations as extreme as that observed should be included in the summation:

$$\text{Upper-tail probability} = \frac{1}{2}\sum_{k=a}^{M_1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} \left(\frac{N_{1i}}{T_i}\right)^{k_i} \left(\frac{N_{0i}}{T_i}\right)^{M_{1i}-k_i} + \sum_{k=a+1}^{M_1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} \left(\frac{N_{1i}}{T_i}\right)^{k_i} \left(\frac{N_{0i}}{T_i}\right)^{M_{1i}-k_i} \qquad [12\text{-}31]$$

$$\text{Lower-tail probability} = \frac{1}{2}\sum_{k=a}^{M_1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} \left(\frac{N_{1i}}{T_i}\right)^{k_i} \left(\frac{N_{0i}}{T_i}\right)^{M_{1i}-k_i} + \sum_{k=0}^{a-1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} \left(\frac{N_{1i}}{T_i}\right)^{k_i} \left(\frac{N_{0i}}{T_i}\right)^{M_{1i}-k_i} \qquad [12\text{-}32]$$

For the data in Example 12-6, the exact upper mid-*P* value would be one-half of 0.000522, which was the probability for the nine possible outcomes with exactly nine exposed cases, plus the probability of all the possible outcomes more extreme than nine exposed cases, which was a total of 0.000078. Therefore, the exact upper mid-*P* value is 0.00034.

*Approximate Hypothesis Testing with Stratified Person-Time Data.* For stratified data, asymptotic test statistics are constructed according to the same principles used for crude data. The test variable is still the number of exposed cases, which is the sum of $a_i$ over the strata. The null expectation and the variance for the number of exposed cases is calculated within each stratum, and these results are summed over the strata. Thus, the null expectation for the number of exposed cases is

$$E(A) = \sum_{i=1}^{N} \frac{N_{1i}M_{1i}}{T_i}$$

and the variance, based on the binomial model, is

$$\text{Var}(A) = \sum_{i=1}^{N} \frac{M_{1i}N_{1i}N_{0i}}{T_i^2}$$

which gives as the test statistic

$$\chi = \frac{\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \frac{N_{1i}M_{1i}}{T_i}}{\sqrt{\sum_{i=1}^{N} \frac{M_{1i}N_{1i}N_{0i}}{T_i^2}}} \qquad [12\text{-}33]$$

Formula 12-33 is identical to formula 11-1 for crude person-time data except that the three components of the test statistic are obtained by summing their stratum-specific contributions over the strata.

For the data in Example 12-6, the test statistic is calculated as follows:

$$A = \text{no. of exposed cases} = 0 + 7 + 2 = 9$$

$$E(A) = 2\left(\frac{1054}{10{,}996}\right) + 12\left(\frac{2665}{18{,}075}\right) + 2\left(\frac{2217}{3747}\right) = 3.14$$

$$\text{Var}(A) = 2\left(\frac{1054}{10{,}996}\right)\left(\frac{9942}{10{,}996}\right) + 12\left(\frac{2665}{18{,}075}\right)\left(\frac{15{,}410}{18{,}075}\right) + 2\left(\frac{2217}{3747}\right)\left(\frac{1530}{3747}\right) = 2.16$$

and

$$\chi = \frac{A - E(A)}{\sqrt{\text{Var}(A)}} = \frac{9 - 3.14}{\sqrt{2.16}} = 3.98$$

which corresponds to a one-tail *P*-value of 0.000034 or a two-tail *P*-value of 0.000069.

The *P*-value calculated from this approximate test statistic, like the exact *P*-value, is very small, but the two *P*-values do not agree closely. The exact mid-*P* value is about 10 times the magnitude of the approximate *P*-value. The discrepancy stems from the small numbers involved but is also related to the fact that the normal approximation is poorer in the extremities of the distribution.

Nevertheless, comparison between the exact and the asymptotic test raises the question of the nature of the applicability criteria for the asymptotic test statistic with regard to the number of observations. There is no simple answer to this question, but one important point should be emphasized: The large-number condition need apply only to the summations involved in formula 12-33, not to each individual stratum. For person-time



data, then, formula 12-33 would apply even if each stratum had only one case, provided that there were enough such strata to allow the distribution of the total number of exposed subjects in all strata to be well enough approximated by a normal distribution. The large-number condition necessary for formula 12-33 to apply, then, could be reached by having few strata with many observations in each one or many strata with sparse data. With one stratum, formula 12-33 reduces to formula 11-1. A stratum with no cases has no information and contributes nothing to A, E(A), or Var(A).

As a second example of the application of formula 12-33, consider the data in Example 12-3. The large number of cases in each of these two strata make it unnecessary to contemplate any exact test. The *P*-value can be determined as follows (considering male gender as "exposed"):

$$A = \text{no. of exposed cases} = 14 + 76 = 90$$

$$E(A) = 24\left(\frac{1516}{3217}\right) + 197\left(\frac{949}{3194}\right) = 69.8$$

$$\text{Var}(A) = 24\left(\frac{1516}{3217}\right)\left(\frac{1701}{3217}\right) + 197\left(\frac{949}{3194}\right)\left(\frac{2245}{3194}\right) = 47.1$$

$$\chi = \frac{90 - 69.8}{\sqrt{47.1}} = 2.94$$

$$P_{(1)} = 0.0017$$

## HYPOTHESIS TESTING WITH STRATIFIED CUMULATIVE INCIDENCE, PREVALENCE OR CASE-CONTROL DATA (2 × 2 TABLES)

*Exact Hypothesis Testing for Stratified 2 × 2 Tables.* As with person-time data, exact hypothesis testing for stratified 2 × 2 tables can be accomplished by enumerating all possible outcomes of the number of exposed cases across strata. The joint probability of each combination is calculated as the product of the hypergeometric probabilities of each 2 × 2 table. The exact *P*-value is determined in the usual way by summing the probabilities in the tail of the distribution. Each 2 × 2 table is considered to have all marginal totals fixed. Using the notation of Table 12-4, we have, for the Fisher *P*-values,

$$\Pr(k \geq a) = \sum_{k=a}^{\min(N_1, M_1)} \prod_{i=1}^{N} \frac{\binom{N_{1i}}{k_i}\binom{N_{0i}}{M_{1i} - k_i}}{\binom{T_i}{M_{1i}}} \qquad [12\text{-}34]$$

for the upper tail, when the effect estimate is greater than the null value, and

$$\Pr(k \leq a) = \sum_{k=\max(0,M_1-N_0)}^{a} \prod_{i=1}^{N} \frac{\binom{N_{1i}}{k_i}\binom{N_{0i}}{M_{1i}-k_i}}{\binom{T_i}{M_{1i}}} \qquad [12\text{-}35]$$

for the lower tail, when the effect estimate is less than the null value. The summations are for $k = \Sigma k_i$, where $k_i$ is the number of exposed cases in each 2 × 2 table, and a, $N_1$, $N_0$, and $M_1$ refer to $\Sigma a_i$, $\Sigma N_{1i}$, $\Sigma N_{0i}$, and $\Sigma M_{1i}$, respectively.

To obtain the exact mid-*P* value, only half of the probability for the equally extreme outcomes should be added to the tail summation:

Upper-tail probability

$$= \frac{1}{2}\sum_{k=a}^{N} \prod_{i=1}^{N} \frac{\binom{N_{1i}}{k_i}\binom{N_{0i}}{M_{1i}-k_i}}{\binom{T_i}{M_{1i}}} + \sum_{k=a+1}^{\min(N_1,M_1)} \prod_{i=1}^{N} \frac{\binom{N_{1i}}{k_i}\binom{N_{0i}}{M_{1i}-k_i}}{\binom{T_i}{M_{1i}}} \qquad [12\text{-}36]$$

$$\text{Lower-tail probability} = \frac{1}{2}\sum_{k=a}^{N} \prod_{i=1}^{N} \frac{\binom{N_{1i}}{k_i}\binom{N_{0i}}{M_{1i}-k_i}}{\binom{T_i}{M_{1i}}}$$

$$+ \sum_{k=\max(0,M_1-N_0)}^{a-1} \prod_{i=1}^{N} \frac{\binom{N_{1i}}{k_i}\binom{N_{0i}}{M_{1i}-k_i}}{\binom{T_i}{M_{1i}}} \qquad [12\text{-}37]$$

These formulas can be used to derive an exact *P*-value for the data in Example 12-5. There were four exposed cases, three in the "young" stratum and one in the "old" stratum. In the young stratum, the number of exposed cases can range from zero to 12 within the constraints imposed by the marginal totals. In the old stratum, the corresponding number can range from zero to 4. Overall, there are 13 × 5 = 65 possible outcomes for the two strata. Of these 65 possibilities, 50 are more extreme positive departures from the null condition than the outcome actually observed; a total of five possibilities, including the observed data, are equally extreme with exactly four exposed cases. These 55 equally or more extreme outcomes are enumerated in Table 12-9.



The probability of observing the actual outcome in Example 12-5 is determined as

$$\Pr(\text{observed data}) = \frac{\binom{107}{3}\binom{1068}{9}\binom{6}{1}\binom{89}{3}}{\binom{1175}{12}\binom{95}{4}}$$

$$= \frac{(107!)\,(1068!)\,(6!)\,(89!)\,(1163!)\,(12!)\,(91!)\,(4!)}{(1175!)\,(95!)\,(104!)\,(3!)\,(1059!)\,(9!)\,(5!)\,(86!)\,(3!)} = 0.0150$$

The probabilities for all five of the possible outcomes resulting in four exposed cases are, in the order the outcomes are listed in Table 12-9, 0.0118, 0.0150, 0.0039, 0.0002, and 0.0000 (the last one is actually 0.00000149), which totals to 0.0309. The corresponding total for the outcome with five exposed cases is 0.0066, and for six exposed cases, 0.0010. The Fisher *P*-value for the upper tail of the distribution can therefore be estimated as 0.0309 + 0.0066 + 0.0010 = 0.0385, assuming that the outcomes with more than six exposed cases do not contribute materially to the summation. The full tail probability is actually 0.0387, based on all 55 possibilities in Table 12-9, so that the truncation after six exposed cases is reasonable. The remaining 10 outcomes with three or fewer exposed cases, which are not listed in Table 12-9, account, in probability terms, for 1 − 0.0387 or about 96 percent of the distribution; in fact, the outcomes 0-0 and 1-0 together account for about 54 percent of the distribution.

To get the exact mid-*P* value for the upper tail, only half of the probability of getting four exposed cases should be included, which results in a *P*-value of 0.0233.

*Approximate Hypothesis Testing for Stratified 2 × 2 Tables.* The extension of formula 11-6 to stratified 2 × 2 tables is analogous to the extension of formula 11-1 for person-time data. As before, the contribution to each of the three components of the test statistic—the number of exposed cases, the null expectation for the number of exposed cases, and the variance—is derived separately for each stratum and then summed over the strata. Thus, the null expectation for the number of exposed cases is

$$E(A) = \sum_{i=1}^{N} \frac{N_{1i}M_{1i}}{T_i}$$

and the variance, based on the hypergeometric model, is

$$\text{Var}(A) = \sum_{i=1}^{N} \frac{N_{1i}N_{0i}M_{1i}M_{0i}}{T_i^2(T_i - 1)}$$

*able 12-9. Enumeration of all possible combinations of exposed cases by ge category, with at least four exposed cases, for the data in example 12-5*

| otal no. xposed ases | Distribution of exposed cases by age category | Total no. exposed cases | Distribution of exposed cases by age category | Total no. exposed cases | Distribution of exposed cases by age category |
|---|---|---|---|---|---|
| 6 | 12–4 | 10 | 10–0 | 6 | 6–0 |
| 5 | 12–3 | 10 | 9–1 | 6 | 5–1 |
| 5 | 11–4 | 10 | 8–2 | 6 | 4–2 |
| 4 | 12–2 | 10 | 7–3 | 6 | 3–3 |
| 4 | 11–3 | 10 | 6–4 | 6 | 2–4 |
| 4 | 10–4 | 9 | 9–0 | 5 | 5–0 |
| 3 | 12–1 | 9 | 8–1 | 5 | 4–1 |
| 3 | 11–2 | 9 | 7–2 | 5 | 3–2 |
| 3 | 10–3 | 9 | 6–3 | 5 | 2–3 |
| 3 | 9–4 | 9 | 5–4 | 5 | 1–4 |
| 2 | 12–0 | 8 | 8–0 | 4 | 4–0 |
| 2 | 11–1 | 8 | 7–1 | 4 | 3–1 |
| 2 | 10–2 | 8 | 6–2 | 4 | 2–2 |
| 2 | 9–3 | 8 | 5–3 | 4 | 1–3 |
| 2 | 8–4 | 8 | 4–4 | 4 | 0–4 |
| 1 | 11–0 | 7 | 7–0 | | |
| 1 | 10–1 | 7 | 6–1 | | |
| 1 | 9–2 | 7 | 5–2 | | |
| 1 | 8–3 | 7 | 4–3 | | |
| 1 | 7–4 | 7 | 3–4 | | |

which gives as the test statistic

$$\chi = \frac{\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \frac{N_{1i}M_{1i}}{T_i}}{\sqrt{\sum_{i=1}^{N} \frac{N_{1i}N_{0i}M_{1i}M_{0i}}{T_i^2(T_i - 1)}}} \qquad [12\text{-}38]$$

The above test statistic, first proposed by Mantel and Haenszel in 1959 and known as the Mantel-Haenszel test, is widely used in epidemiologic analyses and other applications in which stratified 2 × 2 tables are used. It is optimal in statistical power when the odds ratio is uniform across strata, but it is generally the most useful and convenient test even if the odds ratio varies across strata. The $\chi$ takes a value of zero only when the Mantel-Haenszel pooled estimator of the odds ratio (formula 12-26) equals unity, so that the test statistic may be considered a test of the departure of $\widehat{OR}_{MH}$ from unity. The large-number applicability condition does not refer



to individual strata but only to the summations in formula 12-38. Individual strata may each have as few as two subjects as long as no marginal total is zero; if a marginal total is zero, the stratum has no information. The test statistic will be applicable if there is a sufficient number of strata, even with sparse data. As we shall see in Chapter 13, the test is the one that applies even to the analysis of matched-pair data.

The null hypothesis of no relation between tolbutamide and death in the University Group Diabetes Program for the age-stratified data in Example 12-4 can be evaluated with the Mantel-Haenszel test. The number of exposed cases, where "exposed" indicates tolbutamide therapy, is 8 + 22 = 30. The expected number under the null hypothesis is

$$E(A) = \frac{(106)(13)}{226} + \frac{(98)(38)}{183} = 6.10 + 20.35 = 26.45$$

and the variance of the number of exposed cases is

$$\text{Var}(A) = \frac{(106)(120)(13)(213)}{(226)^2(225)} + \frac{(98)(85)(38)(145)}{(183)^2(182)}$$

$$= 3.06 + 7.53 = 10.60$$

The test statistic is

$$\chi = \frac{30 - 26.45}{\sqrt{10.60}} = 1.09$$

which gives a one-tail *P*-value of 0.14, or a two-tail *P*-value of 0.28. Note that since tolbutamide has been considered a preventive of the complications of diabetes, departures from the null value were expected to occur in the direction of preventing death rather than in the opposite direction. Therefore, a one-tail *P*-value should technically be the lower tail of the distribution, in the direction of prevention, rather than the upper tail. Since the data demonstrate a positive association between tolbutamide and death, the one-tail *P*-value should be 1 − 0.14 or 0.86. The two-tail *P*-value is 0.28 whichever the direction of the prior expectation about departures from the null value.

If the Mantel-Haenszel test is applied to the sparse data in Example 12-5, the test statistic is

$$\chi = \frac{4 - \left[\frac{(12)(107)}{1175} + \frac{(6)(4)}{95}\right]}{\sqrt{\frac{(107)(1068)(12)(1163)}{(1175)^2(1174)} + \frac{(6)(89)(4)(91)}{(95)^2(94)}}} = 2.41$$

which gives $P_{(1)} = 0.008$, a result that is considerably different from the exact mid-$P$ value of 0.023. The discrepancy is not surprising in view of the small numbers and the striking asymmetry of the distribution, in which more than half of the probability distribution corresponds to the two most extreme outcomes out of the 17 possibilities for the number of exposed cases.

*Confidence Intervals for Pooled Estimates of Effect*

Confidence intervals for pooled estimates of effect can be calculated exactly from the statistical models adopted to describe the variability of the data, or they can be calculated approximately from asymptotic formulas. The exact calculations are exceedingly complicated and increase quickly in difficulty as the number of observations increases. Nevertheless, ready availability of microcomputers now makes it convenient to calculate exact confidence limits for pooled effect estimates in many applications, since the programming and memory requirements are not great; the calculation time may be long even with a computer, but the cost of computer time for such applications is becoming negligible. In view of the relatively large effort expended on the collection and processing of epidemiologic data, it seems worthwhile to obtain exact confidence limits for sparse data, even if stratified, if the means to do so are at hand. Consequently, exact formulas for confidence limits are presented in the following discussion whenever applicable.

For most situations, on the other hand, it will be preferable to use the straightforward and convenient noniterative approximate formulas for the calculation of confidence limits. The choice of an approximate formula for interval estimation generally depends on the type of point estimator used, since the variance approximation depends on how the point estimate is calculated. Therefore, the description of types of approximate confidence limits is presented according to the types of pooled point estimators described earlier in this chapter.

CONFIDENCE INTERVALS FOR STRATIFIED PERSON-TIME DATA

*Incidence Rate Difference.* No method exists for obtaining exact confidence limits for incidence rate difference because the total number of cases is not independent of the rate difference. Approximate confidence limits can be obtained in several ways according to the method of point estimation.

DIRECTLY WEIGHTED POINT ESTIMATE. The basic approach relies on the general statistical rule that the variance of a sum of independent random variables is the sum of the variance of each random variable. Since the directly pooled estimator for incidence rate difference is a sum of random variables (the stratum-specific estimates of incidence rate difference) multiplied by a constant (the weight for pooling), an overall variance for the pooled estimator would be



$$\text{Var}(\text{pooled } \hat{\text{IRD}}) \doteq \text{Var}\left[\sum_i \frac{w_i}{\sum_i w_i}(\hat{\text{IRD}}_i)\right] \doteq \sum_{i=1}^{N} \frac{w_i^2}{\left(\sum_i w_i\right)^2}\text{Var}(\hat{\text{IRD}}_i)$$

The weight is squared because any constant multiplier of a random variable is squared as a multiplier of variance. Each $w_i$ is taken as the inverse of the variance of $\hat{\text{IRD}}_i$ in pooling, so the overall variance is

$$\text{Var}(\text{pooled } \hat{\text{IRD}}) \doteq \sum_{i=1}^{N} \frac{w_i}{\left(\sum_i w_i\right)^2} = \frac{\sum_i w_i}{\left(\sum_i w_i\right)^2} = \frac{1}{\sum_i w_i} \qquad [12\text{-}39]$$

This variance can be used with the pooled estimator and formula 10-2 to compute approximate confidence limits.

Consider the pooled estimate of incidence rate difference for the data in Example 12-2, $5.95 \times 10^{-4}\text{yr}^{-1}$. From equation 12-39, the variance of the estimate is approximately the inverse of the sum of the weights, or

$$1/(65.16 \times 10^6\text{yr}^2) = 1.535 \times 10^{-8}\text{yr}^{-2}$$

which gives a standard deviation of

$$\sqrt{1.535 \times 10^{-8}\text{yr}^{-2}} = 1.239 \times 10^{-4}\text{yr}^{-1}$$

A 90 percent confidence interval for the pooled estimate is obtainable as

$$5.95 \times 10^{-4}\text{yr}^{-1} \pm 1.645\,(1.239 \times 10^{-4}\text{yr}^{-1})$$
$$= 3.9 \times 10^{-4}\text{yr}^{-1}, 8.0 \times 10^{-4}\text{yr}^{-1}$$

A second method of obtaining approximate confidence limits for incidence rate difference is to compute test-based limits from the point estimate and the $\chi$ from formula 12-33, using formula 10-6. For the data in Example 12-2, the $\chi$ is 3.319, giving, for 90 percent confidence limits,

$$\hat{\text{IRD}}(1 \pm Z/\chi) = 5.95 \times 10^{-4}\text{yr}^{-1}\,(1 \pm 1.645/3.319)$$
$$= 3.0 \times 10^{-4}\text{yr}^{-1}, 8.9 \times 10^{-4}\text{yr}^{-1}$$

a wider result than that obtained above. The test-based approach gives wide results because it does not assign an extremely heavy weight to the youngest age stratum as the direct approach does; the small numbers in the youngest stratum result in a small variance for the incidence rate difference estimated from that stratum.

MAXIMUM LIKELIHOOD POINT ESTIMATE. The maximum likelihood estimation of $\widehat{IRD}$ requires the simultaneous maximum likelihood estimation of the nuisance parameters $R_{0i}$ for each stratum. These fitted or smoothed estimates of $R_{0i}$ can be used in conjunction with the pooled estimate of IRD to get an improved estimate of the variance for the incidence rate difference in each stratum by substituting $\hat{R}_{0i}$ for $b_i/N_{0i}$ and $\hat{R}_{0i} + \widehat{IRD}$ for $a_i/N_{1i}$ in formula 12-1. The improved variance estimate is

$$\widehat{Var}(\widehat{IRD}_i) = \frac{\hat{R}_{0i} + \widehat{IRD}}{N_{1i}} + \frac{\hat{R}_{0i}}{N_{0i}} \qquad [12\text{-}40]$$

The overall variance of the pooled maximum likelihood estimate can be obtained by taking

$$w_i = \frac{N_{1i}N_{0i}}{N_{0i}(\hat{R}_{0i} + \widehat{IRD}) + N_{1i}(\hat{R}_{0i})} \qquad [12\text{-}41]$$

which is the reciprocal of the variance in formula 12-40; these weights are then used to get the overall variance as in equation 12-39.

For the data in Example 12-2, the weights estimated according to formula 12-41 are, from the youngest to the oldest age categories, $57.6 \times 10^6\text{yr}^2$, $4.9 \times 10^6\text{yr}^2$, $0.7 \times 10^6\text{yr}^2$, $0.7 \times 10^6\text{yr}^2$, and $0.2 \times 10^6\text{yr}^2$. The sum of these weights is $64.1 \times 10^6\text{yr}^2$, and therefore the variance is taken as

$$\widehat{Var}(\widehat{IRD}) = \frac{1}{64.1 \times 10^6\text{yr}^2} = 1.56 \times 10^{-8}\text{yr}^{-2}$$

which is close to the value of $1.53 \times 10^{-8}\text{yr}^{-2}$ obtained from the directly pooled weights. A 90 percent confidence interval for the maximum likelihood estimate of $5.91 \times 10^{-4}\text{yr}^{-1}$ is obtained as

$$5.91 \times 10^{-4}\text{yr}^{-1} \pm 1.645\left(\sqrt{1.56 \times 10^{-8}\text{yr}^{-2}}\right)$$
$$= 3.9 \times 10^{-4}\text{yr}^{-1}, 8.0 \times 10^{-4}\text{yr}^{-1}$$

which is virtually identical to the limits obtained from the weights used in direct pooling.

*Incidence Rate Ratio.* Confidence limits for the pooled estimate of the ratio of incidence rates from stratified data can be obtained by exact computation or by approximate methods. The exact computation requires iterative calculation of a complicated sum of probabilities and consequently requires a computer.



EXACT CONFIDENCE LIMITS. Obtaining the exact limits necessitates expressing the tail probability for the observations in terms of the incidence rate ratio (IRR). In stratum i, let the probability that a case is exposed be $u_i$. From formula 11-9 or 11-10 the following relation can be derived:

$$u_i = \frac{N_{0i}(IRR)}{N_{0i}(IRR) + N_{1i}} = \frac{IRR}{IRR + \dfrac{N_{0i}}{N_{1i}}} \qquad [12\text{-}42]$$

Exact Fisher limits for $\underline{IRR}$ and $\overline{IRR}$ can be determined from formula 12-42 and the following modification of formula 12-29:

$$\alpha/2 = \sum_{k=a}^{M_1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} (\underline{u}_i)^{k_i} (1 - \underline{u}_i)^{M_{1i}-k_i} \qquad [12\text{-}43]$$

and

$$1 - \alpha/2 = \sum_{k=a+1}^{M_1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} (\overline{u}_i)^{k_i} (1 - \overline{u}_i)^{M_{1i}-k_i} \qquad [12\text{-}44]$$

in which $k_i$ represents the number of exposed cases in stratum i, $k = \Sigma k_i$, $a = \Sigma a_i$, and $M_1 = \Sigma M_{1i}$. The value for IRR that satisfies equations 12-42 and 12-43 corresponds to the Fisher exact lower confidence bound; using equation 12-44 instead of 12-43 gives the Fisher exact upper confidence bound. Mid-*P* exact limits are obtainable by allotting only half the probability for $k = a$ into the tail:

$$\alpha/2 = \frac{1}{2}\sum_{k=a} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} (\underline{u}_i)^{k_i} (1 - \underline{u}_i)^{M_{1i}-k_i} + \sum_{k=a+1}^{M_1} \prod_{i=1}^{N} \binom{M_{1i}}{k_i} (\underline{u}_i)^{k_i} (1 - \underline{u}_i)^{M_{1i}-k_i} \qquad [12\text{-}45]$$

gives the lower bound, and substituting $1 - \alpha/2$ for $\alpha/2$ will give the upper bound.

With the help of a computer, exact 90 percent confidence limits for the incidence rate ratio using the data in example 12-6 can be calculated. The Fisher-type 90 percent interval is 2.43 to 18.0, and the mid-*P* 90 percent interval is 2.71 to 16.0.

APPROXIMATE CONFIDENCE LIMITS. The formulation of the approximate limits for the three different types of point estimates are as follows:

1. *Directly Weighted Point Estimate.* The variance of the directly pooled incidence rate ratio should be estimated after a logarithmic transforma-

tion, so that the limits can be set on the logarithmic scale. The variance formula resembles 12-39:

$$\text{Var}[\ln(\text{pooled } \widehat{\text{IRR}})] = \frac{1}{\sum_i w_i} \qquad [12\text{-}46]$$

where

$$w_i = \frac{a_i b_i}{a_i + b_i} \qquad [12\text{-}47]$$

For the data in Example 12-3, $w_1 = 140/24 = 5.83$, $w_2 = 9196/197 = 46.7$, and

$$\text{Var}[\ln(\text{pooled } \widehat{\text{IRR}})] = \frac{1}{5.83 + 46.7} = 0.019$$

The directly pooled point estimate is 1.50, giving a 90 percent confidence interval of

$$\exp[\ln(1.50) \pm 1.645\sqrt{0.019}] = 1.20, 1.88$$

2. *Maximum Likelihood Point Estimate.* The asymptotically efficient maximum likelihood estimator of the incidence rate ratio has a variance estimate of

$$\text{Var}(\widehat{\text{IRR}}) = \frac{\widehat{\text{IRR}}}{\sum_{i=1}^{N} \frac{M_{1i}N_{0i}/N_{1i}}{\left(\widehat{\text{IRR}} + \frac{N_{0i}}{N_{1i}}\right)^2}} \qquad [12\text{-}48]$$

It is necessary to divide the above by $(\widehat{\text{IRR}})^2$ to approximate the variance of $\ln(\widehat{\text{IRR}})$:

$$\text{Var}[\ln(\widehat{\text{IRR}})] = \frac{1}{(\widehat{\text{IRR}})\sum_{i=1}^{N} \frac{M_{1i}N_{0i}/N_{1i}}{\left(\widehat{\text{IRR}} + \frac{N_{0i}}{N_{1i}}\right)^2}} \qquad [12\text{-}49]$$

The above formula is identical to formula 12-46 if $a_i/b_i$ is replaced by $(\widehat{\text{IRR}})N_{1i}/N_{0i}$ in the weight given in formula 12-47.

For the data in Example 12-3, the maximum likelihood estimate of IRR is 1.50, and the variance of $\ln(\widehat{\text{IRR}})$ is



$$\frac{1}{(1.50)\dfrac{24(1701/1516)}{(1.50 + 1701/1516)^2} + \dfrac{197(2245/949)}{(1.50 + 2245/949)^2}} = 0.019$$

which gives a 90 percent confidence interval of

$$\exp[\ln(1.50) \pm 1.645\sqrt{0.019}] = 1.20, 1.88$$

This interval estimate is virtually identical to that obtained by the directly weighted approach.

For the data in Example 12-6, the approximate 90 percent confidence interval around the maximum likelihood estimate of IRR, which is 6.55, is 2.77 to 15.5, which agrees quite well with the 90 percent exact mid-*P* interval of 2.71 to 16.0.

3. *Mantel-Haenszel Point Estimate.* The Mantel-Haenszel estimator for incidence rate ratio (formula 12-27) can be considered a weighted average of stratum-specific estimates of the incidence rate ratio with weights equal to $b_iN_{1i}/T_i$ and the approximate confidence limits calculated on this basis. A more stable formula for the variance, however, can be obtained by considering each $a_i$ and $b_i$ to be an independent Poisson variate [Tarone, 1981], or by considering each $a_i$ to be an independent binomial variate conditional on $N_{1i}$ [Greenland and Robins, 1985]. The latter approach yields

$$\text{Var}[\ln(\widehat{\text{IRR}}_{MH})] \doteq \frac{\sum_{i=1}^{N} M_{1i}N_{1i}N_{0i}/T_i^2}{\left[\sum_{i=1}^{N} \frac{a_iN_{0i}}{T_i}\right]\left[\sum_{i=1}^{N} \frac{b_iN_{1i}}{T_i}\right]} \qquad [12\text{-}50]$$

For Example 12-3, the above formula gives $\text{Var}[\ln(\widehat{\text{IRR}}_{MH})] = 0.019$, the same as the result obtained using the directly weighted procedure above. The resulting confidence interval, 1.20–1.88, is likewise identical to the interval for the directly weighted point estimate.

Test-based limits for the data in Example 12-3 can also be obtained for the Mantel-Haenszel point estimate (1.50) and the $\chi$ statistic from formula 12-33 (2.94):

$$1.50^{(1 \pm 1.645/2.94)} = 1.19, 1.87$$

These test-based limits are in close agreement with the results obtained from the other approaches. Test-based limits can also be obtained for the directly weighted point estimate.

For the data in Example 12-6, for which $\widehat{\text{IRR}}_{MH} = 7.30$, the variance

For the data in Example 12-6, for which $IRR_{MH} = 7.30$, the variance calculated from formula 12-50 is 0.344, which gives a 90 percent confidence interval for $IRR_{MH}$ of 2.8 to 19. For the same data, the test-based 90 percent confidence limits can be calculated from the $\chi$ of 3.98 as

$$7.30^{(1 \pm 1.645/3.98)} = 3.2, 17$$

Considering the small numbers involved, both of these approximate intervals are in reasonably good agreement with the mid-$P$ exact 90 percent confidence interval of 2.7 to 16.

CONFIDENCE INTERVALS FOR STRATIFIED CUMULATIVE INCIDENCE DATA

*Risk Difference.* Because of the nuisance parameter $\hat{R}_{0i}$ in each stratum, no approach exists for obtaining exact confidence limits for a pooled risk difference or prevalence difference. Approximate confidence limits can be obtained by methods analogous to those described above for incidence rate data.

DIRECTLY WEIGHTED POINT ESTIMATE. The variance of the pooled risk difference can be expressed in terms of the stratum-specific weights in the same way as that used for incidence rate data (formula 12-39):

$$\text{Var}(\text{pooled } \widehat{RD}) = \frac{1}{\Sigma w_i} \qquad [12\text{-}51]$$

where the weights are those given in formula 12-7:

$$w_i = \frac{N_{1i}^3 N_{0i}^3}{N_{0i}^3 a_i (N_{1i} - a_i) + N_{1i}^3 b_i (N_{0i} - b_i)}$$

The square root of the above variance estimate in formula 12-51 can be used with formula 10-2 to obtain approximate confidence limits. For Example 12-4, the stratum-specific weights are 1,009 and 280 (Table 12-5), so that $\Sigma w_i = 1{,}289$ and the variance estimate is $1/1{,}289 = 0.000776$. The square root is 0.028, so that a 90 percent confidence interval estimate, using the weighted point estimate of 0.034, would be

$$0.034 \pm 1.645(0.028) = -0.012, 0.080$$

An alternative is to use test-based confidence limits (formula 10-6), based on the $\chi$ from formula 12-38. For the data in Example 12-4, the $\chi$ is 1.09 and the test-based confidence interval is

$$0.034(1 \pm 1.645/1.09) = -0.017, 0.086$$

which is slightly wider than the result using the estimates of stratum-specific variance.



MAXIMUM LIKELIHOOD POINT ESTIMATE. Again, the maximum likelihood solutions for the pooled estimate $\widehat{RD}$ and the unexposed risk $\hat{R}_{0i}$ in each stratum can be used to improve the variance estimation for the rate difference in each stratum. The improved estimates can be obtained by substituting $\hat{R}_{0i}$ for $b_i/N_{0i}$ and $\hat{R}_{0i} + \widehat{RD}$ for $a_i/N_{1i}$; these substitutions can be made directly into formula 12-7, giving the improved weights

$$w_i = \frac{N_{1i} N_{0i}}{N_{0i}(\hat{R}_{0i} + \widehat{RD})(1 - \hat{R}_{0i} - \widehat{RD}) + N_{1i}\hat{R}_{0i}(1 - \hat{R}_{0i})} \qquad [12\text{-}52]$$

which can be used in formula 12-51 to get a variance estimate for the maximum likelihood estimate of RD.

For the data in Example 12-4, the weights calculated from formula 12-52 are 1,008 and 280 for strata 1 and 2, respectively, which gives a variance of $1/(1{,}008 + 280) = 0.000776$. Note that the weights and variance estimate are nearly identical to the results obtained from the noniterative directly weighted procedure because the number of observations within each stratum is large. The resulting 90 percent confidence interval is $-0.012$, 0.080 as it was with the directly weighted approach.

*Risk Ratio.* Exact confidence limits for risk ratio are not calculable, since the likelihood equation contains the nuisance parameters $\hat{R}_{0i}$ for each stratum $i$. If risks are small, however, the odds ratio measure may be used to approximate the risk ratio. Since the odds ratio can be estimated without nuisance parameters, the likelihood can be expressed conditionally on all the margins of the $2 \times 2$ table, allowing the calculation of exact confidence limits for the pooled odds ratio. This procedure is described below for case-control data.

Approximate confidence limits for pooled estimates of the risk ratio can be obtained for directly weighted, Mantel-Haenszel, or maximum likelihood point estimators.

DIRECTLY WEIGHTED POINT ESTIMATE. As usual, approximate confidence limits for ratio measures should be set on the logarithmic scale. Formulas 12-46 and 12-10 can be used to obtain the variance of the logarithm of the pooled risk ratio estimate. For the data in Example 12-4, the stratum-specific weights, given in Table 12-6, are 3.25 and 11.6. The $\Sigma w_i = 14.85$ and the variance of the logarithmically transformed point estimate is $1/14.85 = 0.0673$. The weighted average of the logarithms of the stratum-specific estimates of the risk ratio is 0.270, which is the antilogarithm of the pooled estimate of the risk ratio, 1.31. Approximate 90 percent confidence limits can be set as follows:

$$0.270 \pm 1.645 (\sqrt{0.0673}) = -0.157, 0.697$$

$$e^{-0.157}, e^{0.697} = 0.86, 2.0$$

MAXIMUM LIKELIHOOD POINT ESTIMATE. Formula 12-10 can be improved by substituting $N_{1i}\hat{R}_{0i}\hat{RR}$ for $a_i$ and $N_{0i}\hat{R}_{0i}$ for $b_i$, where $\hat{R}_{0i}$ and $\hat{RR}$ are the fitted maximum likelihood estimates. The improved weights are

$$w_i = \frac{(\hat{RR})\hat{R}_{0i}N_{0i}N_{1i}}{(\hat{RR})(1 - \hat{R}_{0i})N_{1i} + (1 - \hat{R}_{0i}\hat{RR})N_{0i}} \qquad [12\text{-}53]$$

which may be used in formula 12-46 to get an estimate for the pooled variance. For Example 12-4, the maximum likelihood point estimate of the risk ratio is 1.31; $\hat{R}_{01} = 0.0504$, $\hat{R}_{02} = 0.1781$, and the improved weights are 3.44 and 11.4. The $\Sigma w_i$ for these improved weights is 14.83, and the variance of the logarithmically transformed point estimate is 1/14.83 = 0.0674, nearly the same result as that obtained from the stratum-specific variance estimates. The 90 percent confidence interval is obtained on the log scale as

$$\ln(1.31) \pm 1.645\ (\sqrt{0.0674}) = -0.156, 0.698$$

and the actual limits are

$$e^{-0.156}, e^{0.698} = 0.86, 2.0$$

MANTEL-HAENSZEL POINT ESTIMATE. The Mantel-Haenszel point estimator of the risk ratio from follow-up data with count denominators takes the same form as the point estimator for the rate ratio with person-time denominators (formula 12-27). The variance for the logarithm of $RR_{MH}$ is approximately [Greenland and Robins, 1985]

$$\mathrm{Var}[\ln(\hat{RR}_{MH})] \doteq \frac{\sum_{i=1}^{N}(M_{1i}N_{1i}N_{0i} - a_ib_iT_i)/T_i^2}{\left[\sum_{i=1}^{N}\frac{a_iN_{0i}}{T_i}\right]\left[\sum_{i=1}^{N}\frac{b_iN_{1i}}{T_i}\right]} \qquad [12\text{-}54]$$

For the data in Example 12-4, the above expression gives $\mathrm{Var}[\ln(\hat{RR}_{MH})] \doteq 0.0671$; coupled with the point estimate of $RR_{MH} = 1.33$, the approximate 90 percent confidence limits are

$$\exp(\ln(1.33) \pm 1.645\ \sqrt{0.0671}) = 0.87, 2.0$$

which are nearly identical to the limits calculated for the directly weighted point estimate and the maximum likelihood point estimate.



CONFIDENCE INTERVALS FOR THE ODDS RATIO FROM STRATIFIED CASE-CONTROL (OR PREVALENCE) DATA

Confidence limits for the pooled estimate of the odds ratio from stratified 2 × 2 tables can be obtained by exact computation or by approximate methods. The exact computation is exceedingly complex for any but the most sparse data and requires a computer program [Thomas, 1975].

*Exact Confidence Limits.* The expression for the probability of the observations in a single set of N 2 × 2 tables, conditional on the marginal totals for each 2 × 2 table and the odds ratio, is

$$\Pr(\text{data}) = \prod_{i=1}^{N} \frac{\binom{N_{1i}}{a_i}\binom{N_{0i}}{b_i}OR^{a_i}}{\sum_{k=\max(0,M_{1i}-N_{0i})}^{\min(M_{1i},N_{1i})}\binom{N_{1i}}{k}\binom{N_{0i}}{M_{1i}-k}OR^{k}} \qquad [12\text{-}55]$$

Tail probabilities for exact confidence limits can be calculated by taking the sum of the probabilities calculated in expression 12-55 for all possible values of $\Sigma a_i$ equal to or greater than the actual value observed, and for all possible combinations of cell frequencies that yield a given value for $\Sigma a_i$. To get mid-*P* exact limits, only one-half the probability determined by expression 12-55 should be added to the summation for every possible combination in which $\Sigma a_i$ equals the observed value. The exact lower confidence limit is obtained by determining through trial and error the value of the odds ratio that produces an upper-tail probability of $\alpha/2$ ($\alpha$ equals the complement of the desired confidence level). The upper confidence limit is obtainable by summing over all values of $\Sigma a_i$ that are less than or equal to the observed value and finding the value of the odds ratio that gives a lower-tail probability of $\alpha/2$.

For the data in Example 12-5, the observed value for $\Sigma a_i$ is 4; the 55 combinations for which $\Sigma a_i \geq 4$ are listed in Table 12-9. Using expression 12-55 to determine the contributions to the tail probability, 90 percent exact mid-*P* confidence limits are found to be 1.30, 9.78. The corresponding Fisher limits are 1.09, 10.9.

*Approximate Confidence Limits.* DIRECTLY WEIGHTED POINT ESTIMATE. Formulas 12-46 and 12-13 can be used to obtain approximate confidence limits for the directly weighted pooled estimate of the odds ratio. As usual with a ratio measure, the limits are first set on the logarithmic scale and then translated back to the original scale. For the data in Example 12-5, using the weights indicated in Table 12-7, the sum of the weights is 2.85,

nd the estimated variance of the logarithm of the pooled odds ratio is 1/ .85 = 0.351. Approximate 90 percent confidence limits are

$$\exp[\ln(3.82) \pm 1.645\sqrt{0.351}] = 1.44, 10.1$$

which differ somewhat from the exact limits, but the discrepancy is tolerble, especially considering the width of the interval. In view of the small umber of cases in the analysis, the approximation seems reasonably ood.

MAXIMUM LIKELIHOOD POINT ESTIMATE. With maximum likelihood point estimation, fitted cell entries in the 2 × 2 tables can be used to derive an stimate of the variance. The values for $\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$, and $\hat{d}_i$ that satisfy equation 2-23 or equation 12-24 and the marginal totals of each 2 × 2 table can e substituted in equation 12-13:

$$w_i = \frac{1}{\frac{1}{\hat{a}_i} + \frac{1}{\hat{b}_i} + \frac{1}{\hat{c}_i} + \frac{1}{\hat{d}_i}} \qquad [12\text{-}56]$$

he variance of the logarithmically transformed odds ratio point estimate s $1/\Sigma w_i$.

For the data in Example 12-5, the cell frequencies for the unconditional naximum likelihood estimate, satisfying equation 12-24, are given in Table 2-10. From these fitted cell frequencies, $w_1 = 2.31$ and $w_2 = 0.54$, which ives $\Sigma w_i = 2.85$ and a variance of 0.351. Approximate 90 percent confidence limits are

$$\exp[\ln(3.79) \pm 1.645\sqrt{0.351}] = 1.43, 10.0$$

Using equation 12-23 rather than 12-24 to calculate the fitted frequencies according to the conditional likelihood is considerably more difficult; for hese data the a cells for the two strata using equation 12-23 are 3.238 and 0.762, which are nearly identical to the unconditional values in Table 12-10 and produce the same approximate confidence interval. Because the computation necessary to get the conditional fitted cell entries from the iterative solution of equations 12-22 and 12-23 is difficult, it is easier to calculate the exact confidence limits instead.

Another approach, which was proposed initially by Cornfield [1956] for a single 2 × 2 table, was extended by Gart [1971] for a set of 2 × 2 tables. The approximate lower limit is the solution to the equation

$$Z_{\alpha/2} = \frac{\Sigma a_i - \Sigma E_i}{\sqrt{\text{Var}(\Sigma a_i)}} \qquad [12\text{-}57]$$



*Table 12-10. Fitted maximum likelihood cell entries for the data in example 12-5; pooled estimate of odds ratio is 3.79*

| | Maternal age < 35 | | | Maternal age 35+ | | |
|---|---|---|---|---|---|---|
| | Spermicide use | | | Spermicide use | | |
| | + | − | Total | + | − | Total |
| Down syndrome | 3.247 | 8.753 | 12 | 0.753 | 3.247 | 4 |
| Control | 103.753 | 1059.247 | 1163 | 5.247 | 85.753 | 91 |
| Total | 107 | 1068 | 1175 | 6 | 89 | 95 |

where $E_i$ is the expected value of $a_i$ conditional on the value of the odds ratio at the lower boundary of the confidence interval,

$$\underline{R} = \frac{E_i(M_{0i} - N_{1i} + E_i)}{(M_{1i} - E_i)(N_{1i} - E_i)}$$

$Z_{\alpha/2}$ is the value of the standard normal statistic that corresponds to the desired level of confidence, and

$$\text{Var}(\Sigma a_i) = \sum_{i=1}^{N}\left[\frac{1}{E_i} + \frac{1}{M_{1i} - E_i} + \frac{1}{N_{1i} - E_i} + \frac{1}{M_{0i} - N_{1i} + E_i}\right]^{-1}$$

To obtain the upper bound to the interval, $Z_{\alpha/2}$ is replaced by $-Z_{\alpha/2}$. Equation 12-57 must be solved iteratively for each limit. In principle this method has the advantage of approximating the variance using the cell frequencies that correspond to the confidence limit value of the odds ratio rather than the point estimate.

For the data in Example 12-5, the fitted cell frequencies for the a cell using equation 12-57 to obtain the lower limit of a 90 percent confidence interval are 1.542 for the young stratum and 0.356 for the older stratum; note that these do not add to $\Sigma a_i = 4$ as in point estimation, though the fitted cell frequencies within each table still conform to the marginal totals of the individual table. The lower confidence limit satisfying equation 12-57 is 1.48. For the upper limit, a separate iterative solution is required to get the fitted frequencies for the a cell of 5.772 and 1.370 for the young and old strata, respectively. The upper confidence limit corresponding to these values is 9.72.

MANTEL-HAENSZEL POINT ESTIMATE. The statistical properties of the Mantel-Haenszel estimator of the odds ratio have been elaborated under two different limiting situations: Either the number of subjects per stratum becomes large, or the number of strata becomes large with few subjects per stratum [Hauck, 1979; Breslow, 1981]. Variance formulas have been pro-

posed for the Mantel-Haenszel odds ratio estimator for each of these limiting situations; Breslow and Liang [1982] proposed weighting the two formulas to derive a combined formula that is generally applicable. More recently, Robins and coauthors [1986] have developed a single variance formula that should be generally applicable for the Mantel-Haenszel odds ratio estimator:

$$\text{Var}[\ln(\widehat{OR}_{MH})] \doteq \frac{\sum_{i=1}^{N} P_i R_i}{2\left[\sum_{i=1}^{N} R_i\right]^2} + \frac{\sum_{i=1}^{N} (P_i S_i + Q_i R_i)}{2\sum_{i=1}^{N} R_i \sum_{i=1}^{N} S_i} + \frac{\sum_{i=1}^{N} Q_i S_i}{2\left[\sum_{i=1}^{N} S_i\right]^2} \quad [12\text{-}58]$$

where

$$P_i = (a_i + d_i)/T_i$$
$$Q_i = (b_i + c_i)/T_i$$
$$R_i = a_i d_i / T_i$$

and

$$S_i = b_i c_i / T_i$$

For the data in Example 12-5, the above formula gives an estimated variance of 0.349, which yields a 90 percent confidence interval of 1.43 to 10.0, a result that is nearly identical to the limits for the directly weighted and maximum likelihood point estimates.

Test-based 90 percent confidence limits for the data in Example 12-5 are obtained as

$$3.78^{(1 \pm 1.645/2.41)} \doteq 1.53, 9.37$$

which are narrower than the limits obtained from formula 12-58. The test-based approach for approximate confidence limits can also be used with the directly weighted point estimate.

## VALUATION AND DESCRIPTION OF EFFECT MODIFICATION

The techniques for deriving a pooled estimate of an effect that is uniform across categories of a third variable should not be applied if it appears unreasonable to assume that the effect is indeed uniform. When an effect is believed to vary across strata—that is, when effect modification is presumed to exist—the focus of data analysis and presentation should shift from the control of confounding to a description of how the effect is modified by the stratification factor. It is important to realize that confounding, when present, is manifest only in the crude measure of effect; when effect modification is present in the data, none of the options for describing the



effect involves the crude measure, so the issue of confounding is superseded by the description of effect modification. Determining whether effect modification is present in the data is clearly an important decision that should be addressed in every stratified analysis.

It is important to inject a note of caution about the methodology for assessing effect modification. The evaluation of effect modification often appears to rest on a seemingly mechanical application of statistical tests. The epidemiologic issues of interaction underlying the statistical evaluation of effect modification are subtle and can become muddled in a purely mechanical approach. These issues are discussed in Chapter 15, which supplies an epidemiologic perspective for the statistical methods described in the following section.

In addition to the epidemiologic considerations, there are statistical considerations that warrant a cautious approach to the statistical evaluation of effect modification. The more general statistical tests for effect modification have low power because the alternatives to the null hypothesis that they test are not very specific. As a result, "nonsignificant" *P*-values are even more difficult to interpret correctly. Furthermore, given the many influences of selection biases, misclassification, confounding, and other biases as well as causal effects, it is seldom that one would expect any effect to be precisely uniform for any scale of measurement. Thus, the null hypothesis of a uniform effect often amounts to no more than a statistical contrivance that at best should be accepted as only an approximation to reality and generally should be regarded with skepticism.

### *Evaluation of Effect Modification*

The first step in evaluating effect modification is to inspect the stratum-specific estimates of effect. While some random variability in stratum-specific estimates is to be expected even when the underlying parameter is uniform, excessive variability or obvious nonrandom patterns of variation may be evident on inspection. The investigator's judgment about effect modification should not be limited to the appearance of the data in hand; when it is available, outside knowledge from previous studies or more general biologic insight should be integrated into the evaluation process.

Typically, however, outside knowledge is scant, and investigators will desire a more formal statistical evaluation of the extent to which the variability of stratum-specific estimates of effect is consistent with purely random behavior. Toward this end, a variety of statistical tests can be applied. Part of the variety derives from the fact that ratio and difference measures require separate evaluations for effect modification, since uniformity of the ratio measure usually implies effect modification of the difference measure and vice versa. The use of statistical tests has been discussed in Chapter 9, especially with regard to assessing "statistical significance," which trivializes the interpretation of otherwise meaningful measures. The use of significance tests is more defensible, however, when an immediate decision

rests on the outcome of a single statistical evaluation. Such may be the case if an investigator is attempting to decide whether the extent of variability in a set of stratum-specific estimates of effect is consistent with the random variation of a uniform effect or, alternatively, whether there is effect modification in the data.

Statistical tests of the null hypothesis that the effect is uniform (i.e., exhibits no effect modification) generally are of two types, one based on a directly pooled estimate of uniform effect and the other on a maximum likelihood estimate.

For the directly pooled estimates, the basic principle of the test is to compare each stratum-specific estimate with the pooled estimate, square the difference, and divide by the variance of the stratum-specific effect estimate. The resulting quotient is summed over all strata, yielding a chi-square statistic with degrees of freedom equal to one less than the number of strata:

$$\chi^2_{N-1} = \sum_{i=1}^{N} \frac{(\hat{R}_i - \hat{R})^2}{\mathrm{Var}(\hat{R}_i)} \qquad [12\text{-}59]$$

For difference measures of effect, the above formula can be applied directly, with $\hat{R}_i$ denoting the stratum-specific difference and $\hat{R}$ denoting the directly pooled estimate of effect. The stratum-specific variances in the denominator are the reciprocals of the weights used to obtain the pooled estimate. For ratio measures of effect, it is desirable to use a logarithmic transformation:

$$\chi^2_{N-1} = \sum_{i=1}^{N} \frac{[\ln(\hat{R}_i) - \ln(\hat{R})]^2}{\mathrm{Var}[\ln(\hat{R}_i)]} \qquad [12\text{-}60]$$

in which $\hat{R}_i$ now denotes the stratum-specific ratio estimate of effect and $\hat{R}$ denotes the directly pooled ratio measure.

As an example of the application of the above test for effect modification of the incidence rate difference, consider the stratified effect estimates presented in Table 12-2. The directly pooled estimate of the incidence rate difference is $5.95 \times 10^{-4}\mathrm{yr}^{-1}$. Using the stratum-specific point estimates and their variances in Table 12-2, formula 12-59 gives a $\chi^2$ of 8.38 with four degrees of freedom. From tables of the chi-square distribution, the corresponding two-tail *P*-value is 0.08, which indicates the degree of consistency of the data in Example 12-2 with the hypothesis that the incidence rate difference is constant across age categories.

For an illustration of formula 12-60 used to evaluate the heterogeneity of a ratio measure of effect, consider the data in Example 12-5 and the calculations derived from them in Table 12-7. The pooled estimate of the odds ratio is 3.8. Formula 12-60 applied to the stratum-specific estimates in Table 12-7 gives a $\chi^2$ with one degree of freedom of 0.14, which corresponds to a two-tail *P*-value of 0.7, showing that the data are consistent with the hypothesis of a uniform odds ratio.



The chi-square tests given in formulas 12-59 and 12-60, like the directly weighted pooled estimators on which they are based, depend on an assumption of large numbers of observations within strata. With small frequencies, the tests are unreliable. With zero cell frequencies, it may not even be possible to obtain stratum-specific variance or effect estimates. An alternative approach is to use a statistical test based on the maximum likelihood estimation of a uniform effect measure. This approach, termed a likelihood-ratio test, constructs the test statistic from a comparison of the likelihood equations for the data under two hypotheses: One hypothesis is that the effect is uniform, and the other is that the effect acquires a different value in each stratum. Although the likelihood-ratio test is also asymptotic, the requirement for large numbers within each stratum is not as stringent as it is in formulas 12-59 and 12-60; the test can be used even when there are small cell frequencies in the data. With zero cell frequencies the test fails, although it can be modified slightly by substituting a small positive value for zero to get a reasonably accurate result in many cases. The tests require previous calculation of the maximum likelihood estimate of a uniform effect, but otherwise require no iteration and involve only simple computation. The likelihood-ratio approach should give more accurate results in testing for effect modification when the data are relatively sparse.

The formulation of the likelihood-ratio tests for effect modification depends on the effect measure and the type of data under consideration. For incidence rate difference (IRD), the test is

$$\chi^2_{N-1} = -2\sum_{i=1}^{N}\left[a_i \ln\left(\frac{(\hat{IR}_{0i} + \hat{IRD})N_{1i}}{a_i}\right) + b_i \ln\left(\frac{\hat{IR}_{0i}N_{0i}}{b_i}\right)\right] \qquad [12\text{-}61]$$

Note that the pooled maximum likelihood estimate of IRD is part of the test formula, as is the maximum likelihood estimate of $IR_0$ in each stratum. The estimates of $IR_0$ must be obtained in the estimation of IRD, so no additional estimation beyond maximum likelihood point estimation is required to apply formula 12-61.

For the data in Example 12-2, the pooled maximum likelihood estimate of the incidence rate difference is $5.91 \times 10^{-4}\mathrm{yr}^{-1}$; stratum-specific maximum likelihood estimates of the incidence among nonsmokers are, from the youngest to the oldest, $8.406 \times 10^{-5}\mathrm{yr}^{-1}$, $1.640 \times 10^{-3}\mathrm{yr}^{-1}$, $6.303 \times 10^{-3}\mathrm{yr}^{-1}$, $1.352 \times 10^{-2}\mathrm{yr}^{-1}$, and $1.917 \times 10^{-2}\mathrm{yr}^{-1}$. Using these estimates in formula 12-61 gives a $\chi^2$ value of 7.4 with four degrees of freedom,

which corresponds to a two-tail *P*-value of 0.12. This value compares with the result of 0.08 from formula 12-59.

For incidence rate ratio, the likelihood ratio test of uniformity is

$$\chi^2_{N-1} = -2\sum_{i=1}^{N}\left[a_i\ln\left(\frac{\hat{IRR}\cdot M_{1i}}{a_i(\hat{IRR} + N_{0i}/N_{1i})}\right) + b_i\ln\left[\left(\frac{M_{1i}}{b_i}\right)\left(1 - \frac{\hat{IRR}}{\hat{IRR} + N_{0i}/N_{1i}}\right)\right]\right] \quad [12\text{-}62]$$

The formula requires the maximum likelihood estimate of IRR, but no nuisance parameters are involved.

For the data in Example 12-2, the maximum likelihood estimate of IRR is 1.42; the chi-square test in formula 12-62 gives a value of 12.1 with four degrees of freedom, which corresponds to a two-tail *P*-value of 0.016. Thus, these data are even less consistent with a uniform incidence rate ratio than they are with a uniform incidence rate difference. For the incidence rate ratio the stratum-to-stratum pattern of variation is extremely regular, decreasing steadily from the youngest to the oldest age category. The regular pattern of variation casts additional doubt on the validity of the assumption of a uniform incidence rate ratio.

For cumulative incidence difference, the likelihood-ratio test analogous to formula 12-61 is

$$\chi^2_{N-1} = -2\sum_{i=1}^{N}\left[a_i\ln\left(\frac{(\hat{R}_{0i} + \hat{RD})N_{1i}}{a_i}\right) + b_i\ln\left(\frac{\hat{R}_{0i}N_{0i}}{b_i}\right) + c_i\ln\left(\frac{(1 - \hat{R}_{0i} - \hat{RD})N_{1i}}{c_i}\right) + d_i\ln\left(\frac{(1 - \hat{R}_{0i})N_{0i}}{d_i}\right)\right] \quad [12\text{-}63]$$

which again involves not only the maximum likelihood estimate of the risk difference but also the nuisance parameters $\{R_{0i}\}$.

For the data in Example 12-4, the maximum likelihood estimate of the risk difference is 0.0343 under the uniformity assumption. Stratum-specific maximum likelihood estimates of the risk among unexposed persons are 0.0415 and 0.1892 in the young and old strata, respectively. The two stratum-specific estimates of risk difference are 0.0338 and 0.0363, showing extremely little variation. Accordingly, the $\chi^2$ from formula 12-63 is 0.001 with one degree of freedom, corresponding to a *P*-value of 0.97; this indicates the extraordinarily high consistency between the data and the statistical hypothesis of a uniform risk difference in these two age categories.

For cumulative incidence ratio, the likelihood-ratio test of uniformity is



$$\chi^2_{N-1} = -2\sum_{i=1}^{N}\left[a_i\ln\left(\frac{\hat{R}_{0i}(RR)N_{1i}}{a_i}\right) + b_i\ln\left(\frac{\hat{R}_{0i}N_{0i}}{b_i}\right) + c_i\ln\left(\frac{(1 - \hat{R}_{0i}RR)N_{1i}}{c_i}\right) + d_i\ln\left(\frac{(1 - \hat{R}_{0i})N_{0i}}{d_i}\right)\right] \quad [12\text{-}64]$$

which differs notably from the corresponding test for incidence rate ratio (formula 12-62). The difference derives from the fact that formula 12-62 is developed from a likelihood expression that is conditional on the total number of cases in each stratum, thereby eliminating the nuisance parameters. For cumulative incidence data, however, it is not correct to condition on the total number of cases in a stratum, and therefore an unconditional likelihood expression must be used; one consequence is that the estimates of the nuisance parameters $\{R_{0i}\}$ are part of the test statistic.

The data in Example 12-5 can be evaluated for uniformity of the risk ratio. The maximum likelihood estimate is 1.311, based on the assumption of uniformity. The nuisance parameter estimates (i.e., maximum likelihood estimates of the risk among unexposed for each stratum) are 0.0504 and 0.1781 in the younger and older categories, respectively. The one degree of freedom $\chi^2$ statistic from formula 12-64 is 0.452, which corresponds to a *P*-value of 0.5. Thus, the data are reasonably consistent with a uniform risk ratio despite the apparent variation in stratum-specific estimates of the risk ratio (1.8 and 1.2).

For case-control (or prevalence) data, the likelihood ratio test of uniformity of the odds ratio is

$$\chi^2_{N-1} = -2\sum_{i=1}^{N}\left[a_i\ln\left(\frac{\hat{a}_i}{a_i}\right) + b_i\ln\left(\frac{\hat{b}_i}{b_i}\right) + c_i\ln\left(\frac{\hat{c}_i}{c_i}\right) + d_i\ln\left(\frac{\hat{d}_i}{d_i}\right)\right] \quad [12\text{-}65]$$

where the fitted cell frequencies $\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$, and $\hat{d}_i$ are the values satisfying equation 12-24,

$$\hat{OR}_{ML} = \frac{\hat{a}_i\hat{d}_i}{\hat{b}_i\hat{c}_i}$$

For the data in Example 12-5, the (unconditional) maximum likelihood estimate of the odds ratio is 3.79. The fitted cell frequencies for the a cell are 3.2473 and 0.7527 for the younger and older strata, respectively; the other fitted cell frequencies are obtained from the margins of each 2 × 2 table by subtraction. Formula 12-65 yields a chi-square of 0.13 with one degree of freedom, which corresponds to a two-tail *P*-value of 0.7. These results are nearly identical to those obtained with formula 12-60.

Another test of uniformity of the odds ratio over a set of 2 × 2 tables

was proposed by Zelen [1971]. Zelen's test calls for summing the chi-square calculated for each 2 × 2 table (the square of formula 11-6 or 11-8) and subtracting from the sum the square of the Mantel-Haenszel chi-square (formula 12-38). Zelen's procedure, however, is not generally valid; counterexamples have been cited in which a uniform odds ratio gives a large chi-square for Zelen's test and a zero chi-square results when stratum-specific odds ratios differ considerably [Mantel et al., 1977]. This procedure is not recommended.

None of the tests considered in this section takes into account the pattern of variability of the effect estimates across strata. The chi-square values calculated in the application of these tests are independent of any ordering of the strata: If the strata were reordered, the test result would not differ. In principle, it is possible to construct more powerful tests directed at specific patterns of variation of the effect estimates over the strata as an alternative to uniformity. To do so, it would be necessary to postulate the pattern. For example, the likelihood test of uniformity of the incidence rate ratio for the data given in Example 12-2 produces a *P*-value of 0.016; the effect estimates decline nearly exponentially, however, so that a more powerful test of uniformity can be constructed using an exponential curve as the alternative pattern of variation. With this more powerful approach, a substantially smaller *P*-value results [Miettinen and Neff, 1971]. The improved test takes into account the declining pattern of the incidence rate ratio estimates with increasing age.

### *Description of Effect Modification*

When the stratum-specific estimates vary enough to indicate that there is likely to be variation in the underlying effect, it is improper to present either the crude estimate of effect or a pooled estimate. The pooled estimate of effect is a weighted average of the stratum-specific estimates, but the weights are intended to promote precision and therefore reflect the number of observations in individual strata. The pooled estimate is consequently potentially misleading unless it is reasonable to assume that the effect estimates vary only randomly around a uniform effect value. If the effect itself varies over strata, then the value of the pooled estimate calculated on the assumption of uniformity will depend on the distribution of subjects over strata in a way that is peculiar to the individual study and difficult to specify. The crude estimate is a worse alternative, since it does not even represent a weighted average of the stratum-specific estimates.

How, then, should the effect be described when the effect is judged to vary over strata? One simple approach is to present the estimates separately for each stratum. The study can be considered a set of individual substudies that are reported separately. Point estimates and confidence intervals can be reported for each stratum. This approach is often used when effect modification occurs for a dichotomous factor such as sex.



### STANDARDIZED EFFECT ESTIMATES

The drawback of reporting the results separately by stratum is that the overall body of data becomes divided, resulting in less precise estimates of effect in individual strata. If the stratification variable has many categories, there will be comparatively little precision for each of the several estimates. Furthermore, a set of many estimates of effect may offer a more detailed description of the effect than would a summary of the overall effect in a single number, but it also provides much less cogency in its detailed description. The entire purpose of data analysis is to reduce inherently complex information into a less complex and therefore more readily interpretable form. With this goal in mind, it is appropriate to consider whether there is not some meaningful way in which a set of stratum-specific estimates might be reduced into a single overall measure. The difficulty with the pooled estimate is the unpredictable or unspecifiable way in which it combines the information over strata. A reasonable way to avoid this difficulty is to combine the information over the strata using a specified system of weights—that is, to standardize the component rates of the effect measure to a standard distribution for the stratification variable. The advantage of standardization is that the weighting of stratum-specific information is easily specified, allowing the averaging of different values of the effect estimate from different strata to occur in a theoretically replicable and epidemiologically meaningful way.

Investigators may occasionally be cautioned to avoid standardization of effect estimates if there is "excessive" variability in the estimates over the strata, since that variability will be obscured in the overall estimate. For example, if effect estimates in two strata point in opposite directions, let us say indicating prevention for males and causation for females, a standardized estimate could indicate prevention, causation, or no effect depending on the choice of standard. This problem exists, however, for any single summary measure. A standardized summary measure at least has the advantage that it weights the divergent estimates in a definable way. It is incorrect to make uniformity of effect a prerequisite for standardization, with uniformity, pooling is preferable to standardization to optimize precision. Standardization is useful principally when the effect does vary over strata. Of course, it is always true that a summary measure can obscure an underlying variability. If the variability is extreme, as it often is when effect estimates point in opposite directions, it may be reasonable to report the stratum-specific details. In other instances, the investigator may properly decide that a summary result will convey enough of the intended message without obfuscating important detail to permit standardization. After all, there is no limit to the process of separating data into levels of detail; even small and apparently homogeneous subgroups represent the aggregate experience of some individuals who experienced the effect of interest and others who did not.

*ble 12-11. Incidence rate ratio estimates of ronary death for smokers relative to nonsmokers nong British male doctors, by age (data of example 12-2)*

| ge | Point estimate of IRR | Exact (mid-*P*) 90 percent confidence interval* |
|---|---|---|
| 5–44 | 5.74 | 1.91, 24 |
| 5–54 | 2.14 | 1.32, 3.63 |
| 5–64 | 1.47 | 1.06, 2.07 |
| 5–74 | 1.36 | 0.98, 1.91 |
| 5–84 | 0.90 | 0.65, 1.28 |

*Confidence intervals calculated from formulas 11-9 and 11-10.

Consider again the incidence rate data given in Example 12-2. It is apparent that the incidence rate ratio is not uniform over age, declining from an estimated value of 5.7 in the youngest stratum to just below unity in the oldest. A reasonable approach to the presentation of these data might be to show the stratum-specific results rather than any summary figure (Table 12-11). On the other hand, despite the interesting variability apparent in the data, a single summary estimate might be desired and could be defended. In that case, a standardized estimate of incidence rate ratio should be used; one reasonable choice for a standard would be the person-year distribution of smoking British male doctors, which would lead to the SMR:

$$\text{SMR} = \frac{630}{[(2/18{,}790)\,(52{,}407) + \ldots]} = \frac{630}{444.41} = 1.42$$

Naturally, a different choice of standard would affect the reported estimate of effect. For example, if each age category were assigned an equal weight, the resulting standardized rate ratio would be

$$\text{SRR} = \frac{\dfrac{32}{52{,}407\text{yr}} + \dfrac{104}{43{,}248\text{yr}} + \ldots}{\dfrac{2}{18{,}790\text{yr}} + \dfrac{12}{10{,}673\text{yr}} + \ldots} = 1.16$$

The relatively great difference in the above two effect estimates reflects only the different choice of weights involved in the selection of a standard. The second approach assigns relatively larger weights to the older age categories in which the effect is small.

Standardized estimates, like pooled estimates, are always weighted averages of stratum-specific effect estimates. For difference measures of effect, the weighting is the same as the standard weights, since the standardized rate difference may be expressed as

$$\text{SRD} = \frac{\Sigma w_i R_{1i}}{\Sigma w_i} - \frac{\Sigma w_i R_{0i}}{\Sigma w_i} = \frac{\Sigma w_i (R_{1i} - R_{0i})}{\Sigma w_i} = \frac{\Sigma w_i (\widehat{RD}_i)}{\Sigma w_i}$$

where $w_i$ is the standard weight for category i, $R_{1i}$ is the rate among exposed in stratum i, $R_{0i}$ is the rate among unexposed in stratum i, and $\widehat{RD}_i$ is the estimate of rate difference in stratum i. For rate ratio measures of effect, however, the standardized rate ratio is a weighted average of stratum-specific values that weights the stratum-specific rate ratios according to the product of the weight from the standard and the rate among the nonexposed:

$$\text{SRR} = \frac{\Sigma w_i R_{1i}/(\Sigma w_i)}{\Sigma w_i R_{0i}/(\Sigma w_i)} = \frac{\Sigma w_i R_{0i} \widehat{RR}_i}{\Sigma w_i R_{0i}} = \frac{\Sigma w_i' \widehat{RR}_i}{\Sigma w_i'} \qquad [12\text{-}66]$$

where $\widehat{RR}_i$ is the estimate of rate ratio in stratum i and $w_i' = w_i R_{0i}$.

The structure of formula 12-66 reveals why the two standardized rate ratios combining the stratum-specific point estimates in Table 12-11 are influenced heavily by the small effect estimates in the older age categories. Since $w_i'$ is a product of the weight for standardization and $R_{0i}$, the value of $R_{0i}$ will influence the weighting of the stratum-specific estimates. As it happens, $R_{0i}$ increases steeply with age for the data of Example 12-2, thereby magnifying the influence of the older age groups on the overall standardized rate ratio. Even the SMR, standardized as it is to the young age distribution of the smokers themselves, is only a modest 1.42 because of the influence of the $R_{0i}$ in expression 12-66.

Equation 12-66 can be applied, under certain conditions, to case-control data to obtain standardized rate ratio estimates from case-control studies [Miettinen, 1972]. Consider the SMR, which, as always, is standardized to the distribution of the exposed population. If $N_{1i}$ is the numerator and $D_{1i}$ is the denominator of the rate for the exposed source population of subjects in category i, then $a_i = f_{ca(i)}\,(N_{1i})$ and $c_i = f_{co(i)}\,(D_{1i})$, where $a_i$ is the number of exposed cases in stratum i of the case-control study, $c_i$ is the number of exposed controls in stratum i, and $f_{ca(i)}$ and $f_{co(i)}$ are the sampling fractions of cases and controls in stratum i of the source population. To standardize to the distribution of the exposed population, $w_i$ should be taken as $D_{1i} = c_i/f_{co(i)}$. $R_{1i}$ may be written as

$$R_{1i} = \frac{a_i/f_{ca(i)}}{c_i/f_{co(i)}}$$

so that $\Sigma w_i R_{1i} = \Sigma a_i/f_{ca(i)}$. Similarly,

$$R_{0i} = \frac{b_i/f_{ca(i)}}{d_i/f_{co(i)}}$$

and

$$\Sigma w_i R_{0i} = \sum_{i=1}^{N} \frac{b_i c_i}{d_i f_{ca(i)}}$$

If the sampling fraction for cases is constant over the strata, which will be true if the cases have been selected independently of the stratification factor, then

$$SMR = \frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} \frac{b_i c_i}{d_i}} \qquad [12\text{-}67]$$

Expression 12-67 has the usual form for an SMR, namely, the ratio of the observed number of exposed cases to an "expected" or null number. The expected number is not identical to the expected number used for statistical hypothesis testing, since hypothesis testing is premised on the correctness of the null hypothesis, which cannot be assumed for estimation. The expected number in expression 12-67 indicates how many exposed cases would have been observed if the exposure had no effect, but it involves no marginal totals that include the a cell, since the a cell is the one cell in each 2 × 2 table that differs from its null value when the exposure has an effect.

It is possible to choose other standards for the standardization of rate ratio estimates in case-control studies. For example, if $w_i$ is taken as the size of the denominator for the unexposed population in category i, equal to $D_{0i} = d_i/f_{co(i)}$, then

$$SRR = \frac{\sum_{i=1}^{N} \frac{a_i d_i}{c_i}}{\sum_{i=1}^{N} b_i} \qquad [12\text{-}68]$$

assuming once again that the sampling fraction for cases is constant over the strata.

Confidence intervals for standardized effect measures can be calculated, but they must reflect the pattern of the weights assigned by the standard. For standardized rate differences, an approximate variance formula is



$$\text{Var(SRD)} \doteq \frac{1}{(\Sigma w_i)^2} \sum_{i=1}^{N} w_i^2 \text{Var}(\text{RD}_i) \qquad [12\text{-}69]$$

where $w_i$ is the weight from the standard for category i and $\text{Var}(\widehat{\text{RD}}_i)$ is obtained from formula 12-1 or 12-6, depending on the type of data. The square root of Var(SRD) can be used for the standard deviation in formula 10-2 to obtain approximate confidence limits. For rate ratio measures, the usual logarithmic transformation should be used. With follow-up data, an approximate variance formula for the logarithm of a standardized rate ratio is

$$\text{Var}[\ln(\text{SRR})] \doteq \frac{\Sigma w_i^2 \text{Var}(R_{1i})}{(\Sigma w_i R_{1i})^2} + \frac{\Sigma w_i^2 \text{Var}(R_{0i})}{(\Sigma w_i R_{0i})^2} \qquad [12\text{-}70]$$

where $\text{Var}(R_{1i})$ and $\text{Var}(R_{0i})$ can be estimated from the first and second terms, respectively, of formula 12-1 or 12-6, depending on the type of data; the square root of Var[ln(SRR)] can be used in expression 10-4 to find approximate confidence limits.

To exemplify the application of formula 12-70, let us determine the approximate 90 percent confidence limits for the SMR calculated from Example 12-2. The weights from the standard, which for an SMR is always the exposed group, and the terms of the necessary sums for the variance of the logarithm of the SMR are given in Table 12-12. The variance is

$$\text{Var}[\ln(\text{SMR})] \doteq \frac{630}{630^2} + \frac{1997.56}{444.41^2} = 0.00159 + 0.01025 = 0.01184$$

The standard deviation is therefore $\sqrt{0.01184} = 0.1088$, and a 90 percent confidence interval around the SMR of 630/444.41 = 1.42 is

$$\exp[\ln(1.42) \pm 1.645(0.1088)] = 1.19, 1.70$$

*Table 12-12. Intermediate calculations for the variance of the logarithm of the SMR for the data of example 12-2*

| Age category | $w_i$ | $w_i R_{1i}$ | $w_i R_{0i}$ | $w_i^2 \text{Var}(R_{1i})$ | $w_i^2 \text{Var}(R_{0i})$ |
|---|---|---|---|---|---|
| 35–44 | 52,407 | 32 | 5.58 | 32 | 15.56 |
| 45–54 | 43,248 | 104 | 48.63 | 104 | 197.03 |
| 55–64 | 28,612 | 206 | 140.30 | 206 | 703.04 |
| 65–74 | 12,663 | 186 | 137.16 | 186 | 671.91 |
| 75–84 | 5,317 | 102 | 112.74 | 102 | 410.02 |
| Total | 142,247 | 630 | 444.41 | 630 | 1997.56 |

The analogous calculations for a standard in which a weight of 1.0 is assigned to each category, which gives a standardized rate ratio of 1.16, would result in a variance of 0.0161, corresponding to a standard deviation of 0.1269. The 90 percent confidence interval for the SRR with uniform weights is

$$\exp[\ln(1.16) \pm 1.645(0.1269)] = 0.94, 1.43$$

No general formulation can be made for the variance of the logarithm of a standardized rate ratio calculated from case-control data, since for case-control data the variance formula itself depends on the choice of a standard. For the SMR (calculated from formula 12-67), which uses the distribution of exposed subjects in the source population as the standard, the variance is approximated by

$$\text{Var}[\ln(\text{SMR})] \doteq \frac{1}{\sum_i a_i} + \frac{\sum_{i=1}^{N} \left(\frac{b_i c_i}{d_i}\right)^2 \left(\frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)}{\left(\sum_{i=1}^{N} \frac{b_i c_i}{d_i}\right)^2}$$

Using the data from Example 12-5, the SMR is calculable as 3.78, and the $\text{Var}[\ln(\text{SMR})] \doteq 0.350$, which gives a 90 percent confidence interval of

$$\exp[\ln(3.78) \pm 1.645(\sqrt{0.350})] = 1.43, 10.0$$

This point estimate and confidence interval happen to agree well in this instance with the (unconditional) maximum likelihood point estimate and the exact mid-*P* 90 percent confidence interval, which were previously calculated to be 3.79 and 1.30 to 9.78, respectively; while such agreement is reasonably common, it is not guaranteed because different principles are involved in weighting the stratum-specific results for the two approaches.

Using the distribution of the nonexposed subjects in the source population as the standard (i.e., formula 12-68), the variance is estimated as

$$\text{Var}[\ln(\text{SRR})] \doteq \frac{1}{\sum_i b_i} + \frac{\sum_{i=1}^{N} \left(\frac{a_i d_i}{c_i}\right)^2 \left(\frac{1}{a_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)}{\left(\sum_{i=1}^{N} \frac{a_i d_i}{c_i}\right)^2} \qquad [12\text{-}71]$$



For the data in Example 12-5, the SRR from formula 12-68 is 3.98 and the $\text{Var}[\ln(\text{SRR})] \doteq 0.381$, giving a 90 percent confidence interval of

$$\exp[\ln(3.98) \pm 1.645(\sqrt{0.381})] = 1.44, 11.0$$

## EFFECT FUNCTIONS

When the effect-modifying factor is measured on a continuous scale such as age it is possible to fit a mathematical equation describing the variation in the effect measure as a function of the effect modifier. For example, if the rate ratio seems to vary linearly with age, it is possible to express the rate ratio as a straight line function of age:

$$\text{RR} = a_0 + a_1(\text{age})$$

where $a_0$ is the "intercept" value and $a_1$ is a coefficient that describes the change in the rate ratio for a unit increment of age. The coefficients $a_0$ and $a_1$ can be estimated by a simple weighted regression procedure.

A linear function is not necessarily a good description of the mathematical relation between the effect measure and the effect modifier; it may be worthwhile to consider transformations that improve the description. The stratum-specific estimates of the incidence rate ratio for the data in Example 12-2 (Table 12-11) illustrate a progressive decline in IRR with age. The stratum-specific estimates are plotted in Figure 12-3. It is evident that a straight line will not provide as good a fit as one might hope. In Figure 12-4, the logarithm of each age-specific estimate of IRR is plotted; for the five age categories from youngest to oldest, these values are 1.7469, 0.7603, 0.3841, 0.3046, and −0.1001, respectively. These values conform better to a linear pattern. For the data of Example 12-2, then, it seems reasonable to describe the effect of smoking, as measured by the incidence rate ratio, as a function of age using a logarithmic transformation of the IRR:

$$\ln(\text{IRR}) = a_0 + a_1(\text{age})$$

The coefficients for this equation can be determined easily by a linear regression procedure. It is important to use a weighted regression that assigns to each age-specific observation a weight that reflects the precision of that estimate; a weight proportional to the reciprocal of the variance of $\ln(\hat{\text{IRR}}_i)$ accomplishes this purpose. The age-specific weights for the weighted regression are calculated as the reciprocals of the variances determined by formula 12-4; the weights are 1.88, 10.76, 24.65, 24.34, and 23.77 from youngest to oldest, respectively. Note the small weight accorded to the youngest age category, for which only two events were observed among the nonsmokers; the small number of events in the denom-

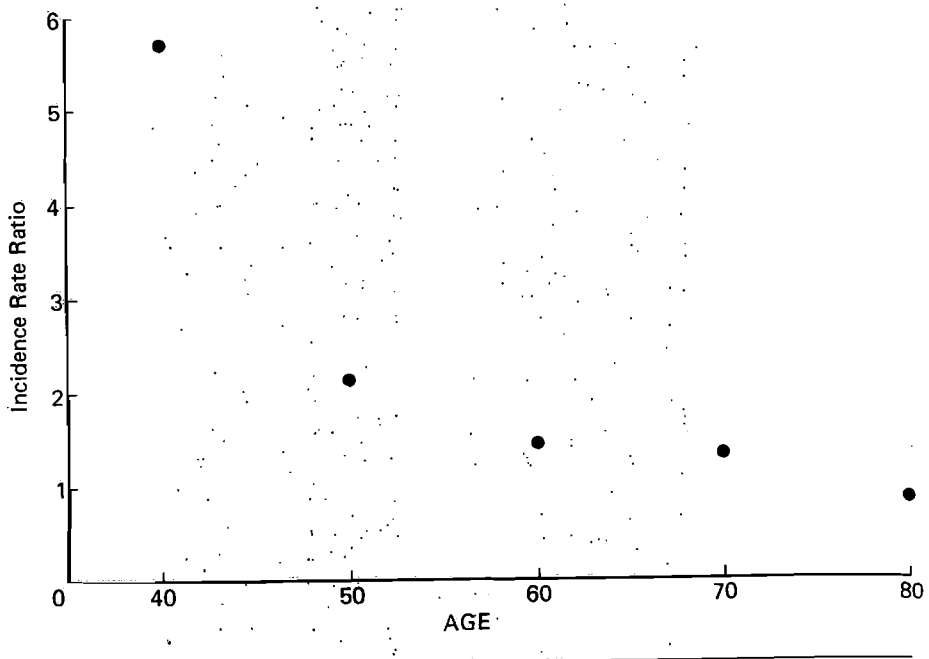


*Fig. 12-3. Estimates of incidence rate ratio for coronary death for smoking British doctors, compared with nonsmoking doctors, by age (data of example 12-2).*

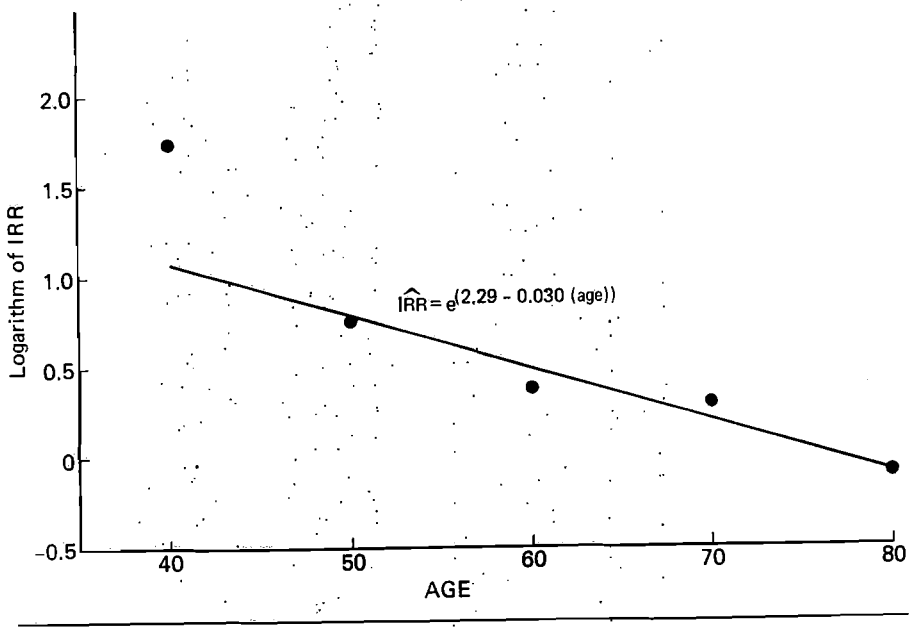


*Fig. 12-4. Logarithm of estimates of incidence rate ratio for coronary death for smoking British doctors, compared with nonsmoking doctors, by age (data of example 12-2), and fitted weighted regression line.*



inator rate of the rate ratio leads to a large variance for the rate ratio. The youngest category consequently does not contribute much to the fitting of the weighted regression line. A least-squares weighted regression analysis [Kleinbaum and Kupper, 1978] using the above weights gives $a_0 = 2.29$ and $a_1 = -0.030$. To express the IRR as a function of age using these results, we can reverse the logarithmic transformation:

$$\widehat{IRR} = e^{(2.29 - 0.030(age))}$$

The above fitted equation describes the incidence rate ratio as a function of age and can be used to estimate the IRR at any given age. For example, at age 65 the estimate of IRR is $\exp(2.29 - 0.030(65)) = 1.43$. At age 40 the estimated value of IRR is 3.0; the predicted and observed values are relatively discrepant at age 40 because the entire set of age categories was used to generate the coefficients, but little weight was contributed by the unstable estimate of IRR in the youngest age category. For age 80, however, the estimated IRR of 0.91 from the regression equation is nearly identical to the observed value of 0.90 because of the greater weights assigned to the older age categories. The overall pattern indicates roughly exponential decline in IRR with age, until the effect disappears entirely between ages 75–80 (the apparent reversal in the direction of the effect at the oldest ages is not striking enough to warrant a biologic interpretation).

## REFERENCES


Breslow, N. Odds ratio estimators when the data are sparse. *Biometrika* 1981;68:73–84.

Breslow, N. E., and Liang, K. Y. The variance of the Mantel-Haenszel estimator. *Biometrics* 1982;38:943–952.

Cornfield, J. A statistical problem arising from retrospective studies. *Proc. Third Berkeley Sympos.* 1956;4:135–148.

Doll, R., and Hill, A. B. Mortality of British doctors in relation to smoking; observations on coronary thrombosis. In W. Haenszel (ed.), *Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases. Natl. Cancer Inst. Mono.* 1966;19:205–268.

Gart, J. J. Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed marginals. *Biometrika* 1970;57:471–475.

Gart, J. J. The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Rev. Int. Stat. Inst.* 1971;39:148–169.

Greenland, S., and Robins, J. M. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 1985;41:55–68.

Haldane, J. B. S. The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Hum. Genet.* 1955;20:309–311.

Hauck, W. W. The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics* 1979;35:817–819.

Hempelmann, L. H., Hall, W. J., Phillips, M., et al. Neoplasms in persons treated with

x-rays in infancy—fourth survey in 20 years. *J. Natl. Cancer Inst.* 1975;55:519–530.
Kleinbaum, D. G., and Kupper, L. L. *Applied Regression Analysis and Other Multivariate Methods.* North Scituate, MA: Duxbury Press, 1978. P. 243.
Lubin, J. H. An empirical evaluation of the use of conditional and unconditional likelihoods for case-control data. *Biometrika* 1981;68:567–571.
Mann, J. I., Inman, W. H. W., and Thorogood, M. Oral contraceptive use in older women and fatal myocardial infarction. *Br. Med. J.* 1968;2:193–199.
Mantel, N., and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 1959;22:719–748.
Mantel, N., Brown, C., and Byar, D. P. Tests for homogeneity of effect in an epidemiologic investigation. *Am. J. Epidemiol.* 1977;106:125–129.
McKinlay, S. M. The effect of nonzero second-order interaction on combined estimators of the odds ratio. *Biometrika* 1978;65:191–202.
Miettinen, O. S. Standardization of risk ratios. *Am. J. Epidemiol.* 1972;96:383–388.
Miettinen, O. S., and Neff, R. K. Computer processing of epidemiologic data. *Hart Bull.* 1971;2:98–103.
Nurminen, N. Asymptotic efficiency of general noniterative estimators of common relative risk. *Biometrika* 1981;68:525–530.
Robins, J. M., Breslow, N., and Greenland, S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large strata limiting models. *Biometrics* 1986;42:311–323.
Rothman, K. J. Spermicide use and Down's syndrome. *Am. J. Public Health* 1982;72:399–401.
Rothman, K. J., and Monson, R. R. Survival in trigeminal neuralgia. *J. Chron. Dis.* 1973;26:303–309.
Rothman, K. J., and Boice, J. D. *Epidemiologic Analysis with a Programmable Calculator.* Brookline, MA: Epidemiology Resources, 1982. (First edition published by U.S. Government Printing Office, Washington D.C., NIH Publication No. 79-1649, June, 1979.)
Tarone, R. E. On summary estimators of relative risk. *J. Chron. Dis.* 1981;34:463–468.
Thomas, D. G. Exact and asymptotic methods for the combination of 2 × 2 tables. *Computers and Biomedical Research* 1975;8:423–446.
University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult onset diabetes. *Diabetes* 1970;19(Suppl. 2):747–830.
Walker, A. M. Small sample properties of some estimators of a common hazard ratio. *Appl. Stat.* 1985;34:42–48.
Woolf, B. On estimating the relation between blood group and disease. *Ann. Hum. Genet.* 1954;19:251–253.
Zelen, M. The analysis of several 2 × 2 contingency tables. *Biometrika* 1971;58:129–137.

# 13. MATCHING

*Matching* refers to the selection of a comparison series—unexposed subjects in a follow-up study or controls in a case-control study—that is identical, or nearly so, to the index series with respect to one or more potentially confounding factors. The mechanics of the matching may be performed subject by subject, which is described as *individual matching,* or for groups of subjects, which is described as *frequency matching.* The general principles that apply to matched data are identical for individually matched or frequency matched data.

## PRINCIPLES OF MATCHING

The topic of matching in epidemiology is beguiling: What at first seems clear is seductively deceptive. Whereas the clarity of an analysis in which confounding has been securely prevented by perfect matching of the compared series seems indubitable and impossible to misinterpret, the intuitive foundation for this cogency attained by matching is a surprisingly shaky structure that does not always support the conclusions that are apt to be drawn. The difficulty is that our intuition about matching springs from knowledge of experiments or follow-up studies, whereas matching is most often applied in case-control studies, which differ enough from follow-up studies to make the implications of matching different and counterintuitive.

Whereas the traditional view, stemming from an understanding based on follow-up studies, has been that matching enhances validity, in case-control studies the effectiveness of matching as a methodologic tool derives from its effect on study efficiency, not on validity. Indeed, for case-control studies it would be more accurate to state that matching introduces confounding rather than that it prevents confounding.

The different implications of matching for follow-up and case-control studies are easy to demonstrate. Consider a source population of 2,000,000 individuals, distributed by exposure and sex as indicated in Table 13-1. Both the exposure and male gender are risk factors for the disease: For the exposure the relative risk is 10, and for males relative to females it is 5. There is also substantial confounding, since 90 percent of the exposed individuals are male and only 10 percent of the unexposed are male. The crude relative risk in the source population, comparing exposed with unexposed, is 32.9, considerably different from the unconfounded value of 10.

Now consider what happens if a follow-up study is planned by drawing the exposed cohort from the exposed source population and matching the unexposed cohort to the exposed cohort for sex. Suppose 10 percent of the exposed source population were included in the follow-up study; if these subjects were selected independently of gender, we would have approximately 90,000 males and 10,000 females in the exposed cohort. A