11. ANALYSIS OF CRUDE DATA

The simplest type of epidemiologic analysis, which is based on crude (i.e., unstratified) data, applies when it is not necessary to take into account any factors beyond the exposure and the disease of interest. Although it is not unusual to see data presented solely in crude form, typically the investigator needs first to explore more complicated analyses using stratification or multivariate methods to evaluate the role of other factors. Vigorous restriction by covariates in subject selection (so as to prevent confounding) will often lead to a simple or crude analysis. Clinical trials using random allocation of subjects also can often be analyzed satisfactorily in crude form if the investigators are persuaded that the randomization has successfully prevented confounding. A crude analysis, because of its simplicity, possesses an appealing cogency that is lacking in more complicated analyses.

HYPOTHESIS TESTING WITH CRUDE DATA

The epidemiologist, in conceptualizing types of epidemiologic data, tends to separate follow-up data from case-control data. For statistical hypothesis testing, however, statistical modeling leads to a different kind of separation according to whether the data consist of person-time units or persons as the basic observations. Whereas the units of observation are measured as person-time only in follow-up studies, not all follow-up studies are presented with the data expressed as incidence rates with person-time denominators. If all subjects are followed for a constant period, it may be convenient to express the incidence rates as risk estimates, that is, cumulative incidence data, in which the number of cases is related not to an amount of person-time experience but to the total number of people who were followed. Clinical trials are often presented in this manner. When the denominators of incidence measures are presented as counts of persons rather than as measures of person-time experience, the statistical model that applies for hypothesis testing is the same one that applies to case-control data, in which all observations also are counts of persons.

Hypothesis Testing with Person-Time Data

Crude incidence-rate data consist of the total number of cases and person-time units for both exposed and unexposed categories. We shall use the notation in Table 11-1. The apparent simplicity of this table may mask analytic subtleties that must be considered. Specifically, the person-time experience in the "exposed" column should be defined according to a plausible or tentative model for induction time. Before an individual becomes exposed, all of that individual's person-time experience is, naturally, unexposed person-time (though it is often not included as such in an analysis). If exposure occurs at a point in time and the induction-time model being evaluated calls for a minimum induction time of 5 years, then the 5 years

Table 11-1. Notation for crude incidence-rate data with person-time denominators

	Exposed	Unexposed	Total
Cases	a	b	M_1
Person-time	N_1	N_0	T

after the point of exposure for each individual is likewise unexposed person-time experience rather than exposed, because according to the induction-time model it relates back to a period of time when exposure was absent. Tallying the person-time units into the appropriate exposure categories is a task that must be done subject by subject and may involve complicated rules if the exposure is chronic. Incident cases are tallied into the same category to which the concurrent person-time units are being added—for example, an incident case occurring 4 years after exposure would be tallied in the "unexposed" category if the induction-time model specified a minimum induction time of 5 years.

The statistical model used for hypothesis testing of person-time data is the *binomial distribution* [Shore et al., 1976]. A random event that has only two possible outcomes, X and Y, that occur with fixed probabilities is referred to as a *Bernoulli trial*. Flipping a coin is an example. Let the probability of one of the two outcomes, say X, be p. The probability distribution of the total number of Xs occurring in N independent Bernoulli trials with p constant is referred to as a binomial distribution. Mathematically, the probability is expressed as

$$\Pr(\text{total number of Xs} = x) = \binom{N}{x} p^x (1 - p)^{(N-x)}$$

$$\text{where} \quad \binom{N}{x} \quad \text{is} \quad \frac{N!}{x!\,(N - x)!}$$

The mean of the binomial distribution is Np, and the variance is Np(1 − p).

In applying the binomial model to crude person-time data, each case is considered to be an independent Bernoulli trial, having as its "outcome" the two possibilities of exposed or unexposed. According to the null hypothesis that exposure is unrelated to disease, the probability that a given case will be classified as exposed or unexposed depends only on the proportion of the total person-time experience that is allocated to the exposed category; that is, each case has a probability equal to N_1/T of being classified as exposed under the null hypothesis.

The M_1 cases are thus considered to be M_1 independent Bernoulli trials,

and the distribution of exposed cases has a binomial distribution under the null hypothesis, with $p = N_1/T$. The probability of the observed data under the null hypothesis can be written as

$$\Pr(\text{number of exposed cases} = a) = \binom{M_1}{a} \left(\frac{N_1}{T}\right)^a \left(\frac{N_0}{T}\right)^b$$

An exact one-tail Fisher *P*-value can be obtained as

$$\sum_{k=a}^{M_1} \Pr(\text{number of exposed cases} = k)$$

The above summation gives the upper tail of the distribution; the lower tail can be obtained by summing k over the range from 0 to a. To obtain the mid-*P* value instead of the traditional Fisher *P*-value, only one-half the probability of the observed data should be added to the summation for each tail. When the mid-*P* values are calculated, the lower and upper tails of the distribution have the desirable property of summing to unity.

With large numbers, these exact calculations are unnecessary because an asymptotic test statistic will give accurate approximations for the *P*-value. The test statistic is computed from formula 11-1, using the number of exposed cases as the random variate. Based on the formulas for the mean and variance of the number of successes in a binomial distribution, the null expectation for the number of exposed cases is N_1M_1/T, and the variance is $M_1N_1N_0/T^2$, which gives

$$\chi = \frac{a - N_1M_1/T}{\sqrt{\dfrac{M_1N_1N_0}{T^2}}} \qquad [11\text{-}1]$$

The χ values can then be translated into *P*-values from tables of the standard normal distribution.

For Example 11-1, the probability of the observed data under the null hypothesis may be calculated as

$$\Pr(41 \text{ exposed cases}) = \binom{56}{41} \left(\frac{28{,}010}{47{,}027}\right)^{41} \left(\frac{19{,}017}{47{,}027}\right)^{15} = 0.0122$$

An exact upper-tail Fisher *P*-value may be calculated by repeating the calculation for the more extreme positive outcomes up through 56 exposed cases. For 42 exposed cases, the calculation gives

$$\Pr(42 \text{ exposed cases}) = \binom{56}{42} \left(\frac{28{,}010}{47{,}027}\right)^{42} \left(\frac{19{,}017}{47{,}027}\right)^{14} = 0.0064$$

Example 11-1. Breast cancer cases and person-years of observation for women with tuberculosis repeatedly exposed to multiple x-ray fluoroscopies, and women with tuberculosis not so exposed [Boice and Monson, 1977]

	Radiation exposure		
	Yes	No	Total
Breast cancer	41	15	56
Person-years	28,010	19,017	47,027

Similarly, Pr(43 exposed cases) = 0.0031, Pr(44 exposed cases) = 0.0013, and Pr(45 exposed cases) = 0.0005. The small magnitude of this last probability indicates that it should not be necessary to calculate the additional terms in the summation, since their contribution would be even smaller and therefore would not affect the sum materially. The one-tail *P*-value thus equals 0.0122 + 0.0064 + 0.0031 + 0.0013 + 0.0005 = 0.024. The one-tail mid-*P* would have ½(0.0122) as the first term, giving 0.017 as the *P*-value (it is actually 0.0174 and would be rounded to 0.018 if the summation were carried a few terms more). Two-tail *P*-values can be obtained simply by doubling the corresponding one-tail *P*-values.

The numbers in the example are large enough to use the normal approximation in formula 11-1, which is a simpler calculation:

$$\chi = \frac{41 - 28{,}010\left(\frac{56}{47{,}027}\right)}{\sqrt{\frac{(56)\,(28{,}010)\,(19{,}017)}{(47{,}027)^2}}} = \frac{41 - 33.35}{\sqrt{13.49}} = \frac{7.65}{3.67} = 2.08$$

From tables of the standard normal distribution, a χ value of 2.08 corresponds to a one-tail *P*-value of 0.019, which agrees closely with the exact one-tail mid-*P* value.

Hypothesis Testing with Count Data

Follow-up data or prevalence data with denominators consisting of the number of persons at risk can be treated like case-control data for statistical hypothesis testing. For each of these types of data, the basic information can be displayed in a 2 × 2 table in which all four cells of the table are frequencies of subjects classified according to the presence or absence of exposure and disease. The notation we shall use is given in Table 11-2.

Superficially, Table 11-2 resembles Table 11-1 except for the addition of an added row for noncases. The denominators in Table 11-2, however, are frequencies, or counts, of subjects rather than person-time accumulations.

Table 11-2. Notation for crude 2 × 2 table

	Exposed	Unexposed	Total
Cases	a	b	M_1
Noncases	c	d	M_0
Total	N_1	N_0	T

Again, the apparent simplicity of the table may mask some subtleties in determining the classification of subjects.

For case-control data, classification according to exposure depends on an appropriate and meaningful definition of exposure according to a biologic model of induction time that specifies the timing of exposure in relation to disease. In a study of oral cavity cancer, for example, patients with cancer may tend to use mouthwash regularly more frequently than controls, but such use may occur as a consequence of early symptoms of the disease or of its subsequent treatment (radiotherapy in the region of the oropharynx tends to shrink the salivary glands and cause foul breath). A meaningful model for induction time should classify as unexposed only those individuals who were exposed to the agent outside the time window during which exposure might have been etiologically related to the disease.

For follow-up data analyzed with a 2 × 2 table, presumably all subjects were free of disease at the beginning of the follow-up period; the classification of exposure refers to the time of initiation of follow-up, and the classification of disease refers to the time of completion of follow-up. Disease occurrence should not count as such unless it occurs during the time window specified by a meaningful induction-time model. An instance of the illness of interest occurring before or after the hypothesized induction time window should be ignored; if illness occurs before the time window, it may be reasonable to exclude the subject as not being free of disease at the start of the relevant period of follow-up. If the follow-up period has been so long that a substantial proportion of subjects have been lost or have died from causes unrelated to the outcome of interest, it is preferable to use person-time denominators rather than to analyze the data with a 2 × 2 table.

The 2 × 2 table can be considered as representing two independent series of observations: For case-control studies the observations are exposure observations and the two independent series of subjects are the cases and the controls; for follow-up studies the observations are disease observations and the two independent series are the exposed and unexposed groups. The observations made on each of the two independent series can be considered as conforming to the model of a binomial distribution; under the null hypothesis, the probability of a "positive" observa-

tion in each of the two independently observed binomial series is the same.

Consider a follow-up study of N_1 exposed subjects and N_0 unexposed subjects. In the exposed series, "a" subjects develop disease, and in the unexposed series, "b" subjects develop disease. The probability that exactly a and b subjects will develop disease among the exposed and unexposed, respectively, is, according to the binomial model,

Pr(a exposed cases and b unexposed cases)

$$= \binom{N_1}{a}(p_1)^a(1 - p_1)^c \cdot \binom{N_0}{b}(p_0)^b(1 - p_0)^d \quad [11\text{-}2]$$

which is the product of the binomial probabilities for each of the two independent groups, exposed and unexposed. The probability of developing disease among the exposed is p_1; among the unexposed, it is p_0. Under the null hypothesis, these two probabilities are equal: $p_1 = p_0 = p$, which gives

Pr(a exposed cases and b unexposed cases)

$$= \binom{N_1}{a}\binom{N_0}{b} \cdot (p)^{M_1}(1 - p)^{M_0} \quad [11\text{-}3]$$

To calculate the value of expression 11-3 for a particular 2 × 2 table, it is necessary to have an estimate of p. Usually p is estimated directly from the data, using the overall disease proportion from the margins of the table, M_1/T. Substituting M_1/T for p gives

Pr(a exposed cases and b unexposed cases)

$$= \binom{N_1}{a}\binom{N_0}{b} \cdot (M_1/T)^{M_1}(M_0/T)^{M_0} \quad [11\text{-}4]$$

From expression 11-4 it is possible to obtain a *P*-value that represents an exact test based on two independent binomial distributions, provided that it is clear how the departures from the null state that are more extreme than those observed are calculated. Let us assume that a positive association is observed between exposure and disease. Assume that a and b are the number of exposed and unexposed cases actually observed. For other possible realizations of the data in which the number of exposed cases exceeds a while the number of unexposed cases is b or less, the overall departure from the null condition is more extreme than that actually observed. Similarly, if the number of exposed cases is a but the number of unexposed cases is less than b, again the departure from the null would be more extreme than that actually observed. The preceding possibilities are easy to classify, but what if the number of exposed cases were a + 1

and the number of unexposed cases were b + 1? What about other combinations such as a + 1 and b + 2? It is difficult to say whether these possibilities represent situations that depart from the null to a greater extent than the actual observations. To decide definitively if a departure is more extreme, it would be necessary to evaluate an effect measure for each hypothetical outcome of the data and compare that measure with the effect measure calculated from the actual observations. Interestingly, the decision about which outcomes are more extreme would depend on which effect measure was used.

To illustrate, consider example 11-2. The "observed" data indicate an estimated risk difference of 0.05, a risk ratio of 1.11, and an odds ratio of 1.22. Variations 1 and 2 are two other possible outcomes for the data, presuming that the same number of exposed and unexposed subjects are studied. Using the risk difference measure to determine departures from the null, variation 2, but not variation 1, is a more extreme departure from the null. Using the risk ratio measure, neither variation 1 nor variation 2 is more extreme. For the odds ratio measure, both variations are more extreme.

The different measures each designate a distinct set of outcomes as more extreme. This ambiguity makes it problematic to use two independent binomial distributions as a model for hypothesis testing for a 2 × 2 table. Another problem with the use of two binomials is the large number of possible outcomes. For example, if $N_1 = N_0 = 25$, there are 676 possible outcomes for the data $[(N_1 + 1) \cdot (N_0 + 1)]$. To simplify the calculation, an assumption can be made that addresses both of these problems. The assumption, for follow-up or prevalence data, is that the total number of cases actually observed is taken to be a constant [Mantel and Hankey, 1971]. For case-control data, the two binomial distributions refer not to the exposed and unexposed series but to the case and control series, and the corresponding assumption is that the total number of exposed subjects is constant. These assumptions essentially fix all the marginal totals of the 2 × 2 table; therefore, if the a cell increases, the b and c cells must each decrease an equivalent amount, and the d cell increases by the same amount. With all the margins held constant, there is only one random variable to describe: variation in any cell of the 2 × 2 table with fixed marginal totals is locked together with concomitant variation in each of the other cells. Usually, then, the focus becomes simply the a cell of the table, which is taken to be the random variable.

The assumption that all the marginal totals are fixed in a 2 × 2 table can be justified methodologically as a means of focusing the problem (of hypothesis testing) directly on the association between exposure and disease. In the jargon of statistics, the "nuisance parameter" is removed by fixing the marginal totals: Testing the null hypothesis using a model of two independent binomials requires assessing the values for *two* parameters, the proportions p_1 and p_0, whereas the analytic problem can be reduced

Example 11-2. Hypothetical data illustrating ambiguity of definition for departures from the null state using the two-binomial model

	"Observed data"		Variation 1		Variation 2	
	Exposed	Unexposed	Exposed	Unexposed	Exposed	Unexposed
Cases	10	45	16	76	13	59
Noncases	10	55	4	24	7	41
Totals	20	100	20	100	20	100
Estimate of						
Risk difference		0.05		0.04		0.06
Risk ratio		1.11		1.05		1.10
Odds ratio		1.22		1.26		1.29

Example 11-3. History of chlordiazopoxide use in early pregnancy for mothers of children born with congenital heart defects and mothers of normal children [Rothman et al., 1979]

	Chlordiazopoxide use		
	Yes	No	Total
Case mothers	4	386	390
Control mothers	4	1250	1254
Totals	8	1636	1644

to assessing the value of a single measure. That measure is the odds ratio, equal to $p_1(1 - p_0)/[p_0(1 - p_1)]$, which is completely determined by the value of the a cell if the marginal totals are taken as fixed. Testing a departure of the odds ratio from unity is equivalent to testing a departure of p_1 from p_0, since the null condition of $p_1 = p_0$ is equivalent to an odds ratio of unity, but fixing the margins of the 2 × 2 table greatly simplifies the calculations by reducing the number of parameters in the model from two to one.

The statistical model that describes the variability of the a cell in a 2 × 2 table with fixed marginal totals is the hypergeometric distribution. The probability of a exposed cases occurring under the assumption that the null hypothesis is correct can be expressed simply as follows [Fisher, 1935]:

$$\Pr(\text{a exposed cases}) = \frac{\binom{N_1}{a}\binom{N_0}{b}}{\binom{T}{M_1}} \qquad [11\text{-}5]$$

For the data in example 11-3, the hypergeometric probability for four exposed cases is

$$\Pr(\text{4 exposed cases}) = \frac{\binom{8}{4}\binom{1636}{386}}{\binom{1644}{390}} = \frac{390!\ 1254!\ 8!\ 1636!}{1644!\ 4!\ 4!\ 386!\ 1250!}$$

$$= \frac{(390)\,(389)\,(388)\,(387)\,(1254)\,(1253)\,(1252)\,(1251)\,(8)\,(7)\,(6)\,(5)}{(1644)\,(1643)\,(1642)\,(1641)\,(1640)\,(1639)\,(1638)\,(1637)\,(4)\,(3)\,(2)} = 0.0748$$

The probability for an outcome more extreme, five exposed cases, under the hypergeometric model, would be

$$\Pr(5 \text{ exposed cases}) = \frac{\binom{8}{5}\binom{1636}{385}}{\binom{1644}{390}} = 0.0185$$

The probability for six exposed cases would be 0.0028; for seven exposed cases, 0.0002; and for the most extreme outcome, eight exposed cases, 0.000009. The total one-tail *P*-value calculated according to Fisher would be 0.0748 + 0.0185 + 0.0028 + 0.0002 + 0.000009 = 0.096. The one-tail mid-*P* would be 0.0374 + 0.0185 + 0.0028 + 0.0002 + 0.000009 = 0.059. The two-tail *P*-value, either Fisher or mid-*P*, could be obtained by doubling the one-tail *P*-value.

For this example, the hypergeometric model requires the calculation of only five probabilities. Had the model of two independent binomial distributions, with 390 cases and 1,254 controls as the two independent series, been used instead, thousands of calculations would be necessary to determine which outcomes were equally or more extreme, and then the probability of each of these outcomes would have to be calculated as well. The simplifying assumption of the hypergeometric distribution, which fixes all the marginal totals, reduces the complexity of the calculations enormously.

The reasonableness of the hypergeometric assumption, even for data such as those given in example 11-3 in which two of the four cell frequencies are small, is evident by comparing the results with the results obtained by using the two-binomial model. Using the two-binomial model and using the magnitude of the odds ratio to determine which outcomes are equally or more extreme departures from the null, the Fisher *P*-value was found to be 0.094, and the mid-*P*, 0.071. (This calculation took several hours using a BASIC program on a microcomputer.) The agreement between the two approaches is striking when one considers that only five separate probabilities are included in the hypergeometric calculation, whereas thousands are included in the two-binomial model. With larger cell frequencies, the agreement between the results obtained from the different models improves. Whatever disagreement exists between the results from the two approaches does not indicate any inaccuracy with the hypergeometric approach; since the assumption of fixed marginal totals yields a valid test, even if the margins were not actually fixed by the study design, a test of the null hypothesis based on the hypergeometric model is just as valid as a test based on the two-binomial model. Since the hypergeometric approach is extraordinarily simpler, it is clearly the preferred model.

Unfortunately, even the hypergeometric model can require an onerous number of calculations if all the cell frequencies are sizable. In most applications, therefore, an asymptotic test statistic is used to calculate the *P*-

value. The asymptotic test statistic can be derived starting from either the two-binomial model or the hypergeometric model. With the hypergeometric model, the random variable would be the a cell, the number of exposed cases. The null expectation for the number of exposed cases is N_1M_1/T, and the hypergeometric variance for the number of exposed cases is $M_1M_0N_1N_0/[T^2(T-1)]$. The χ statistic is

$$\chi = \frac{a - N_1M_1/T}{\sqrt{\frac{M_1M_0N_1N_0}{T^2(T-1)}}} \qquad [11\text{-}6]$$

which appears similar to equation 11-1 for person-time data. If an asymptotic test statistic were derived from the two-binomial model rather than from the hypergeometric, one would compare the two observed binomial proportions, a/N_1 and b/N_0. Under the null hypothesis, the expectation of the difference between these proportions is zero. The variance might be estimated in several ways; the usual way is to use a pooled common variance for the two binomial proportions, since under the null hypothesis the binomial probabilities for the two binomial distributions are equal. Thus, M_1/T is taken to be an estimate of the pooled binomial probability, and the variance of the difference in proportions can be expressed as

$$\left(\frac{M_1}{T}\right)\left(\frac{M_0}{T}\right)\left[\frac{1}{N_1} + \frac{1}{N_0}\right]$$

which gives

$$\chi = \frac{\frac{a}{N_1} - \frac{b}{N_0}}{\sqrt{\left(\frac{M_1}{T}\right)\left(\frac{M_0}{T}\right)\left[\frac{1}{N_1} + \frac{1}{N_0}\right]}} \qquad [11\text{-}7]$$

Algebraic manipulation of equation 11-7 gives an expression nearly identical to equation 11-6:

$$\chi = \frac{a - N_1M_1/T}{\sqrt{\frac{M_1M_0N_1N_0}{T^3}}} \qquad [11\text{-}8]$$

The only difference between the two formulas is the $T - 1$ in the denominator expression in equation 11-6, which is replaced by T in equation 11-8. Since neither formula is applicable unless T is large, for practical purposes these formulas are identical.

If an asymptotic test statistic had been used to calculate the *P*-value for the data in example 11-3, we would have obtained, using equation 11-6,

$$\chi = \frac{4 - (8)(390)/(1644)}{\sqrt{\dfrac{(8)(1636)(390)(1254)}{(1644)^2(1643)}}} = 1.75$$

which gives a one-tail *P*-value of 0.040. As one would expect, the *P*-value resulting from the asymptotic test is closer to the mid-*P* exact value than to the Fisher exact value (see Chap. 10), but the approximation is not very good. Notice that under the hypergeometric model there are only nine possible outcomes for the a cell; evidently the number of outcomes is too few for the normal approximation to be valid. A rule of thumb that is often used is that the asymptotic test statistic should be applied only when the smallest null expectation of any cell in the 2 × 2 table, based on the marginal totals, is greater than about 3. If there is any doubt, however, about the adequacy of the asymptotic approximation, it is best to evaluate the *P*-value exactly.

ESTIMATION OF EFFECTS WITH CRUDE DATA

Estimation with Follow-up Data

POINT ESTIMATION

Point estimation of either difference or ratio measures of effect involves taking the difference or ratio of the observed values of incidence or risk. Thus, the point estimate of incidence rate difference (IRD) would be

$$\widehat{IRD} = \frac{a}{N_1} - \frac{b}{N_0}$$

and the point estimate of incidence rate ratio (IRR) would be

$$\widehat{IRR} = \frac{a/N_1}{b/N_0}$$

Similarly, for risk (cumulative incidence) data, in which denominators are counts rather than measures of person-time, the point estimate of risk difference (RD) would be

$$\widehat{RD} = \frac{a}{N_1} - \frac{b}{N_0}$$

and the point estimate of risk ratio (RR) would be

$$\widehat{RR} = \frac{a/N_1}{b/N_0}$$

If the object of inference is the ratio of incidence rates rather than risk ratio, then the ratio of risks that is directly calculable from risk data using the above formula leads to an underestimate of the effect. The degree of underestimation depends on the level of the risks, being slight for small risks and greater for large risks (see Chap. 4 and Table 6-1). An alternative approach to point estimation with count denominators is to use the odds-ratio formula

$$\widehat{IRR} = \frac{ad}{bc}$$

which overestimates the effect to roughly the same extent that the risk ratio underestimates it (Table 6-1) but has the advantage of being the same estimator used in case-control studies (formula 6-1).

INTERVAL ESTIMATION

Exact Interval Estimation with Follow-up Data. Interval estimation can be exact or approximate. For exact interval estimation, like the calculation of an exact *P*-value, an appropriate statistical model must be used to describe the probability distribution of the data. The model will generally be an extension of the model used for calculation of an exact *P*-value. For testing the null hypothesis, an effect of zero is assumed and incorporated into the statistical model; for calculation of exact confidence limits, the statistical model must be able to accommodate nonzero effects.

INCIDENCE RATE (PERSON-TIME) DATA. For incidence rate difference, a difficulty arises in attempting to postulate a statistical model from which an exact confidence interval can be calculated. For hypothesis testing, the binomial model rests on the assumption that M_1, the total number of cases, is a constant. This assumption is analogous to the assumption of fixed marginal totals in the hypergeometric model for 2 × 2 tables. For interval estimation, the problem with assuming M_1 to be constant is that, with respect to incidence rate difference, the value of M_1 is not simply a "nuisance parameter" that statistically has no bearing on the effect measure; the value of M_1 imposes a limit on the magnitude of the incidence rate difference (a small value of M_1 is compatible only with small values of the rate difference), therefore requiring the sampling variability of M_1 to be taken into account in estimating the incidence rate difference. Thus, the single-binomial model with a fixed M_1 cannot be used to calculate exact confidence limits for incidence rate difference with person-time data. The counterpart

for person-time data of the more general two-binomial model for 2 × 2 tables would be a model of two independent Poisson distributions, in which exposed and unexposed cases each occurred independently with frequencies described by a Poisson distribution. The Poisson distribution, however, has no upper limit for the number of events (i.e., cases) that can occur, so it cannot be used for the above calculations without arbitrary truncation. For these reasons, exact interval estimation for incidence rate difference is not easily possible.

It is appropriate, however, to fix M_1 for estimation of incidence rate ratio because the ratio measure depends on the ratio of exposed to unexposed cases, not on the absolute magnitude of the frequencies. Therefore, M_1 can be considered a nuisance parameter that is statistically independent of the rate ratio measure. For estimation, the simple single binomial model used for hypothesis testing must be modified to accommodate a nonzero effect. This can be accomplished by noting that the probability that a case is exposed, given M_1, is related to incidence rate ratio as follows:

$$\text{IRR} = \frac{N_0 \cdot \Pr(\text{case is exposed})}{N_1 \cdot \Pr(\text{case is unexposed})}$$

Exact confidence limits for IRR can be obtained by setting the tail probability of the binomial distribution equal to $\alpha/2$ and $1 - \alpha/2$, where $1 - \alpha$ equals the desired level of confidence. If we denote $\underline{u}$ as the lower confidence bound for the probability that a case is exposed, and $\bar{u}$ as the exact upper confidence bound, then

$$\underline{\text{IRR}} = \frac{\underline{u}N_0}{(1 - \underline{u})N_1} \qquad [11\text{-}9]$$

and

$$\overline{\text{IRR}} = \frac{\bar{u}N_0}{(1 - \bar{u})N_1} \qquad [11\text{-}10]$$

where $\underline{u}$ and $\bar{u}$ are the solutions to the following equations (for Fisher limits):

$$\alpha/2 = \sum_{k=a}^{M_1} \binom{M_1}{k} \underline{u}^k (1 - \underline{u})^{M_1 - k}$$

and

$$1 - \alpha/2 = \sum_{k=a+1}^{M_1} \binom{M_1}{k} \bar{u}^k (1 - \bar{u})^{M_1 - k}$$

The preceding equations assume that $\hat{\text{IRR}} > 1$, and consequently calculate the upper tail of the distribution. If $\hat{\text{IRR}} < 1$, then the lower end of the distribution could be used to calculate the tail probabilities:

$$\alpha/2 = \sum_{k=0}^{a} \binom{M_1}{k} \bar{u}^k (1 - \bar{u})^{M_1 - k}$$

and

$$1 - \alpha/2 = \sum_{k=0}^{a-1} \binom{M_1}{k} \underline{u}^k (1 - \underline{u})^{M_1 - k}$$

If calculations are performed based on the mid-*P* exact *P*-value, then the tail probabilities are calculable as

$$\alpha/2 = \frac{1}{2}\binom{M_1}{a} \underline{u}^a (1 - \underline{u})^b + \sum_{k=a+1}^{M_1} \binom{M_1}{k} \underline{u}^k (1 - \underline{u})^{M_1 - k}$$

and

$$1 - \alpha/2 = \frac{1}{2}\binom{M_1}{a} \bar{u}^a (1 - \bar{u})^b + \sum_{k=a+1}^{M_1} \binom{M_1}{k} \bar{u}^k (1 - \bar{u})^{M_1 - k}$$

These equations must be solved iteratively, by choosing trial values for $\underline{u}$ and $\bar{u}$ and calculating the tail probability repeatedly until it is equal to $\alpha/2$ or $1 - \alpha/2$. Notice the similarity to the calculation of an exact *P*-value, which involves taking $u = N_1/T$ and calculating the tail probability once. For exact confidence limits, the value of u is adjusted until the tail probability equals the predefined values, $\alpha/2$ or $1 - \alpha/2$.

Consider again the person-time data in example 11-1. Exact Fisher-type 90 percent confidence limits for the IRR would be calculated from equations 11-9 and 11-10 as follows:

$$0.05 = \sum_{k=41}^{56} \binom{56}{k} \underline{u}^k (1 - \underline{u})^{56 - k}$$

$$0.95 = \sum_{k=42}^{56} \binom{56}{k} \bar{u}^k (1 - \bar{u})^{56 - k}$$

These calculations are best done by computer or by a shortcut method that involves the F-distribution [Rothman and Boice, 1982; Brownlee, 1965]. A trial and error solution of the preceding equations gives $\underline{u} = 0.618$ and $\bar{u} = 0.827$, which gives $\underline{\text{IRR}} = 1.10$ and $\overline{\text{IRR}} = 3.25$.

If the limits are calculated based on the mid-*P* exact *P*-value, then the equations to be solved are

$$0.05 = \frac{1}{2}\binom{56}{41} \underline{u}^{41} (1 - \underline{u})^{15} + \sum_{k=42}^{56} \binom{56}{k} \underline{u}^k (1 - \underline{u})^{56 - k}$$

and

$$0.95 = \frac{1}{2}\binom{56}{41}\bar{u}^{41}(1-\bar{u})^{15} + \sum_{k=42}^{56}\binom{56}{k}\bar{u}^{k}(1-\bar{u})^{56-k}$$

The upper and lower 90 percent confidence limits determined by the above mid-*P* based equations are $\underline{u} = 0.626$ and $\bar{u} = 0.8205$, corresponding to $\underline{IRR} = 1.14$ and $\overline{IRR} = 3.10$.

CUMULATIVE INCIDENCE DATA. If the denominators are counts rather than person-time units, an exact confidence interval for risk difference could theoretically be calculated from the two-binomial model. The calculation, however, would involve iterative determination of the exact tail probability based on two independent binomials and is therefore not readily feasible.

Confidence limits for the risk ratio measure are also subject to the same computational difficulty because the value of the measure is dependent on the total number of cases and requires the use of two independent binomials. If, however, the odds ratio measure is used for estimation, both margins of the 2 × 2 table may be considered fixed, and the calculations can be greatly simplified because the odds ratio measure is independent of the total number of cases. Because the odds ratio is only an approximation of the risk ratio, the calculation of exact limits for the odds ratio does not produce exact confidence limits for the risk ratio. The approximation is good only if the risks are small, in which case the exact confidence interval for the odds ratio can be used as a reasonable surrogate confidence interval for the risk ratio.

The statistical model that describes the variation of the a cell in a 2 × 2 table with fixed margins is the hypergeometric, but for the non-null situation the "noncentral" form of the hypergeometric distribution must be used. The noncentral hypergeometric is more complicated than the null form of the hypergeometric distribution given in formula 11-5 because it accommodates the strength of association between exposure and disease measured by the odds ratio. Given the value of the odds ratio, R, the probability of observing a exposed cases is [Fisher, 1935; Gart, 1971]

$$\Pr(\text{a exposed cases}) = \frac{\binom{N_1}{a}\binom{N_0}{b}R^a}{\sum_{k=\max(0,M_1-N_0)}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}R^k}$$

When R = 1, the above formula reduces to expression 11-5. Exact confidence limits for R with a confidence level of $1 - \alpha$ can be calculated from the formulas

$$\alpha/2 = \frac{\sum_{k=a}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}\underline{R}^k}{\sum_{k=\max(0,M_1-N_0)}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}\underline{R}^k} \qquad [11\text{-}11]$$

and

$$1-\alpha/2 = \frac{\sum_{k=a+1}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}\bar{R}^k}{\sum_{k=\max(0,M_1-N_0)}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}\bar{R}^k} \qquad [11\text{-}12]$$

for the Fisher limits and

$$\alpha/2 = \frac{\frac{1}{2}\binom{N_1}{a}\binom{N_0}{b}\underline{R}^a + \sum_{k=a+1}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}\underline{R}^k}{\sum_{k=\max(0,M_1-N_0)}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}\underline{R}^k} \qquad [11\text{-}13]$$

$$1-\alpha/2 = \frac{\frac{1}{2}\binom{N_1}{a}\binom{N_0}{b}\bar{R}^a + \sum_{k=a+1}^{\min(N_1,M_1)}\binom{N_1}{k}\binom{N_0}{M_1-k}\bar{R}^k}{\sum_{k=\max(0,M_1-N_0)}^{\min(N_1,M_1)} + \binom{N_1}{k}\binom{N_0}{M_1-k}\bar{R}^k} \qquad [11\text{-}14]$$

for mid-*P* limits. The solution of the foregoing equations can be time-consuming, since each iteration in the process calls for calculating a complicated sum, but it is not nearly as complicated as the calculations that would be required using a statistical model of two independent binomials.

The data in example 11-4 describe partial results from a follow-up study evaluating risk of diarrhea in breast-fed infants in Bangladesh during an

Example 11-4. Diarrhea during a 10-day follow-up period in 30 breast-fed infants colonized with Vibrio cholerae 01, *according to antilipopolysaccharide antibody titers in mother's breast milk [Glass et al., 1983]*

	Antibody level	
	High	Low
Diarrhea	7	12
No diarrhea	9	2
Totals	16	14

11-day period following the determination of various antibody titers in the mothers' breast milk. An exact *P*-value of the null hypothesis of no association gives $P_{(1)} = 0.02$ for the Fisher *P*-value and $P_{(1)} = 0.01$ for the mid-*P* value. The estimate of relative risk from these data comparing the group exposed to high titers with the group exposed to low titers is [7/16]/[12/14] = 0.51. The odds ratio estimate differs considerably from the relative risk estimate because the risks are so high; the odds ratio estimate is [7 × 2]/[12 × 9] = 0.13. An exact confidence interval can be calculated for the odds ratio using formulas 11-11 and 11-12 to set the Fisher limits. The 90 percent confidence limits are 0.017 and 0.751, obtained by the trial-and-error solution of equations 11-11 and 11-12. If the mid-*P* exact limits were desired instead, these could be obtained from equations 11-13 and 11-14 as 0.025 and 0.608. These exact limits for the odds ratio, however, cannot be used as confidence limits for the risk ratio, since the odds ratio is a poor approximation to the risk ratio with these data.

Approximate Interval Estimation with Follow-up Data. Approximate interval estimation from crude follow-up data is straightforward.

INCIDENCE RATE (PERSON-TIME) DATA. Consider first incidence rate data with person-time denominators. Two effect-measures can be estimated, rate difference and rate ratio. Because the rate-difference measure has a symmetric sampling distribution, no scale transformation is needed to obtain accurate approximate confidence limits. The number of exposed and unexposed cases can each be assumed to have a Poisson distribution, from which the variance for each rate can be estimated as a/N_1^2 and b/N_0^2 for exposed and unexposed groups, respectively. The standard deviation of the rate difference, then, is the square root of the sum of the variances of each rate,

$$\text{SD(Incidence rate difference)} \doteq \sqrt{\frac{a}{N_1^2} + \frac{b}{N_0^2}} \qquad [11\text{-}15]$$

From the data in example 11-1, we can estimate the rate difference as

$$\frac{41}{28{,}010\ \text{yr}} - \frac{15}{19{,}017\ \text{yr}} \doteq \frac{6.75}{10{,}000}\ \text{yr}^{-1}$$

with a standard deviation for the rate difference of

$$\text{SD} \doteq \sqrt{\frac{41}{(28{,}010\ \text{yr})^2} + \frac{15}{(19{,}017\ \text{yr})^2}} = \frac{3.06}{10{,}000}\ \text{yr}^{-1}$$

To obtain an approximate 90 percent confidence interval, the standard deviation is multiplied by 1.645 to get the limits as follows:

$$\frac{6.75}{10{,}000}\ \text{yr}^{-1} \pm 1.645\left[\frac{3.06}{10{,}000}\ \text{yr}^{-1}\right] = \frac{1.7}{10{,}000}\ \text{yr}^{-1}, \frac{11.8}{10{,}000}\ \text{yr}^{-1}$$

Another approach would be to use formula 10-6 for test-based limits,

$$\hat{\text{RD}}\ (1 \pm Z/\chi)$$

Earlier we calculated χ to be 2.08. Using that value in the above formula with Z = 1.645 yields an approximate 90 percent confidence interval of 1.4/(10,000 yr), 12.1/(10,000 yr), which compares well with the other approximation.

For the estimation of rate ratio, it is desirable to use a logarithmic transformation to compensate for the asymmetric sampling distribution. By taking confidence limits that are symmetric about the logarithm of the rate ratio and then reversing the transformation by taking antilogarithms, much greater accuracy can be achieved than by taking limits calculated symmetrically around the rate ratio itself. Thus, we calculate

$$\exp\{\ln(\hat{\text{RR}}) \pm Z \cdot \text{SD}[\ln(\hat{\text{RR}})]\} \qquad [11\text{-}16]$$

The standard deviation of the incidence rate ratio can be approximated by

$$\text{SD}[\ln(\hat{\text{RR}})] \doteq \sqrt{\frac{1}{a} + \frac{1}{b}}$$

Again using the data from example 11-1, we can estimate the incidence rate ratio to be

$$\frac{41/28{,}010\ \text{yr}}{15/19{,}017\ \text{yr}} = 1.86$$

and ln(1.86) = 0.618. The standard deviation of the log-transformed point estimate is

$$\sqrt{1/41 + 1/15} \doteq \sqrt{0.091} = 0.302$$

A 90 percent confidence interval for the ln(RR) would then be

$$0.618 \pm 1.645(0.302) = 0.12, 1.1$$

which, after taking antilogarithms to reverse the transformation, gives a confidence interval of 1.1 to 3.0. The whole process can be summarized as follows:

$$\exp[\ln(1.86) \pm 1.645\sqrt{1/41 + 1/15}] = 1.1, 3.0$$

These limits agree well with the exact mid-*P* 90 percent limits calculated previously as 1.1 and 3.1.

An alternative approach would be to use the test-based formula

$$\hat{RR}^{(1 \pm Z/\chi)}$$

in which χ has the value of 2.08 for the data in example 11-1. Using Z = 1.645 for 90 percent limits, the test-based approach gives an interval of 1.1 to 3.0, which is also in excellent agreement with the exact mid-*P* limits.

CUMULATIVE INCIDENCE DATA. To get approximate limits for follow-up data with denominators consisting of persons rather than person-time, slightly different formulas are needed to estimate the standard deviations. For the risk difference, the standard deviation is derived from the sum of two binomial variances and is estimated as

$$\text{SD(Risk difference)} \doteq \sqrt{\frac{a(N_1 - a)}{N_1^3} + \frac{b(N_0 - b)}{N_0^3}} \qquad [11\text{-}17]$$

From the data in example 11-4, the point estimate of rate difference is

$$\frac{7}{16} - \frac{12}{14} = -0.42$$

with an approximate 90 percent confidence interval of

$$\begin{aligned} &-0.42 \pm 1.645\sqrt{\frac{(7)(9)}{(16)^3} + \frac{(12)(2)}{(14)^3}} \\ &= -0.42 \pm 1.645(0.155) \\ &= -0.68, -0.16 \end{aligned}$$

Alternatively, the test-based calculation gives

$$-0.42\left[1 \pm \frac{1.645}{-2.34}\right] = -0.71, -0.12$$

Considering the small numbers involved in these calculations, the agreement between these two approaches seems good.

For the risk ratio, it is again desirable to use a logarithmic transformation, that is, to apply formula 11-16. The standard deviation, however, is estimated as

$$\text{SD}[\ln(\hat{RR})] \doteq \sqrt{\frac{c}{aN_1} + \frac{d}{bN_0}}$$

For example 11-4, the risk ratio estimate is [7/16]/[12/14] = 0.51, and

$$\ln(\hat{RR}) = -0.673$$

The estimated standard deviation of $\ln(\hat{RR})$ is

$$\sqrt{\frac{9}{7 \cdot 16} + \frac{2}{12 \cdot 14}} = 0.304$$

and the 90 percent confidence interval is

$$\exp[-0.673 \pm 1.645(0.304)] = 0.31, 0.84$$

Alternatively, test-based limits could be calculated as

$$\hat{RR}^{(1 \pm Z/\chi)} = 0.51^{(1 \pm 1.645/-2.34)} = 0.32, 0.82$$

Once again the two approximate methods for confidence interval estimation are in good agreement.

If the inference from follow-up data with count denominators is to be based on the odds ratio rather than on the risk ratio, then the formula for standard deviation is [Woolf, 1955]

$$\text{SD}[\ln(\text{odds ratio})] \doteq \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

For example 11-4, the odds ratio is [7·2]/[9·12] = 0.13 and the logarithm is ln(0.13) = −2.04. The standard deviation is

$$\text{SD}[\ln(\text{odds ratio})] \doteq \sqrt{\frac{1}{7} + \frac{1}{9} + \frac{1}{12} + \frac{1}{2}} = 0.915$$

and the approximate 90 percent confidence interval is

$$\exp[\ln(0.13) \pm 1.645(0.915)] = 0.03, 0.58$$

The test-based confidence limits are calculated as

$$0.13^{(1 \pm 1.645/-2.34)} = 0.03, 0.55$$

Considering the very small numbers involved in the calculations, the above two approximate interval estimates for the odds ratio agree tolerably well not only with one another but also with the exact mid-*P* confidence interval for the odds ratio, calculated previously to be 0.025 to 0.608.

The Cornfield [1956] approach, which is described in greater detail in Chapter 12, is a theoretically preferable approximate technique since it involves recalculating the standard error using fitted cell frequencies that correspond to the value of the confidence limit. Thus, the procedure is iterative and involves substantially more calculation than the other approximate methods. For the data of example 11-4, the Cornfield approach gives a 90 percent confidence interval of 0.03 to 0.55, agreeing in this instance with the test-based approach.

Case-Control Data

For case-control data, the epidemiologic measure of central interest is the odds ratio, the point estimator for which is

$$\hat{R} = \frac{ad}{bc}$$

Exact confidence interval estimation for the odds ratio is identical for case-control and follow-up data and is based on formulas 11-11 through 11-14. Approximate confidence intervals for the odds ratio from case-control data are determined using the same method used for follow-up data, using the logarithmic transformation with one of the following formulas:

$$\hat{R}^{(1 \pm Z/\chi)} \quad \text{or} \quad \exp\{\ln(\hat{R}) \pm Z \cdot SD[\ln(\hat{R})]\}$$

where

$$SD[\ln(\text{odds ratio})] \doteq \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Consider example 11-3. Exact 90 percent confidence limits for the odds ratio, using equations 11-11 and 11-12 for the Fisher limits, are 0.77, 13.6; using equations 11-13 and 11-14 for the mid-*P* limits, the results are 0.94,

11.1. Approximate 90 percent confidence limits can be determined as follows:

$$\ln(\hat{R}) = \ln(3.24) = 1.175$$

$$SD[\ln(\text{odds ratio})] \doteq \sqrt{\frac{1}{4} + \frac{1}{386} + \frac{1}{4} + \frac{1}{1250}}$$

$$= 0.71$$

$$\exp[\ln(3.24) \pm 1.645(0.71)] = 1.0, 10.4$$

or, using the test-based approach,

$$3.24^{(1 \pm 1.645/1.75)} = 1.1, 9.8$$

As expected, these results agree better with the mid-*P* exact limits than with the Fisher exact limits. The approximation is not perfect, but neither is it very poor considering that two of the four cells of the 2 × 2 table have observed frequencies of only four. The Cornfield method gives a 90 percent interval of 1.1 to 9.8, identical to that given by the test-based approach.

REFERENCES

Boice, J. D., and Monson, R. R. Breast cancer in women after repeated fluoroscopic examinations of the chest. *J. Natl. Cancer Inst.* 1977;59:823–832.

Brownlee, K. A. *Statistical Theory and Methodology in Science and Engineering.* New York: Wiley, 1965.

Cornfield, J. A statistical problem arising from retrospective studies. In J. Neyman (ed.), *Proceedings Third Berkeley Symposium,* Vol. 4. Berkeley: University of California Press, 1956, pp. 135–148.

Fisher, R. A. The logic of inductive inference. *J. R. Stat. Soc.,* Series A, 1935;98:39–54.

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.

Glass, R. I., Svennerholm, A. M., Stoll, B. J., et al. Protection against cholera in breast-fed children by antibiotics in breast milk. *N. Engl. J. Med.* 1983;308:1389–1392.

Mantel, N., and Hankey, B. F. Programmed analysis of a 2 × 2 contigency table. *Am. Stat.* 1971;25:40–44.

Rothman, K. J., and Boice, J. D. *Epidemiologic Analysis with a Programmable Calculator* (2nd ed.). Brookline, MA: Epidemiology Resources Inc., 1982.

Rothman, K. J., Fyler, D. C., Goldblatt, A., et al. Exogenous hormones and other drug exposures of children with congenital heart disease. *Am. J. Epidemiol.* 1979;109:433–439.

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