International Journal of Epidemiology
© International Epidemiological Association 1995

Vol. 24, No. 1
Printed in Great Britain

LEADING ARTICLE

Basic Models for Disease Occurrence in Epidemiology

W DANA FLANDERS AND DAVID G KLEINBAUM

Flanders W D (Emory University School of Public Health , Division of Epidemiology, 1599 Clifton Rd, NE, Atlanta, GA 30329, USA) and Kleinbaum D G. Basic models for disease occurrence in epidemiology. *International Journal of Epidemiology* 1995; **24**: 1–7.

Background. One of the epidemiologist's most basic tasks is estimation of disease occurrence. To perform this task, the epidemiologist frequently models variability in disease occurrence using one of three distributions—the binomial, the Poisson or the exponential distribution. Although epidemiologists often use them and their properties appear in standard texts, we know of no text or review that compares and contrasts epidemiological application of these distributions.

Methods. In this commentary, we discuss these three basic distributions. We note key assumptions as well as limitations, and compare results from analyses based on each distribution.

Results and Conclusions. We illustrate that the three distributions, although superficially different, often lead to similar results. We argue that epidemiologists should often obtain similar results regardless of which distribution they use. We also point out that application of all three distributions can be inappropriate if assumptions of independence or homogeneity of risks fail to hold. Finally, we briefly review how these basic distributions can be used to justify use of other distributions, such as the Gaussian distribution, for studying disease-exposure associations.

Assessment of disease occurrence is one of the epidemiologist's most basic tasks. To perform this task, the epidemiologist typically uses either of two basic measures: cumulative incidence, which reflects the average risk of disease in a population over a specified time, or incidence rate, which reflects the average rate of disease occurrence per unit of person time. Each measure has well-developed, accepted methods for estimation.[1–5]

Although these basic measures allow the epidemiologist to describe disease occurrence in a population, disease occurrence is not fixed, but varies from place to place, from time to time, and from group to group. This variability reflects differences in biological phenomena, genetic differences, environmental and social differences, differences in medical care, and random and poorly understood phenomena. Sampling variability, when one selects subjects from a larger universe, also contributes.

In many applications, the epidemiologist will need to account for this inherent variability, perhaps by calculating confidence intervals or p-values. In surveillance activities, for example, the epidemiologist may wish to assess whether the observed increase in rates is likely to indicate a real increase in morbidity or is consistent with random variation. In aetiological studies, the epidemiologist may wish to assess if the difference in observed rates between two subgroups is likely to reflect a real difference or is compatible with chance.

This important step—accounting for variability in disease occurrence—generally involves use of a statistical distribution to model the variability. Distributions commonly used for this purpose include the binomial, the Poisson and the exponential.[1–3] Although application of these distributions is described separately in standard texts and papers, we are unaware of any text, review or commentary in which epidemiological application of these distributions is compared, contrasted and integrated.

This review has four purposes. In the first section, we review the binomial, Poisson and exponential distributions, highlighting key assumptions and indicating the relationship of model parameters with basic measures of epidemiology. In the second section, we discuss limitations of these three models. In the third section, we indicate the inter-relationships between these three

Emory University School of Public Health, Division of Epidemiology, 1599 Clifton Rd, NE, Atlanta, GA 30329, USA.

basic statistical approaches. And finally, we discuss the relationship of these models for disease occurrence to some of the standard methods of assessing disease-exposure associations.

Our review supports two general conclusions. First, application of the three models will tend to yield the same analytical conclusions, an expected result since the underlying assumptions are similar. Second, the epidemiologist may need to use other models when checking to see whether misspecification of the model has affected the conclusions, since the assumptions that underlie these three models are alike. These results should help epidemiologists understand the relationships between different models, their limitations and applications.

BASIC STATISTICAL DISTRIBUTIONS

In this section, we discuss the binomial, Poisson and exponential distributions, note key assumptions, indicate the relationship of model parameters with basic measures of epidemiology, and illustrate their application.

Binomial Distribution

Epidemiologists often use the binomial distribution to model variability of disease frequency in follow-up studies or cross-sectional studies. Specifically, they use it to model the variability in the number of people, X, observed to develop disease in a cohort of size N. In symbols, the binomial distribution states that

$$P(X = x) = \binom{N}{x} p^x (1-p)^{N-x},$$

where N is the number of people at risk, X is the number of 'cases', and p is the probability of disease for each person.

We can justify the use of the binomial distribution by assuming that: 1) disease occurs independently in different people; 2) disease occurs with probability p in each person. This homogeneity of risks can also hold if the disease probability for each person can be modelled as arising randomly by independent selection from an underlying distribution of risks. In this situation, each person's risk is:

$$p = \sum_{i=1}^{n} p_i \cdot P(p_i),$$

where p_i denotes the i^{th} disease probability, and $P(p_i)$ the probability of selecting this disease probability. On the other hand, use of the binomial may not be justified if the length of follow-up varies since risks may then be heterogeneous—longer follow-up should associate with greater risk. Given these assumptions, disease occurrence in the population is consistent with a series of independent, identically distributed Bernoulli trials implying that the total number of cases follows a binomial distribution with parameters N and p.

TABLE 1 *Data from follow-up study of Doll and Hill,[7] summarized using counts*

	Smokers	Non-smokers	Total
Deaths	1582[a]	166	1748
Non-disease	27 116	5630	32 746
Total	28 698[b]	5796	34 494

[a] We apportioned the smoking status of 34 deaths among men under age 35 (smoking status not reported) according to the proportion of smokers among the 1714 deaths among older men (smoking status reported).
[b] We calculated the number of smokers in the entire cohort from the total number in the cohort, and the distribution of person-years across smoking categories.

Interpretation of the binomial parameter as a measure of disease frequency depends on the type of study. In a follow-up study, the parameter p typically corresponds to cumulative incidence over the period of observation. For a cross-sectional study, the parameter p corresponds to the disease prevalence in the study population.

The maximum likelihood estimate[6] of p is A/N with associated estimate variance $A(N-A)/N^3$, where A is the number of cases and N is the size of the cohort. For small cohorts, we can base exact confidence limits for the cumulative incidence on the binomial distribution.

Example 1. Consider the follow-up study of smoking among British physicians reported by Doll and Hill[7] in which each subject, initially disease-free, was followed for 4 years 5 months to detect subsequent occurrence of disease. In Table 1, we summarize data separately for the smoking and non-smoking cohorts. As just noted, we can model variability of the number of deaths (cases) in each cohort (smoking and non-smoking) by using binomial distributions, if independence and homogeneity of risks are reasonable assumptions. Based on the binomial distribution, the maximum likelihood estimate of the cumulative incidence among the smokers is 0.0551 with associated standard deviation estimate of 0.00097. Since the sample is large, the maximum likelihood estimator of cumulative incidence has an approximate Gaussian distribution, so that an approximate 95% confidence interval is 0.0532 to 0.0570. Similarly, the estimated cumulative incidence among the non-smokers is 0.0286 with confidence interval from 0.0243 to 0.0329.

Poisson Distribution

Epidemiologists frequently use the Poisson distribution to analyse data from follow-up studies when the summary data involve counts of cases. Typical applications include studies of cancer, cardiovascular disease, other chronic diseases, and mortality. Data required for analysis consist of the total number of cases and the person-time of follow-up for each subgroup of interest. Use of the Poisson distribution to model variability in counts of cases is probably reasonable if we can assume that the disease occurs independently in different people and in the same person at different points in time, that the likelihood that a new case will occur in a short period is proportional to the number of people, and that disease risks are homogeneous across people and time. These assumptions have a more formal statement, called Poisson postulates,[6,8] which statisticians use to justify formal application of the Poisson distribution. In symbols, the Poisson distribution states that

$$P(X = x) = \exp(-\mu)\,\mu^x/x!,$$

where X is the total number of cases that occur during the follow-up period and μ is a parameter to be estimated.

With this formulation, the Poisson parameter μ is readily interpretable as the incidence rate multiplied by the person-time. The maximum likelihood estimate (MLE) of μ is A, with associated variance estimate A, and the MLE for the incidence rate (IR) is A/PT, where A is the number of cases and PT is the person-time of observation of the cohort during the follow-up period. With this approach, we typically treat the person-time as though it were a constant. For small cohorts, 'exact' confidence limits for μ and p-values are readily calculable from tables of the Poisson distribution. For large cohorts, approximate confidence limits and p-values can be based on the Gaussian approximation to the Poisson distribution.[8] The following example illustrates application of the Poisson distribution to analyse count data from a follow-up study.

Example 2. Consider again the follow-up study of smoking reported by Doll and Hill.[7] As summarized in Table 2, they reported 1582 deaths among smokers, who they followed for 123 436 person-years (py), and 166 deaths among non-smokers, who they followed for 25 250 py. Applying Poisson distributions to model variability in the case counts in each cohort, the MLE for the expected (mean) number of deaths among smokers is 1582 with associated standard deviation estimate of 39.8. Since the sample is large, we can treat the estimate as Gaussian, so that the approximate 95% confidence interval is 1504 to 1660. Similarly, the MLE of the mean number of deaths among non-smokers is 166 with confidence interval, 141 to 191. Dividing by the appropriate number of person-years gives an estimated incident rate among smokers of 0.0128 cases per person-year (c/py), with 95% confidence interval from 0.0122 c/py to 0.0134 c/py, and an incidence rate among non-smokers of 0.00657 c/py with 95% confidence interval from 0.00557 c/py to 0.00757 c/py.

TABLE 2 *Data from follow-up study of Doll and Hill,[7] summarized using person-time*

	Smokers	Non-smokers
Disease	1582	166
Person-years	123 436	25 250

Exponential Distribution

Epidemiologists also use the exponential distribution to model disease occurrence in follow-up studies. Specifically, the exponential distribution models time until disease occurrence, an approach which is particularly useful for survival analyses when censoring has occurred. Data required for analysis consist of the duration of follow-up and disease status for each subject. Use of the exponential distribution to analyse data from a follow-up study is probably reasonable if we can assume that disease occurs independently (over different people), and that disease risks for a given length of time are homogeneous across people and across different points in time. In symbols, the exponential distribution states that

$$P(t \leq T < t + dt) = \lambda \cdot \exp(-\lambda \cdot t) \cdot dt,$$

where T is the time until disease occurs and λ is a parameter to be estimated.

The exponential distribution's parameter, λ, is readily interpretable, given the homogeneity assumption, as the incidence rate. The MLE of λ is A/PT with associated variance estimate A/PT^2; large sample confidence limits for λ can be based on the Gaussian distribution, as illustrated in the following example.

Example 3. Consider the follow-up study of smoking and British physicians reported by Doll and Hill.[7] We can summarize the duration of follow-up and disease status which they report, as shown in Table 3. Using the exponential distribution to model disease occurrence, the MLE of the incidence rate among the exposed is 0.0128 c/py with associated standard deviation estimate

TABLE 3 *Data from follow-up study of Doll and Hill,[7] summarized using (average) individual follow-up time*

Number of subjects	Average years of follow-up each person[a]	Smoking status	Disease
27 116	4.42	yes	no
1582	2.32	yes	yes
5630	4.42	no	no
166	2.32	no	yes

[a] Calculated from data presented by Doll and Hill by assuming that the average follow-up of subjects who died was the same among smokers and non-smokers. The follow-up of other subjects was 4 years 5 months.

of 0.00032. Since the study is large, we can treat the estimate of the incidence rate as approximately Gaussian, so that the approximate 95% confidence interval is 0.0122 to 0.0134. The corresponding estimate for the unexposed is 0.00657 with 95% confidence interval from 0.00557 c/py to 0.00757 c/py.

LIMITATIONS

Application of these distributions to model variability is limited, in part, because of the need to assume independence and homogeneity. For studies of communicable disease, application may be inappropriate because of lack of independence. For example, if one person in a group develops the 'flu', others in that group have higher risk, reflecting dependency of disease occurrence. In situations like these, the investigator may attempt to modify the model to account for the dependency.[9]

Application of stochastic models for disease occurrence may also be limited by violations of the homogeneity assumption.[10] The homogeneity assumption probably does not hold in many situations since, even after accounting for recognized risk factors, unrecognized risk factors presumably subject different people to different risk. Under heterogeneity, we can still estimate the cumulative incidence by the number of cases over the cohort size (A/N) and estimate the incidence rate by the number of cases over person-time (A/PT), but must recognize that these estimates represent 'average' risks and incidence rates in the population.

Although the simple estimators may still apply for estimation of average risks and incidence, usual variance estimates may be biased when risks are heterogeneous[11] (personal communication, S Greenland). To exemplify this bias, consider the year-to-year variation in estimated cumulative incidence of death in a particular county. We suppose the county population consists of two subgroups, one with annual risk p_1 and size n_1, the other with annual risk p_2 and size n_2. The average annual cumulative incidence, ignoring subgroups is $p_\cdot = (n_1p_1 + n_2p_2)/(n_1 + n_2)$, and the associated year-to-year variance is $(n_1p_1(1-p_1) + n_2p_2(1-p_2))/(n_1 + n_2)^2$, assuming the composition of the county changes negligibly from year to year. The simple variance estimate based on the average risk, however, is $p_\cdot(1-p_\cdot)/(n_1+n_2)$, an overestimate of the actual variance.

Biased estimation of the variance can also result from correlation of outcomes between people (lack of independence). To exemplify this bias, consider the year-to-year variation in estimated risk of a communicable disease for which prior illness confers no immunity. We suppose that precisely a proportion p_0 of the population fall ill in non-epidemic years, and a proportion p_1 in epidemic years. If an epidemic occurs in a given year with probability r because one or more index cases arise, then the average cumulative incidence is $p_\cdot = (1-r)p_0 + rp_1$., and the variance (year-to-year) is $(1-r)(p_0)^2 + r(p_1)^2 - (p_\cdot)^2$. Simple estimation of the variance, based on the binomial distribution and $p_\cdot$, is $p_\cdot(1-p_\cdot)/n$ where n is the county size. The latter estimate substantially underestimates variance for any reasonable size county. In this example, the disease risk each year is the same for each person (homogeneous risks), but lack of independence invalidates the simple binomial model. Thus, violations of the assumptions can lead to either conservative or anti-conservative conclusions.

INTER-RELATIONSHIPS BETWEEN MODELS

We have considered three statistical distributions that epidemiologists often use to model variability in disease occurrence. In this section, we show that these distributions typically lead to the same analytic result, a similarity which results in part because the underlying assumptions are similar. In other words, these different models are often consistent with one another, and the epidemiologist may have a choice—any of two or three models might appropriately form the basis of analysis if the underlying assumptions are met. On the other hand, if risks are not homogeneous or disease does not occur independently, then the epidemiologist will probably not be able to use any of the three distributions considered here since, without specific modifications, each depends on independence and homogeneity.

First, consider the Poisson and exponential distributions. The MLE of the incidence rate and associated variance estimates which result from application of the exponential distribution are the same as those which result from use of the Poisson distribution, illustrating the

agreement of analytic results for these two distributions. Moreover, if the number of deaths in a given population has a Poisson distribution, the time until the first death and the time between deaths is known to follow an exponential distribution, further illustrating the close relationship between these two distributions.[6,12]

The binomial distribution is closely related to these two distributions, a correspondence which is particularly easy to see for rare disease. As noted above, the binomial distribution leads to p = A/N as the estimated cumulative incidence. We can convert this to an estimate of the incidence rate, assuming the rate to be constant, by using the usual[1,2] relationship between risk and rates: IR = −ln(1−Risk)/t ≈ Risk*t, where the approximation holds for rare disease and where t is the length of follow-up. Substituting the cumulative incidence estimate, A/N, into this expression and combining results gives: IR ≈ A/N*t ≈ A/PT (Table 4). The associated variance estimate is A/PT^2, again assuming rare disease. Thus, the estimated incidence rate and the corresponding variance associated with the binomial distribution under these conditions are approximately the same as those derived from the Poisson and exponential distributions. The following example further illustrates that the different distributions often lead to the same analytic result.

Example 4. Consider again the follow-up study considered in *Example 1* (Table 1). As noted previously, the MLE of the cumulative incidence accrued over 4.42 years in smokers is 0.0551 and that in non-smokers is 0.0286. If we assume a constant incidence rate and treat the deaths as rare (this is borderline), we can use the approximations in Table 4 to obtain the related incidence rate estimates. Thus, based on the binomial distribution, the estimated incidence rate among smokers is −ln(1−0.0551)/4.42 ≈ 0.0128 c/py. We can estimate a lower 95% confidence limit for the incidence rate by using this same risk-to-incidence rate-transformation on the lower limit for the cumulative incidence: −ln(1−0.0532)/4.42 = 0.0124 c/py. The associated upper limit for the incidence rate is −ln(1−0.0227)/2 = 0.0133 c/py. Among non-smokers, the corresponding incidence rate estimate is 0.00657 c/py, with 95% confidence interval from 0.00557 c/py to 0.00757 c/py.

Comparing these results with those from Examples 2 and 3 shows that estimates based on the binomial, the Poisson and the exponential distribution all agree closely. This example illustrates that the epidemiologist should often obtain similar and consistent results, regardless of which of these three basic distribution he or she chooses to model variability in disease occurrence.

TABLE 4 *Summary of incidence rate estimators, basic distributions*

Distribution	Maximum likelihood estimator for incidence rate	Variance estimator
Exponential	A/PT	A/PT^2
Poisson	A/PT	APT^2
Binomial	−ln(1−A/N)/t ≈ A/PT*	A/PT^2

*Approximately, assuming rare disease

In illustrating the relationships of the binomial to the exponential and Poisson distributions, we used a rare disease assumption. However, we do not require this assumption and now show that the close relationships between these distributions hold even for outcomes that are not rare.

Specifically, we suppose that the time until disease occurrence follows an exponential distribution. As noted previously (Table 4), the exponential and the Poisson distribution yield A/PT as the maximum likelihood estimator for the incidence rate. For large samples, maximum likelihood theory implies that these estimators will converge to the true value λ. On the other hand, the binomial distribution leads to IR_B = −ln (1−A/N)/t as the (binomial) estimator for IR. With these assumptions, the expected value for A is Np = N(1−exp (−λt)) so that for large samples, IR_B, like estimators based on the Poisson and exponential distributions, will converge to λ. Thus, we expect close agreement for large samples between results based on these three distributions. Importantly, this argument, as pointed out by a reviewer, does not depend on rare disease.

For small samples, the observed values and A and PT may differ from their expected values. If so, methods of analysis based on one of these three distributions may yield results that differ from those based on either of the other distributions.

RELATIONSHIP TO OTHER DISTRIBUTIONS

In this section we discuss how the three basic distributions discussed here relate to other distributions, such as the Gaussian and hypergeometric, which are often used to study disease-exposure associations.

We often use two binomial distributions to assess the association between disease and a dichotomous exposure when data can be summarized in 2 × 2 tables, as illustrated in Table 1. A limitation of this approach, however, is that two parameters are involved, p_1 and p_2, whereas interest may centre primarily on one—the odds ratio—which summarizes the strength of disease-exposure association. For a cohort study, this odds ratio is the

cumulative incidence odds among the exposed divided by that among the unexposed $(p_1/(1-p_1)) + (p_2/(1-p_2))$. To focus on this measure, statisticians sometimes treat the marginal totals of the summary 2 × 2 table as fixed.[1,2] This statistical device, sometimes called a conditional analysis, eliminates the 'nuisance parameters' and leads us to use of the (non-central) hypergeometric distribution which involves only the parameter of interest, the odds ratio. The conditional argument which leads to use of the hypergeometric distribution, however, is based on the underlying assumption that the number of exposed cases and the number of unexposed cases each have a binomial distribution. The hypergeometric distribution also arises naturally from certain structural models.[13]

For large studies in which the outcome is not too rare, the epidemiologist can base hypothesis tests and confidence intervals on the Gaussian distribution[1,2] since the hypergeometric distribution is approximately Gaussian for large samples. Since the square of a Gaussian test statistic is a χ^2 statistic, a slight extension of the argument justifies use of χ^2 tests. Thus the binomial distribution actually underlies, and can be used to justify, use of the hypergeometric, Gaussian, and χ^2 distributions for analyses of disease-exposure associations in follow-up studies that, like Example 1, involve counts of cases and non-cases.

Similar conditional arguments show that the Poisson distribution actually underlies some applications of the binomial distribution as a method for analysing disease-exposure associations in follow-up studies if summary data involve counts and person-time, like those summarized in Table 2. In particular, the probability that A cases occur in the exposed, given that A+B cases have occurred altogether, is treated under the null hypothesis of no disease-exposure association as binomial, with parameter N = A+B and $p = PT_1/(PT_1 + PT_2)$. The binomial distribution can be derived by assuming that the number of cases in the exposed is Poisson, that the number in the unexposed is Poisson, and then conditioning on the total number of cases. Under the null hypothesis, the ratio of exposed to total cases should occur in proportion to the corresponding ratio of person-time.

In summary, the basic models for disease variability considered here, such as the binomial and Poisson distributions, are consistent with and, in fact, can be used to justify use of the hypergeometric, Gaussian and other distributions for studying disease-exposure associations.[1,2]

Example 5. We illustrate use of conditional analyses to estimate the odds ratio using count data as summarized in Table 1. Assuming that deaths among smokers as well as those among non-smokers follow binomial distributions, the conditioning arguments cited above lead to use of the hypergeometric distribution to estimate the odds ratio. The estimated odds ratio comparing smokers to non-smokers is 1.98 with 95% confidence limits from 1.68 to 2.33. Since the odds ratio approximates the risk ratio for rare disease, this odds ratio is nearly the same as the estimated risk among smokers divided by that among non-smokers: 0.0551 + 0.0286 = 1.93.

Alternatively, if we use person-time data (Table 2) and assume Poisson distributions for the numbers of deaths of smokers and non-smokers, conditional arguments lead to use of the binomial distribution for estimating the incidence rate ratio.[1] The estimated rate ratio is 1.95 with 95% confidence limits from 1.66 to 2.29. These results, based on person-time data and underlying Poisson distributions, show reasonably close agreement with those based on count data and underlying binomial distributions, i.e. 1.98 versus 1.95 and (1.68 to 2.33) versus (1.66 to 2.29).

DISCUSSION

We have reviewed three statistical distributions which epidemiologists commonly use to model variability in disease frequency. Importantly, the three models often lead to similar analytic results and have similar limitations, reflecting the similar underlying assumptions. The similarities become apparent by accounting for the relationship between parameters of the respective distributions. Moreover, these basic models underlie many applications of the Gaussian, hypergeometric and other distributions used to study disease-exposure associations. In summary, then, many of the statistical distributions and approaches used to model disease are closely related, and can often be expected to yield similar results.

In actual applications, the different models lead to different numerical results. As noted previously, different results could arise by chance in small samples, but would tend to be minor for large samples. Differences may also reflect violation of one or more of the underlying assumptions. Moreover, if lack of independence or homogeneity is of concern, then none of these three basic distributions may be appropriate. Careful evaluation and consideration of assumptions could indicate which modifications or alternative models might address the problems and yield a valid result.

We have argued that the binomial, Poisson and exponential distributions have common attributes that often lead to similar results when used to model variability in disease occurrence in one or two exposure groups. Many similarities carry over even when the models are

expanded to incorporate covariates that reflect potential confounding or effect modification. In particular, even though the three distributions are not mathematically equivalent, they do belong to a family of distributions called the 'exponential family'. Statisticians have developed a generalized theory of regression modelling which is based on this exponential family in which these three basic distributions arise as special cases. We can use this approach, called generalized linear models, to study disease occurrence by specifying an appropriate member of the exponential family, such as the binomial distribution, to model variability.[14] Related methods of analysis based on quasi-likelihood approaches or generalized estimating equations,[14,15] those based on models for infectious diseases or dependent events,[9,15,16] and those based on random effects models[17,18] may allow the analyst to proceed if the homogeneity or independence assumptions fail to hold.

REFERENCES

[1] Rothman K J. *Modern Epidemiology*. Boston: Little, Brown, 1986.

[2] Kleinbaum D G, Kupper L L, Morgenstern H. *Epidemiologic Research*. New York, NY: Van Nostrand Reinhold Publications, 1982.

[3] Cox D R, Oakes D. *Analysis of Survival Data*. New York: Chapman and Hall, 1984

[4] Fleiss J L. *Statistical Methods for Rates and Proportions. 2nd Edn*. New York: John Wiley & Sons, 1981.

[5] Breslow N E, Day N E. *Statistical Methods in Cancer Research. Vol. II—The Design and Analysis of Cohort Studies*. Lyon: International Agency for Research on Cancer, 1987.

[6] Hogg R V, Craig A T. *Introduction to Mathematical Statistics. 3rd Edn*. London: Macmillan, 1970.

[7] Doll R, Hill A B. Lung cancer and other causes of death in relation to smoking. *Br Med J* 1956; **5001:** 1071–81.

[8] Parzen E. *Stochastic Processes*. San Francisco: Holden-Day, 1962.

[9] Longini I M, Koopman J S, Haber M, Cotsonis G. Statistical inference for infectious diseases. *Am J Epidemiol* 1988; **128:** 845–59.

[10] Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990; **1:** 421–29.

[11] Cox D R, Snell E J. *Analysis of Binary Data*. New York: Chapman and Hall, 1989.

[12] Cox D R, Isham V. *Point Processes*. New York: Chapman and Hall, 1980.

[13] Greenland S. On the logical justification of conditional tests for two by two contingency tables. *Am Statist* 1991; **45:** 248–51.

[14] McCullagh P, Nelder J A. *Generalized Linear Models*. New York: Chapman and Hall, 1989.

[15] Zeger S L, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42:** 121–30.

[16] Bonney G E. Logistic regression for dependent binary observations. *Biometrics* 1987, **43:** 951–73

[17] Self G S, Prentice R L. Incorporating random effects into multivariate relative risk regression models. In: Moolgavkar S H, Prentice R L (eds) *Modern Statistical Methods in Chronic Disease Epidemiology*. New York: John Wiley & Sons, 1986.

[18] Laird N M, Ware J H. Random-effects models for longitudinal data. *Biometrics* 1982; **38:** 963–74

(Revised version received July 1994)