DETECTION OF CONFOUNDING AND TESTING FOR CROSS-OVER EFFECT IN EPIDEMIOLOGICAL STUDIES

by

Regina C. Elandt-Johnson

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1459

April 1984

DETECTION OF CONFOUNDING AND TESTING FOR CROSS-OVER EFFECT IN EPIDEMIOLOGICAL STUDIES

Regina C. Elandt-Johnson *)
Department of Biostatistics, University of
North Carolina, Chapel Hill, N.C. 27514

ABSTRACT

This is a semi-expository paper on theoretical bases, applicability and applications of X^2-ANOVA-like analysis of incidence rates or prevalence proportions, where the data are presented in the form of a sequence of 2×2 tables corresponding to levels (strata) of a specified variable (risk factor) X. The distributional assumptions for incidence rates (prospective studies) and prevalence proportions (retrospective studies) are reviewed with the emphasis on conditional inference, in each situation. A novel feature of this paper is the finding that the so-called X^2_{homog} (denoted here by X^2_{Diff}) is *not* a test of "homogeneity" in the sense that rate ratios are constant over all strata, but it is a test for *crossing-over*, at least once, of the rate functions in two defined populations. Several examples are given to illustrate the techniques of constructing the tests as well as emphasizing the applicability of stratified X^2-analysis to various problems in epidemiological studies.

Key Words: Incidence rate; Poisson distribution; Binomial distribution; Confounding factor; Multiplicative model; Interaction; Effect modifying factor; Cross-over effect; Homogeneity.

---

*) This work was supported by U.S. National Heart, Lung, and Blood Institute contract NIH-NHLI-712243 from the National Institutes of Health.

1. INTRODUCTION

1.1. Consider an epidemiological or medical comparative study, in which the effects of a specified factor (X) on incidence or prevalence of certain event ($\mathcal{D}$) is investigated in two populations, P_1 and P_2. Thus, formally, three random variables, P, D, and X can be observed on each individual.

Population variable P takes values 1 or 0 and determines two study groups, P_1 and P_2, respectively. They might represent two demographic populations, individuals exposed and non-exposed to occupational hazard, two groups of people with different life styles, diet habits, treatment and control groups in medical investigation, etc.

Response variable D also takes values 1 or 0, which refer to occurrence or non-occurrence of event ($\mathcal{D}$). In epidemiological studies, event $\mathcal{D}$ may represent a certain disease or health disorder; in mortality studies, it represents death, or death from a specific cause.

Concomitant variable X is a characteristic which may be, in some way, associated with P or D or both, and therefore is a reasonable subject for investigation. It can be a discrete variable (e.g. smoking or non-smoking), or a continuous variable such as age, blood pressure, level or serum cholesterol, etc.

By *association* of random variables we understand their stochastic (statistical) *dependence*. We now examine possible associations among P, D, and X.

(i) If *X and D* are associated, then X may be considered as a *prediction* or *prognostic factor* for event $\mathcal{D}$. In particular, if $\mathcal{D}$ is a

harmful event and increase in X increases (decreases) the probability of occurrence of D, then X is a risk (beneficial) factor. It is, however, customary in epidemiological literature to use the term *risk factor* in a broader sense to mean: "X is associated with D."

(ii) Association *between X and P* implies that the *distributions* of X in P_1 and P_2 are *different*. Care should be taken to recognize how this may happen; X might be differently distributed in P_1 and P_2 because genetic and environmental factors affecting X are differently distributed in P_1 and P_2, or the association between X and P may arise from the design of the experiment in selecting P_1 and P_2.

(iii) Of special interest in epidemiological and medical experiments is the association *between X and both D and P*. In this case, X is a *confounding factor*. Or, in other words, if the response (D) depends on factor X, and the distributions of X in P_1 and P_2 are different, then X is a confounding factor.

1.2. Comparison of incidence rates or prevalence proportions in two populations almost always encounters the problem of adjusting for confounding. Often data are stratified by a confounding factor and are represented in the form of a series of 2×2 contingency tables.

First, one may be interested in a sort of 'global test' for comparison of X-adjusted rates (Armitage (1966), Gart (1978)) or X-adjusted proportions Cochran (1954), Mantel and Haenszel (1959)). In Section 2, some of the issues in constructing such tests will be reviewed, emphasizing various assumptions and interpretations of the results.

1.3. Our interest, however, might also be in the analysis of behavior (patterns) of rates in two populations over individual strata. One way of approaching this problem would be by fitting multiplicative (log-linear) models. If the data fit a multiplicative model, then there is no interaction between X and P. (See Section 3.2 (iii).) In Section 4, we present a special approach--a X^2-ANOVA-like method--to detect whether cross-over of the rate functions in two populations occurs.

2. COMPARISON OF TWO SURVIVAL FUNCTIONS: SOME TESTS BASED ON POISSON MODELS

2.1. For convenience, and without loss of generality, we assume that the event ($\mathcal{D}$) of interest is death, and the factor X is *age*, which is clearly associated with mortality. We consider mortality experiences of two study populations, P_1 and P_2, with the data grouped in fixed age intervals. In age determined cohort studies, grouping might be according to follow-up time rather than age. For convenience we will occasionally refer to the ith grouping unit as the ith *stratum*.

For the gth experience (briefly, *sample* from the gth population, g = 1,2) and ith stratum, let A_{gi} denote the amount of person-years (or person-time units) exposed to risk, d_{gi}--the observed number of deaths, and $\hat{\lambda}_{gi} = d_{gi}/A_{gi}$--the observed death rate. For the ith stratum we display the data in the following form

Sample 1			Sample 2		
A_{1i}	d_{1i}	$\hat{\lambda}_{1i}$	A_{2i}	d_{2i}	$\hat{\lambda}_{2i}$

, (2.1)

Let the random variable D_{gi} denote the number of deaths in the gth sample and the ith stratum. If the D_{gi}'s are small and the A_{gi}'s are

large, then (conditional on the A_{gi}'s), the D_{gi}'s can be treated as *independent Poisson variables* with means

$$\mu_{gi} = A_{gi}\lambda_{gi} \quad (g = 1,2;\ i = 1,2,\ldots,I) \ , \tag{2.2}$$

where the λ_{gi}'s are the true death rates.

2.2. Conditional test in the ith stratum. Consider first the single (ith) stratum as shown in (2.1). Assume that D_{1i} and D_{2i} are independent Poisson variates with means given in (2.2). The sum $D_{\cdot i} = D_{1i} + D_{2i}$ also has a Poisson distribution, with mean $\mu_{\cdot i} = \mu_{1i} + \mu_{2i}$. Conditional on $D_{1i} + D_{2i} = d_{\cdot i}$, the distribution of D_{1i} is *binomial* with parameters $d_{\cdot i}$ and $\pi_i = \mu_{1i}/\mu_{\cdot i}$; that is, $D_{gi} \frown b(d_{\cdot i},\pi_i)$. This is a well-known result, first obtained--to the best of my knowledge--by Przyborowski and Wilenski (1939). Such distributions will be used in constructing tests for various hypotheses concerning comparison of two rate functions.

Consider first the single (ith) stratum alone and a null hypothesis

$$H_{i0}\colon\ \lambda_{1i} = \lambda_{2i} \text{ or } \lambda_{1i}/\lambda_{2i} = \gamma_i = 1 \ , \tag{2.3}$$

against *one-sided* alternative

$$H_{iA}\colon\ \lambda_{1i} > \lambda_{2i} \text{ or } \lambda_{1i}/\lambda_{2i} = \gamma_i > 1 \ . \tag{2.4}$$

Let λ_{i0} denote the (hypothetical) death rate when H_{i0} is valid. Then under H_{i0}, $\mu_{gi0} = A_{gi}\cdot\lambda_{i0}$ $(g = 1,2)$, and

$$\pi_{i0} = \mu_{1i0}/\mu_{\cdot i0} = A_{1i}/A_{\cdot i} \ . \tag{2.5}$$

Note that λ_{i0} (which is a nuisance parameter) is here not relevant; π_{i0} is expressed entirely in terms of amounts of person-years exposed to risk.

The null hypothesis (2.3) is then equivalent to

$$H_{i0}: \pi_i = \pi_{i0} = A_{1i}/A_{\cdot i}, \tag{2.6}$$

against the alternative

$$H_{iA}: \pi_i > \pi_{i0}. \tag{2.7}$$

Three tests can be constructed for testing H_{i0}.

(a) *Exact test* based on the binomial distribution, $b(d_{\cdot i}, \pi_{i0})$. For a given significance level α, let $k_i = k_i(\alpha)$ denote the smallest k_i such that

$$\Pr\{D_{1i} \geq k_i | d_{\cdot i}, \pi_{i0}\} = \sum_{r=k_i}^{d_{\cdot i}} \binom{d_{\cdot i}}{r} \pi_{i0}^r (1-\pi_{i0})^{d_{\cdot i}-r} \leq \alpha . \tag{2.8}$$

The null hypothesis, H_{i0}, is rejected if $d_{1i} \geq k_i$.

(b) *Approximate z-test.* For sufficiently large $d_{\cdot i}$, and π_{i0} not too small, a normal approximation to the binomial can be used.

The expected value and variance of D_{1i} conditional on $d_{\cdot i}$ and under H_{i0} are

$$E(D_{1i} | d_{\cdot i}, \pi_{i0}) = E_{1i} = d_{\cdot i}\pi_{i0}, \tag{2.9}$$

and

$$\mathrm{Var}(D_{1i} | d_{\cdot i}, \pi_{i0}) = V_i = d_{\cdot i}\pi_{i0}(1-\pi_{i0}), \tag{2.10}$$

respectively.

The statistic

$$Z_i = (D_{1i}-E_{1i})/\sqrt{V_i} , \tag{2.11}$$

is approximately distributed as unit normal, when H_{i0} is true.

(c) *Approximate X_i^2 test.* Equivalently, the statistic

$$X_i^2 = (D_{1i}-E_{1i})^2/V_i \tag{2.12}$$

is approximately distributed as χ^2 with 1 d.f.

2.3. Stratified analysis. We consider two experiences, P_1 and P_2, in which mortality data are grouped in I age intervals (strata). We restrict ourselves here to the situation in which the age specific rates in P_1 are not smaller than the corresponding rates in P_2; that is, we *assume a model*

$$M: \lambda_{1i} \geq \lambda_{2i} \text{ or } \lambda_{1i}/\lambda_{2i} = \gamma_i \geq 1 \text{ for } all \text{ i}. \tag{2.13}$$

Under this assumption, we wish to test the hypothesis

$$H_0: \lambda_{1i} = \lambda_{2i} \text{ or } \lambda_{1i}/\lambda_{2i} = 1 \text{ for } all \text{ i}, \tag{2.14}$$

against the alternative

$$H_A: \lambda_{1i} > \lambda_{2i} \text{ or } \lambda_{1i}/\lambda_{2i} = \gamma_i > 1 \text{ for at least some i}. \tag{2.15}$$

Assuming that (conditionally on the A_{gi}'s), the D_{gi}'s are independent Poisson variates, we construct a test based on the statistic $D_{1\cdot} = \sum_i D_{1i}$. In fact, we propose two versions of this test.

2.3.1. Version 1 (Statistic X^2_{Comb}). We recall that, conditionally on $d_{\cdot i}$, the variate D_{1i} is binomial with mean $E_{1i} = d_{\cdot i}\pi_{i0}$ and variance $V_i = d_{\cdot i}\pi_{i0}(1-\pi_{i0})$, where $\pi_{i0} = A_{1i}/A_{\cdot i}$. Thus, *conditionally on the sets* $\{d_{\cdot i}\}$ *and* $\{\pi_{i0}\}$, the random variable $D_{1\cdot}$ is approximately normal with mean

$$E_{1\cdot} = \sum_{i=1}^{I} E_{1i} = \sum_{i=1}^{I} d_{\cdot i}\pi_{i0}, \tag{2.16}$$

and variance

$$V = \sum_{i=1}^{I} V_i = \sum_{i=1}^{I} d_{\cdot i}\pi_{i0}(1-\pi_{i0}) \ . \tag{2.17}$$

The statistic

$$Z_{Comb} = (D_{1\cdot} - E_{1\cdot})/\sqrt{V} \ , \tag{2.18}$$

("Comb" for "combined") is approximately distributed as unit normal, and the statistic

$$X^2_{Comb} = (D_{1\cdot} - E_{1\cdot})^2/V \tag{2.19}$$

is approximately distributed as χ^2 with 1 d.f., when H_0 is valid. Either of these can be used in testing H_0.

2.3.2. Version 2 (Statistic X^{*2}_{Comb}). We propose a test based, again, on the statistic $D_{1\cdot}$, but *conditionally on the total number of deaths,* $d.. = d = \sum_g \sum_i d_{gi}$.

As in Section 2.2, let λ_{i0} $(i = 1,2,\ldots,I)$ denote the true death rates, when H_0 is valid. Since the D_{gi}'s are independent Poisson variates, so are $D_{1\cdot}$ and $D_{2\cdot}$. Under H_0 (see (2.14)), we have

$$\mu_{g\cdot} = \sum_{i=1}^{I} A_{gi}\lambda_{i0} \ (g = 1,2), \tag{2.20}$$

and

$$\mu.. = \mu = \sum_{i=1}^{I} A_{\cdot 1}\lambda_{i0} \ . \tag{2.21}$$

Thus, *conditional* on $D_{1\cdot} + D_{2\cdot} = d$, the random variable $D_{1\cdot}$ has a binomial distribution with parameters d and π_0, where

$$\pi_0 = \mu_{1\cdot}/\mu = (\sum_{i=1}^{I} A_{1i}\lambda_{i0})/(\sum_{i=1}^{I} A_{\cdot i}\lambda_{i0}) \ . \tag{2.22}$$

Now, π_0 *does depend* on the nuisance parameters λ_{i0}'s so that, in principle, the test can be affected by the choice of the λ_{i0}'s. If there

is no rationale for using pre-specified λ_{i0}'s, a common and convenient procedure is to use the maximum likelihood estimates for the λ_{i0}'s. In our case, we have

$$\hat{\lambda}_{i0} = d_{\cdot i}/A_{\cdot i}, \tag{2.23}$$

so that

$$\hat{\mu}_{1i} = E_{1i} = A_{1i}\hat{\lambda}_{1i} = d_{\cdot i}\frac{A_{1i}}{A_{\cdot i}} = d_{\cdot i}\pi_{i0}, \tag{2.24a}$$

and similarly,

$$\hat{\mu}_{2i} = E_{2i} = d_{\cdot i}(1-\pi_{i0}). \tag{2.24b}$$

Hence, under H_0

$$\hat{\mu}_{1\cdot} = E_{1\cdot} = \sum_{i=1}^{I} d_{\cdot i}\pi_{i0}, \tag{2.25a}$$

(see also (2.16)), and

$$\hat{\mu}_{2\cdot} = E_{2\cdot} = \sum_{i=1}^{I} d_{\cdot i}(1-\pi_{i0}), \tag{2.26b}$$

so that

$$\hat{\mu}_{\cdot\cdot} = \hat{\mu} = E_{1\cdot} + E_{2\cdot} = d . \tag{2.26}$$

Substituting into (2.22), we obtain

$$\hat{\pi}_0 = E_{1\cdot}/(E_{1\cdot}+E_{2\cdot}) = \frac{1}{d}\sum_{i=1}^{I} d_{\cdot i}\pi_{i0} . \tag{2.27}$$

Therefore, under H_0 and *conditionally on* $D_{1\cdot} + D_{2\cdot} = d$, the distribution of $D_{1\cdot}$ is approximately binomial, $b(d,\hat{\pi}_0)$. Using this distribution, an 'exact' critical region can be constructed in a similar manner as for the ith stratum (cf. (2.8)).

We notice that

$$\text{Est. } E(D_{1\cdot}|d,\hat{\pi}_0) = d\hat{\pi}_0 = \sum_{i=1}^{I} d_{\cdot i}\lambda_{i0} = E_{1\cdot} \ , \tag{2.28}$$

and

$$\text{Est. } \text{Var}(D_{1\cdot}|d,\hat{\pi}_0) = V^* = d\hat{\pi}_0(1-\hat{\pi}_0) = E_{1\cdot}E_{2\cdot}/d = E_{1\cdot}E_{2\cdot}/(E_{1\cdot}+E_{2\cdot}) \tag{2.29}$$

If $\hat{\pi}_0$ is sufficiently large, normal approximation can be used.

The statistic

$$Z^*_{\text{Comb}} = (D_{1\cdot}-E_{1\cdot})/\sqrt{V^*} \ , \tag{2.30}$$

or equivalently,

$$X^{*2}_{\text{Comb}} = (D_{1\cdot}-E_{1\cdot})^2/V^* = d\,\frac{(D_{1\cdot}-E_{1\cdot})^2}{E_{1\cdot}E_{2\cdot}} \tag{2.31}$$

can be used in testing H_0.

We also notice that

$$X^{*2}_{\text{Comb}} = d\,\frac{(D_{1\cdot}-E_{1\cdot})^2}{E_{1\cdot}E_{2\cdot}} = \frac{(D_{1\cdot}-E_{1\cdot})^2}{E_{1\cdot}} + \frac{(D_{2\cdot}-E_{2\cdot})^2}{E_{2\cdot}} \ . \tag{2.32}$$

The following remarks might be useful.

(i) The form of X^{*2}_{Comb} in (2.32) resembles the Peto and Peto (1972) logrank statistic, which was derived for ungrouped data, by using non-parametric techniques.

(ii) Statistic (2.31) was also considered by Armitage (1966). He showed that conditional on the set $\{d_{\cdot i}\}$ (though this is not clearly spelled out in his paper), the statistic (2.31) has to be modified (corrected) to have an approximate χ^2-distribution. In the present context, (2.31) is derived conditionally on the total number of deaths, d.

(iii) The techniques presented here are essentially equivalent to *indirect standardization*. If the λ_{i0}'s are pre-specified, this corresponds to external standardization; when the $\hat{\lambda}_{i0}$'s given by (2.23) are used, this corresponds to internal standardization. Note that our hypothesis H_0 is concerned with comparison of two (non-crossing over) rate functions, and not with the estimation of population effects, so it is legitimate to use standardization.

(iv) It can be shown (the algebra is a bit tedious) that

$$V \leq V^*, \tag{2.33}$$

so that tests constructed in "Version 1," reject H_0 more often than those in "Version 2." If the numbers of events (the $d_{\cdot i}$'s) are small, the statistics in the two versions differ but little. If this is not the case, the discrepancies might be too big to be neglected; in such cases Version 1 is the more appropriate (Haybittle and Freedman (1979)).

(v) Clearly, similar analyses can be used with the person-years exposed to risk, A_{gi}, replaced by the numbers of individuals, n_{gi}, in the midyear population (samples).

(vi) Of course, the situation

$$\left.\begin{aligned} M&: \lambda_{1i} \leq \lambda_{2i} \text{ for all } i\ , \\ H_0&: \lambda_{1i} = \lambda_{2i} \text{ or } \lambda_{1i}/\lambda_{2i} = 1 \text{ for all } i\ , \\ H_A&: \lambda_{1i} < \lambda_{2i} \text{ or } \lambda_{1i}/\lambda_{2i} = \gamma_i < \text{ for some } i \end{aligned}\right\} \tag{2.34}$$

can be treated in the same fashion.

3. DETECTING CONFOUNDING AND INTERACTION

3.1. Confounding. We first consider the problem whether X is a confounding factor; that is, whether the distributions of X in P_1 and P_2 are different.

3.1(i) Analysis of the π_{i0}'s. Suppose that

$$A_{1i}/A_{2i} = c \text{ for } \textit{all } i, \tag{3.1}$$

where c is a constant. It follows that

$$A_{1i}/A_{2i} = \pi_{i0} = c/(1+c) \text{ for } \textit{all } i \tag{3.2}$$

is also constant. If the data represent two independent random samples from two populations, one may construct a test of the null hypothesis: $\pi_{i0} = \pi_0$ for *all* i. However, in this article inferences are made *conditional* on sets $\{A_{1i}\}$, $\{A_{2i}\}$, so that the λ_{i0}'s are true binomial proportions. Of course, it is also necessary to judge what sizes of differences among the π_{i0}'s have a practical significance.

3.1(ii) Pooling the strata. Another way of detecting confounding is to pool the strata, and calculate the test statistic as if *there were a single stratum* (cf. Section 2.2).

In this case, we calculate

$$\pi_0' = A_{1.}/(A_{1.}+A_{2.}), \tag{3.3}$$

$$E(D_{1.}|d,\pi_0') = d\,\pi_0' = E_{1.}' \tag{3.4}$$

and

$$\text{var}(D_{1.}|d,\pi_0') = d\,\pi_0'(1-\pi_0') = E_{1.}'E_{2}'/d\ . \tag{3.5}$$

The statistics

$$Z' = (D_{1.}-E_{1.}')/\sqrt{V'}\ , \tag{3.6}$$

or

$$X'^2 = (D_{1.}-E_{1.}')^2/V' \tag{3.7}$$

can be used to test the hypothesis H_0': $\lambda_1 = \lambda_2$, that is, that the *crude* (overall) rates are the same.

On the other hand, the statistics Z_{Comb} (Z^*_{Comb}) or equivalently X^2_{Comb} (X^{*2}_{Comb}) can be used in testing the hypothesis that the *age adjusted* rates are equal. Notice, that with this formulation of the null hypothesis, the assumption that $\lambda_{1i}/\lambda_{2i} \geq 1$ for *all* i is not required.

If, however, we have proportionate distributions in P_1 and P_2 as defined in (3.1), then

$$\pi_0' = \pi_0 = c/(1+c) \ , \tag{3.8}$$

so that

$$E_{1\cdot}' = E_{1\cdot} \text{ and } V' = V = V^* \ . \tag{3.9}$$

Clearly, if (3.8) holds, then the observed statistics are identical, that is

$$Z' = Z_{Comb} = Z^*_{Comb} \ , \tag{3.10}$$

and

$$X'^2 = X^2_{Comb} = X^{*2}_{Comb} \ . \tag{3.11}$$

Thus, sufficiently large discrepancy between X'^2 and X^2_{Comb} (or X^{*2}_{Comb}) indicates that confounding exists. No formal test involving size of this discrepancy is here suggested; the bigger the discrepancy the more support for using stratification (see Example 1).

3.2. Interaction and effect modification. In our model (M), we only assumed that $\lambda_{1i}/\lambda_{2i} = \gamma_i \geq 1$ for *all* i. If, however, additionally, $\lambda_{1i}/\lambda_{2i} = \gamma \geq 1$, for all i, where γ is a constant, then we conclude that there is *no interaction* between X (age) and D (mortality) (see Section 3.2(iii) and also Elandt-Johnson (1984).)

If the γ_i's are not the same (even if the model $\lambda_{1i}/\lambda_{2i} = \gamma_i \geq 1$ for *all* i holds), then there is an interaction and X is considered as an *effect modifying factor* for response $\mathcal{D}$. How do we test the hypothesis $H_0^{\dagger}$: $\lambda_{1i}/\lambda_{2i} = \gamma$ for *all* i, against the alternative $H_A^{\dagger}$: $\lambda_{1i}/\lambda_{2i} = \gamma_i \neq \gamma$ for at least some i?

(i) A heuristic approach would be to examine the observed ratios $\hat{\lambda}_{1i}/\hat{\lambda}_{2i} = \hat{\gamma}_i$, to obtain some idea from the data whether the hypothesis that X is not a modifying factor for $\mathcal{D}$ might not be rejected.

(ii) Another way of looking at this problem would be to analyze the individual X_i^2's. Of course, their values depend also on the total number of deaths, $d_{\cdot i}$, in each stratum. A better idea can be obtained by comparing the "phi coefficients"

$$\varphi_i = \sqrt{X_i^2/d_{\cdot i}} \quad , \tag{3.12}$$

where $0 \leq \varphi_i \leq 1$ is a kind of correlation coefficient for the ith 2×2 contingency table (see, for example, Fleiss (1981), Section 5.2).

(iii) A formal method would be to fit a *Poisson multiplicative model*

$$\lambda_{gi} = \delta_g \varepsilon_i \quad , \; g = 1,2; \; i = 1,2,\ldots,I \quad , \tag{3.13}$$

(with $\delta_1 + \delta_2 = 1$), where δ_g is the effect of the gth population and ε_i is the effect of the ith stratum. If model (3.13) is valid, then there is *no interaction* between P and X (Elandt-Johnson (1984)). Model (3.13) can be equivalently represented in the form

$$\lambda_{gi} = \exp(\alpha_g + \beta_i). \tag{3.14}$$

Several authors (Bishop *et al* (1975), Breslow and Day (1975), Gail (1978), Gart (1971, 1978), Osborn (1975) among others) have considered

such models, usually for more than two populations. An elegant (theoretical and practical) treatment of Poisson multiplicative models and of inference based on these models for G ($\geq$ 2) populations is given by Anderson (1977).

If the data fit the model (3.13), then we have

$$\lambda_{1i}/\lambda_{2i} = \delta_1/\delta_2 = \gamma \text{ for } all \text{ i} . \tag{3.15}$$

Also, under the assumption that the data fit a multiplicative model, our hypothesis H_0 in (2.13) is equivalent to H_0: $\delta_1 = \delta_2 = \frac{1}{2}$ for *all* i.

4. CROSSING-OVER OF RATE FUNCTIONS. X^2-ANOVA-LIKE ANALYSIS

4.1. Incidence data. So far, we have assumed that the model M: $\lambda_{1i} \geq \lambda_{2i}$ for *all* i holds. Suppose, however, that this assumption cannot be made, nor do the observed λ_{gi}'s support this assumption, that is, we observe $\hat{\lambda}_{1i} > \hat{\lambda}_{2i}$ for some i, and $\hat{\lambda}_{1i'} < \hat{\lambda}_{2i'}$ for some i' $\neq$ i.

The analysis may be the following. We consider again mortality data in two populations, grouped in fixed age intervals.

We first recall a well-known algebraic relationship. Let y_i be an observed value of a variable y, and w_i be a 'weight' attached to y_i. We have

$$\sum_{i=1}^{I} w_i y_i^2 - (\sum_{i=1}^{I} w_i y_i)^2/(\sum_{i=1}^{I} w_i) = \sum_{i=1}^{I} w_i (y_i - \bar{y})^2 , \tag{4.1}$$

where $\bar{y} = (\sum_{i=1}^{I} w_i y_i)/(\sum_{i=1}^{I} w_i)$.

For our data, let

$$y_i = (D_{1i} - E_{1i})/V_i \tag{4.2}$$

be a 'score,' and

$$w_i = V_i \tag{4.3}$$

be the 'weight,' so that

$$\sum_{i=1}^{I} w_i y_i^2 = \sum_{i=1}^{I} (d_{1i}-E_{1i})^2/V_i = \sum_{i=1}^{I} X_i^2 = X^2_{Total} . \tag{4.4}$$

Note that *conditional* on the set $\{d_{\cdot i}\}$, the statistic X^2_{Total} is approximately distributed as χ^2 with I d.f. The X^2_{Total} defined in (4.4) reflects--to some extent--the variation from stratum to stratum, irrespective of the signs (+ or -) of individual scores.

Also the statistic

$$(\sum_{i=1}^{I} w_i y_i)^2/(\sum_{i=1}^{I} w_i) = (D_{1\cdot}-E_{1\cdot})^2/V = X^2_{Comb} , \tag{4.5}$$

(cf. (2.19)) *conditional on* the set $\{d_{\cdot i}\}$, is approximately distributed as χ^2 with 1 d.f. This is a test statistic for *overall* effect of population P_1 relative to P_2 (without assuming that the model M is correct).

The difference

$$X^2_{Diff} = X^2_{Total} - X^2_{Comb}, \tag{4.6}$$

is approximately distributed as χ^2 with (I-1) d.f. and can be used in testing the hypothesis

$$H_0^{(S)}: (\lambda_{1i}/\lambda_{2i} - 1) \text{ does not change sign as i varies,} \tag{4.7}$$

against the alternative

$$H_A^{(S)}: \left.\begin{array}{l} \lambda_{1i} > \lambda_{2i} \quad \text{for some i,} \\ \lambda_{1i'} < \lambda_{2i'} \text{ for some i'.} \end{array}\right\} \tag{4.8}$$

In other words, X^2_{Diff} is a test statistic for detecting whether the rate functions *cross-over at least once* over the strata; it provides a test for a special kind of heterogeneity. (See discussion in Section 6.)

Also notice that X^2_{Comb} and X^2_{Diff} are approximately independent.

If the rates are small, we often have

$$X^{*2}_{Comb} \doteq X^2_{Comb} , \tag{4.8}$$

where X^{*2}_{Comb} is the logrank test statistic for grouped data, defined in (2.31). Although algebraically

$$X^2_{Total} - X^{*2}_{Comb} = X^2_{Diff} \tag{4.9}$$

is correct, we notice that X^2_{Total} is a relevant test statistic if inference is conditional on the set $\{d_{\cdot i}\}$, while X^{*2}_{Comb} is relevant for inference conditional on $d = \sum_i d_{\cdot i}$.

The three statistics X^2_{Total}, X^2_{Comb} and X^2_{Diff} should be considered and interpreted jointly, as will be shown in Examples (Section 5); some questions of joint behavior of these statistics will be discussed in more detail in Section 6.

4.2. Prevalence data. The tests we have discussed in Sections 2, 3, and 4.1 are appropriate for comparisons of *incidence rates* of an event $\mathcal{D}$ in prospective studies. Similar methods can also be used in analysis of prevalence proportions in retrospective epidemiological experiments.

Let n_{gi} be the number of individuals and d_{gi} the number who experienced an event $\mathcal{D}$ in the gth population and the ith stratum. Let $\hat{q}_{gi} = d_{gi}/n_{gi}$ be the estimated prevalence (probability, proportion) of an event $\mathcal{D}$ in the

(gi)th class, and q_{gi} be the corresponding true (but unknown) probability of this event. If the q_{gi}'s are small, then the binomial distribution, $b(n_{gi},q_{gi})$, can be approximated by the Poisson distribution with mean $\mu_{gi} = n_{gi}q_{gi}$. Thus, the methods described in this paper are approximately applicable to prevalence data. If, however, the q_{gi}'s are rather large, these methods might be inappropriate. In such cases, the *conditional* distribution of D_{1i} given $D_{1i} + D_{2i} = d_{\cdot i}$ is *hypergeometric*, and the test criteria for H_0: $q_{1i} = q_{2i}$ for *all* i, should be based on this distribution.

Analysis analogous to those used in deriving X^2_{Comb} (Section 2.3.1) leads to the Mantel-Haenszel (1959) (briefly, M-H) procedure; the only formal difference is the formula for variance in each stratum. It turns out, however, that the variance in the M-H procedure, $V_{i(M-H)}$, can be expressed in terms of the variance V_i defined in (2.10) by the relation

$$V_{i(M-H)} = \frac{n_{\cdot i}-d_{\cdot i}}{n_{\cdot i}} V_i \ . \tag{4.10}$$

The M-H test statistic for the ith stratum is

$$X^2_{i(M-H)} = (D_{1i}-E_{1i})^2/V_{i(M-H)} \ , \tag{4.11}$$

and for all strata combined

$$X^2_{Comb(M-H)} = (D_{1\cdot}-E_{1\cdot})^2/V_{(M-H)} \ , \tag{4.12}$$

where

$$V_{(M-H)} = \sum_{i=1}^{I} V_{i(M-H)} \ . \tag{4.13}$$

Of course,

$$V_{(M-H)} < V < V^* \ . \tag{4.14}$$

(For application, see Example 4.)

4.3. In our analysis of *rates*, we have assumed that these are very small, so that the assumption that the D_{gi}'s are Poisson variables is approximately valid. Suppose, however, that this is not the case. What should we do in such situations? We may 'convert' the rates $\hat{\lambda}_{gi}$'s into proportions $\hat{q}_{gi}$'s, by calculating the 'effective number of initial exposed to risk,' n'_{gi} (Elandt-Johnson and Johnson (1980) Chapter 8).

If n_{gi} is the size of the gth midperiod population (sample) in the ith age interval (stratum), then

$$n'_{gi} \doteq n_{gi} + \frac{1}{2} d_{gi} \quad . \tag{4.15}$$

If the person-years exposed to risk, A_{gi}, over an age (or time) interval of length h_i is given, then

$$n'_{gi} \doteq \frac{1}{h_i}(A_{gi} + \frac{1}{2} h_i d_{gi}). \tag{4.16}$$

Using the n'_{gi}'s *as if they were* integers in binomial distributions, $b((n'_{gi}, q_{gi})$ for D_{gi}, the M-H procedure can be applied as discussed in Section 4.2.

5. EXAMPLES

Three examples are given in this section, mainly for the purpose of illustrating various aspects of inferences and conclusions from the X^2-analysis. The significance level $\alpha = 0.05$ will be used in these analyses.

EXAMPLE 1. The data in Table 1 represent the mortality experience of white males selected at random from two locations (called here, briefly, "cities") and followed, on the average, for 5 years. City 1 is an industrial city with a greater proportion of younger people, while City 2 is a kind of retirement community, with greater proportion of older people.

TABLE 1

COMPARISON OF MORTALITY FROM ALL CAUSES IN TWO USA CITIES

Stratum (i)	Age Group	City 1				City 2				Total						
		A_{1i}	d_{1i}	$\hat{\lambda}_{1i}$	E_{1i}	A_{2i}	d_{2i}	$\hat{\lambda}_{2i}$	E_{2i}	$A_{\cdot i}$	$d_{\cdot i}$	$\hat{\lambda}_{i0}$	$\hat{\gamma}_i$	π_{i0}	V_i	X_i^2
1	30-55	875.06	4	.00457	2.54	503.65	0	0	1.46	1378.71	4	.00290	-	.63469	0.9274	2.30
2	55-60	196.24	4	.02038	2.87	145.73	1	.00686	2.13	341.97	5	.01462	2.97	.57385	1.2227	1.04
3	60-65	184.49	5	.02710	3.47	187.49	2	.01067	3.53	371.98	7	.01882	2.54	.49597	1.7499	1.34
4	65-70	156.87	6	.03825	4.64	282.45	7	.02478	8.36	439.32	13	.02959	1.54	.35707	2.9845	0.62
5	70-75	118.49	6	.05064	2.91	369.52	6	.01624	9.09	488.01	12	.02459	3.12	.24280	2.2062	4.33
6	75+	58.61	6	.10237	5.15	305.78	26	.08503	26.85	364.39	32	.08782	1.20	.16084	4.3192	0.17
TOTAL		1589.76	31		21.58	1794.62	42		51.42	3384.38	73				13.4099	9.80

$X^2_{Total} = 9.80$ (NS); $X^2_{Comb} = 6.62$ (S); $X^2_{Diff} = 3.18$ (NS); $X'^2 = 0.60$

(For details, see Lipid Research Clinics Program (1974).) Clearly (as also can be seen from the π_{i0} column), age is a confounding factor. The resulting X^2-tests are:

Version 1: $X^2_{Total} = 9.80$ (NS); $X^2_{Comb} = 6.62$ (S); $X^2_{Diff} = 3.18$ (NS);

Version 2: $X^2_{Total} = 9.80$ (NS); $X^{*2}_{Comb} = 5.84$ (S); $X^{*2}_{Diff} = 3.96$ (NS).

(S-significant, NS-not significant)

In this example, the results X^2_{Comb} and X^{*2}_{Comb} are similar, each indicating that the age specific mortality rates over the age range 30 to 75+ are higher in City 1 than in City 2. The observed values of the γ_i's are greater than 1 for *all* i, which indicate that the model M: $\lambda_{1i} \geq \lambda_{2i}$ for *all* i, is not contradicted by the data. It is also worthwhile noting that pooling the data in a single stratum, we obtain $X'^2 = 0.60$, which is far different from X^2_{Comb}, and indicates (together with the π_{i0}'s) that the age (X) here is, indeed, a confounding factor.

EXAMPLE 2. The data used in this example are also Lipid Research Clinics Program Follow-Up Study, but the samples were here purposive, with fairly large proportions of individuals with high cholesterol and/or high triglycerides (for detail, see LRCP (1974)). The average follow-up time was about 7 years. Our interest here is in the effect of higher vs. lower level of cholesterol on mortality from Coronary Heart Disease (CHD) (see Table 2) and mortality from Cancer (Table 3) in white males of age $x \geq 30$. The higher level of cholesterol was defined as being the third tertile (approximately, Chol $\geq$ 67th percentile), and the lower level was corresponding to the remaining two first tertiles (approximately, Chol < 67th percentile).

TABLE 2

CHD MORTALITY IN WHITE MALES WITH HIGH AND LOW LEVELS OF CHOLESTEROL

Stratum (i)	Age Group	Cholesterol ≥ 67th PRC				Cholesterol < 67th PRC				Total						
		A_{1i}	d_{1i}	$\hat{\lambda}_{1i}$	E_{1i}	A_{2i}	d_{2i}	$\hat{\lambda}_{2i}$	E_{2i}	$A_{\cdot i}$	$d_{\cdot i}$	$\hat{\lambda}_{i0}$	$\hat{\gamma}_i$	π_{i0}	V_i	X_i^2
1	30-55	6568.98	16	.002436	7.66	13149.86	7	.000532	15.34	19718.74	23	.001166	4.58	.33313	5.1096	13.61
2	55-70	2470.80	23	.009309	13.56	4997.76	18	.003602	27.44	7468.56	41	.005490	2.58	.33083	9.0766	9.82
3	70+	615.54	10	.016246	7.72	1219.34	13	.010662	15.28	1834.88	23	.012535	1.52	.33547	5.1273	1.01
Total		9655.32	49		28.94	19366.96	38		58.06	29022.28	87				19.3135	24.44

$X^2_{Total} = 24.44$ (S); $X^2_{Comb} = 20.84$ (S); $X^2_{Diff} = 3.60$ (NS); $X'^2 = 20.83$

TABLE 3

CANCER MORTALITY IN WHITE MALES WITH HIGH AND LOW LEVELS OF CHOLESTEROL

Stratum (i)	Age Group	Cholesterol ≥ 67th PRC				Cholesterol < 67th PRC				Total						
		A_{1i}	d_{1i}	$\hat{\lambda}_{1i}$	E_{1i}	A_{2i}	d_{2i}	$\hat{\lambda}_{2i}$	$2i$	$A_{\cdot i}$	$d_{\cdot i}$	$\hat{\lambda}_{i0}$	$\hat{\gamma}_i$	π_{i0}	V_i	X_i^2
1	30-55	6568.98	4	.000609	4.33	13149.86	9	.000684	8.67	19718.84	13	.000659	0.89	.33313	2.8880	0.04
2	55-70	2470.80	4	.001619	9.59	4997.76	25	.005002	19.41	7468.56	29	.003883	0.32	.33083	6.4200	4.86
3	70+	615.54	7	.011372	9.06	1219.34	20	.016402	17.94	1834.88	27	.014715	0.69	.33547	6.0190	0.70
Total		9655.32	15		22.98	19366.96	54		46.02	29022.28	69				15.3270	5.61

$X^2_{Total} = 5.61$ (NS); $X^2_{Comb} = 4.15$ (S); $X^2_{Diff} = 0.46$ (NS); $X'^2 = 4.14$

It would be desirable to have age specific population tertiles; lacking these, we estimate them from the data. Also, since there were no repeated measurements of cholesterol during the follow-up time, the tertiles were estimated from the frequency distribution at entry. This is not quite correct, but lacking proper data, it seems reasonable, at this stage.

First we notice that for these data age does not appear to be a confounding factor; the π_{i0}'s are almost the same for the two groups, and the X^2_{Comb} and X'^2_{Comb} are practically the same. However, the effects of cholesterol level on the CHD and on cancer mortality are quite different.

(a) CHD mortality. The mortality rates are greater for the higher level of cholesterol for all ages ($\hat{\gamma}_i > 1$ for *all* i); the $X^2_{Comb} = 20.84$ is highly significant. However, the $\hat{\gamma}_i$'s (and the X^2_i's) decrease substantially with age, indicating that high level of cholesterol as a potential risk factor in CHD mortality is more important in younger ages. It implies that there might be a chol × age interaction.

(b) Cancer mortality. Here we observe an opposite effect. Lower level of cholesterol is a potential risk factor in cancer mortality ($\hat{\gamma}_i < 1$ for *all* i). It seems that this may be more apparent in age group 55-70, but since mortality data are rather sparse, they do not support strongly this view.

EXAMPLE 3. The data in Table 4 represent prevalence of cardiac event (a manifestation of possible ischemic heart disease) in white males and white females selected for the Follow-Up Study in the Lipid Research Clinics Program. The π_{i0}'s indicate that age x is, in these data, a confounding factor. For ages 30-59, $\hat{\gamma}_i > 1$, while for ages ≥ 60, $\hat{\gamma}_i < 1$, which suggests that there is a cross-over effect. The M-H X^2-statistics calculated for these data

TABLE 4

PREVALENCE OF CARDIAC EVENT IN WHITE FEMALES AND WHITE MALES

Stratum (i)	Age Group	White Females				White Males				Total							
		n_{1i}	d_{1i}	$\hat{q}_{1i}$	E_{1i}	n_{2i}	d_{2i}	$\hat{q}_{2i}$	E_{2i}	$n_{\cdot i}$	$d_{\cdot i}$	$\hat{q}_{i0}$	$\hat{\gamma}_i$	π_{i0}	V_i	$V_{i(M-H)}$	$X^2_{i(M-H)}$
1	30-35	316	12	.03797	8.58	384	7	.01823	10.42	700	19	.02714	2.08	.45143	4.7052	4.5775	2.56
2	35-40	284	23	.08099	14.22	355	9	.02535	17.78	639	32	.05008	3.19	.44444	7.9013	7.5056	10.27
3	40-45	305	12	.03934	9.00	373	8	.02145	11.00	678	20	.02950	1.83	.44985	4.9497	4.8037	1.87
4	45-50	319	16	.05016	12.96	321	10	.03115	13.04	640	26	.04062	1.61	.49844	6.5000	6.2358	1.48
5	50-55	248	22	.08871	18.35	333	21	.06306	24.65	581	43	.07401	1.41	.42685	10.5199	9.7413	1.37
6	55-60	243	24	.09877	21.56	253	20	.07905	22.44	496	44	.08871	1.25	.48992	10.9955	10.0201	0.59
7	60-65	146	12	.08219	15.03	126	16	.12698	12.97	272	28	.10294	0.65	.53676	6.9622	6.2455	1.47
8	65-70	121	13	.10744	15.88	100	16	.16000	13.12	221	29	.13122	0.67	.54751	7.1846	6.2418	1.33
9	70-75	81	3	.03704	7.95	82	13	.15854	8.05	163	16	.09816	0.23	.49693	3.9999	3.6072	6.79
10	75-80	28	6	.21429	8.52	18	8	.44444	5.48	46	14	.30435	0.48	.60870	3.3346	2.3197	2.74
11	80+	19	4	.21053	3.06	12	1	.08333	1.94	31	5	.16129	2.53	.61290	1.1863	0.9949	0.89
TOTAL		2110	147		135.11	2357	129		140.89	4467	276				68.2392	62.2931	31.36

$X^2_{Total} = 31.36$ (S); $X^2_{Comb} = 2.27$ (NS); $X^2_{Diff} = 29.09$ (S); $X'^2 = 4.29$

confirm this hypothesis; $X^2_{Comb} = 2.27$ is not significant, while $X^2_{Diff} = 29.09$ (with 10 d.f.) is highly significant. The cardiac event in white males aged 30-59 is observed less often (and in white females, more often) than expected, while the situation is reversed for ages ≥ 60.

The X^2-analysis restricted to ages 30-59 gives the following M-H X^2-statistics:

$$X^2_{Total(M-H)} = 18.14 \text{ (NS; } X^2_{Comb(M-H)} = 13.80 \text{ (S)}$$

$$X^2_{Diff(M-H)} = 4.34 \text{ (NS), and } X'^2 = 13.84.$$

The interpretation is straightforward: there is an excess (statistically significant) of cardiac events in females, for ages 30-59.

Similar results, but in the opposite direction, are obtained from analysis of these data for ages ≥ 60.

6. DISCUSSION

6.1. Our main concern in this paper is with comparison of rates (or more precisely, rate functions) in two populations, but most of the remarks below apply also to proportions in stratified analysis of categorical data.

First, the X^2_{Comb} (or X^{*2}_{Comb}) statistic is often termed X^2_{assoc} ("Chi-square for association"). In this context, this means association between D and P. If the model M ($\lambda_{1i} \geq \lambda_{2i}$ for *all* i) is valid, and X^2_{Comb} is significant, then this, indeed, indicates that the incidence rates in P_1 are higher than in P_2. If, however, model M does not hold, and X^2_{Comb} is not significant, it does not imply that there is lack of association; in fact, there is often crossing-over of rate functions, as can be seen from Example 3. Therefore, the term "association" seems to be not quite relevant; in this paper, the term

"combined" indicates only cumulative effect over strata as opposed to X'^2 (Chi-square 'pooled') in which the strata are pooled into a single stratum.

6.2. A second, even more important problem, arises with the interpretation of X^2_{Diff}, which is also commonly termed X^2_{homog} ("*of* homogeneity"), sometimes without even defining what "homogeneity" means.

Consider the model (3.13)

$$\lambda_{gi} = \delta_g \varepsilon_i \ , \ g = 1,2; \ i = 1,2,\ldots,I \ . \qquad (6.1)$$

This model implies that there is a constant (multiplicative) population effect, δ_g, over all strata, and there is no interaction between P and D. It follows that

$$\lambda_{1i}/\lambda_{2i} = \delta_1/\delta_2 = \delta \text{ for } \textit{all } i \ , \qquad (6.2)$$

so that the *relative risk* over all strata is constant. In this sense, the model may be regarded as *homogeneous*. An appropriate goodness of fit test of a multiplicative model would then be the test *of* homogeneity (*for* heterogeneity; cf. Section 3.2 (iii)). The X^2_{Diff} is not a suitable test statistic for detecting heterogeneity (Examples 1 and 2(a)); it is constructed to detect cross-over effect of rate functions.

Also, some authors use the term "homogeneity" in a narrower sense: that there is no population effect, that is, $\delta_1 = \delta_2$ in (6.1) (e.g. Armitage (1966).)

The above remarks apply also to proportions and to the use of M-H procedure.

6.3. Mantel *et al.* (1977) choose to define "homogeneity" in terms of *odds ratios*. Using the notation of our Section 4.2, the odds ratio for the ith stratum is

$$o_i = \frac{q_{1i}}{1-q_{1i}} \cdot \frac{1-q_{2i}}{q_{2i}} = \frac{q_{1i}}{q_{2i}} \cdot \frac{1-q_{2i}}{1-q_{1i}} \ . \qquad (6.3)$$

If q_{1i} and q_{2i} are very small, then $o_i \doteq q_{1i}/q_{2i}$, that is, the odds ratio approximates the relative risk. Otherwise, o_i has no sensible meaning, in my opinion, and should be avoided, even though it is a mathematically "handy" index.

In the examples given by Mantel *et al.*(1977), the $\hat{q}_{1i}$ and $\hat{q}_{2i}$ are large (of order 0.50 to 0.99) and so the discussion about use of Zelen's (1971) test (which appears to be based on our $X^2_{Diff(M-H)}$ is confusing and misleading. *If* Zelen intended to use this test for equality of odds ratios over all strata, it would, indeed, not be appropriate for this purpose; but it is *not* "an invalid test for any purpose," as Mantel *et al.* (1977) conclude at the end of their discussion.

6.4. In the X^2-analysis, commonly encountered patterns are the following:

	X^2_{Total}	X^2_{Comb}	X^2_{Diff}	
(1)	NS	NS	NS	No crossing-over of rate functions
(2)	NS	S	NS	
(3)	S	S	NS	
(4)	S	NS	S	Crossing-over of rate functions .
(5)	S	S	S	

The question arises: Does the pattern (5) *always* indicate substantial crossing-over? There are two situations where this might not be so.

(i) If the mortality data (the $d_{\cdot i}$'s) are large, any small difference between observed and expected values would lead to big X^2-value, even if such a difference is not practically important.

(ii) A more complicated situation can arise, even when the $d_{\cdot i}$'s are rather small. To investigate this question formally, consider, for simplicity, only two strata, 1 and 2. It is easy to show that the right-hand side of (4.1) can be written in the form

$$\sum_{i=1}^{2} w_i(y_i-\bar{y})^2 = \frac{w_1 w_2}{w_1+w_2}(y_1-y_2)^2 \quad . \tag{6.4}$$

Expressing this in terms of our variables defined in (4.2) and (4.3), we obtain

$$X^2_{Diff} = \frac{V_1 V_2}{V_1+V_2}\left(\frac{D_{11}-E_{11}}{V_1} - \frac{D_{12}-E_{12}}{V_2}\right)^2 \quad . \tag{6.5}$$

Suppose that $D_{11} - E_{11} > 0$ and $D_{12} - E_{12} < 0$, that is, crossing over is observed in the sample, and suppose that X^2_{Diff} is significant. If $|D_{11}-E_{11}|/V_1$ is relatively large as compared to $|D_{12}-E_{12}/V_2$, then changing the sign of $(D_{12}-E_{12})$ would have a minimal effect on X^2_{Diff}, that is, X^2_{Diff} still could be significant even if $D_{12} - E_{12} > 0$. From my experience, however, $|D_{11}-E_{11}|/V_1$ must be *very* large as compared to $|D_{12}-E_{12}|/V_2$ for this to occur. In practice such situations are very rare, so we have to use an artificial example to produce such a situation.

EXAMPLE 4. Consider the following data (below, in Table 5).

TABLE 5

Stratum	Sample 1			Sample 2		
(i)	n_{1i}	d_{1i}	$\hat{q}_{1i}$	n_{2i}	d_{2i}	$\hat{q}_{2i}$
1	100	20	0.20	100	2	0.02
2	100	5	0.05	100	3	0.03

Note that the observed proportion in sample 1 and stratum 1, $q_{11} = 0.20$, is much bigger than the remaining q's, so that the interaction populations × strata is almost obvious. The resulting X^2's are as follows.

$X^2_{Total(M-H)} = 17.05(S)$, $X^2_{Comb(M-H)} = 14.46(S)$ and $X^2_{Diff} - 2.39(NS)$; this means that the observed model M: $\hat{q}_{11} > \hat{q}_{21}$, $\hat{q}_{12} > \hat{q}_{22}$, is not contradicted by the analysis, even with this obvious interaction effect.

(ii) Suppose, however, that $d_{11} = 70$, while the remaining d_{gi}'s are the same as in Table 5, so that there is still no observed crossing over. In this case, $X^2_{Total(M-H)} = 100.35(S)$, $X^2_{Comb(M-H)} = 91.15(S)$ and $X^2_{Diff(M-H)} = 9.20(S)$. Now, X^2_{Diff} is significant because the unusual structure in stratum 1 ($\hat{q}_{11} = 0.70$) as opposed to stratum 2 ($\hat{q}_{21} = 0.05$). Of course, such data are unlikely to occur in a real situation, and *if* they were to occur, no statistical test would be required.

(iii) Also note that if in Table 5 we would have $d_{11} = 20$, $d_{21} = 2$, but $d_{21} = 3$ and $d_{22} = 5$ (that is, d_{21} and d_{22} have exchanged their positions, so that $\hat{q}_{11} > \hat{q}_{21}$ but $\hat{q}_{21} < \hat{q}_{22}$), then $X^2_{Total(M-H)} = 17.05(S)$, $X^2_{Comb(M-H)} = 9.38(s)$, and $X^2_{Diff(M-H)} = 7.67(S)$. It appears that X^2_{Diff} is rather sensitive for detecting even small cross-over phenomenona.

REFERENCES

1. Andersen, E.B. (1977). Multiplicative Poisson models with unequal cell rates. Scand. J. Statist. 4, 153-158.

2. Armitage, P. (1966). The χ^2 test for heterogeneity of proportions after adjustment for stratification. J. Roy. Statist. Soc. Ser. B, 28, 150-163

3. Bishop, Y.M.M., Fienberg, S.G., and Holland, P.W. (1975). Discrete Multivariate Analysis. Cambridge, Mass., MIT Press.

4. Breslow, N. and Day N.E. (1975). Indirect standardization and multiplicative models for rates with reference to the age adjustment of cancer incidence and relative frequency data. J. Chron. Dis. 28, 289-303.

5. Cochran, W.G. (1954). Some methods of strengthening the common X^2-tests. Biometrics 10, 417-451.

6. Elandt-Johnson, Regina C. (1984). Statistical interaction revisited. Inst. of Statistics Mimeo Series, No. 1457, March 1984, Dept. of Biostatistics, UNC, Chapel Hill, N.C.

7. Elandt-Johnson, Regina C. and Johnson, Norman L. (1980). Survival Models and Data Analysis, (Chapter 8) J. Wiley and Sons, New York.

8. Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions (Chapter 10). J. Wiley and Sons, New York.

9. Gail, M. (1978). The analysis of heterogeneity for indirect standardized mortality ratios. J. R. Statist. Soc. Ser. A, 141, 224-234.

10. Gart, J.J. (1971). The comparison of proportions: a review of significance tests, confidence intervals and adjustments for stratification. Rev. Inter. Statist. Inst. 39, 148-169.

11. Gart, J.J. (1978). The analysis of ratios and cross-product ratios of Poisson variates with application to incidence rates. Commun. Statist. (Theory and Methods) A7 (10), 917-937.

12. Haybittle, J.L. and Freedman, L.S. (1979). Some comments on the log-rank test statistic in clinical trials applications. The Statistician, London 28, 199-208.

13. Lipid Research Clinics Program (LRCP) (1974). Protocol of the LIpid Research Clinics Prevalence Study. Central Patient Registry and Coordinating Center, Dept. of Biostatistics, UNC, Chapel Hill, N.C.

14. Mantel, N., Brown, Ch., and Byar, D.P. (1977). Test for homogeneity of effect in an epidemiologic intestigation. Amer. J. Epid. 106, 125-129.

15. Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst. 22, 719-748.

16. Osborn, J. (1975). A multiplicative model for the analysis of vital statistics rates, Appl. Statist. 24, 75-84.

17. Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). J. R. Statist. Soc. Ser. A, 135, 185-198.

18. Przyborowski, J. and Wilenski, H. (1939). Homogeneity of results in testing samples from Poisson series with application for testing clover seeds for dodder. Biometrika 31, 313-323.

19. Zelen, M (1971). The analyses of several 2 × 2 contingency tables. Biometrika 58, 129-137.