CONFIDENCE LIMITS, BASED ON PROSPECTIVE STUDIES, FOR RISK RATIOS AND ETIOLOGIC FRACTIONS STANDARDIZED FOR CONFOUNDING VARIABLES
I. THEORY

by

Eugene Sobel

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1061

April 1976

CONFIDENCE LIMITS, BASED ON PROSPECTIVE STUDIES, FOR RISK RATIOS AND ETIOLOGIC FRACTIONS STANDARDIZED FOR CONFOUNDING VARIABLES

I. THEORY

Eugene Sobel*

*Eugene Sobel is a research fellow in the Departments of Epidemiology and Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, N.C. 27514. This research was supported by Grant No. 1-T32-HL007055, National Heart and Lung Institute from the National Institutes of Health.

ABSTRACT

It is shown that the usual estimators of numerator and denominator of standardized risk ratios and etiologic fractions are jointly asymptotically normal when the data come from prospective studies. Standardization for a categorical confounding variable with an arbitrary number of categories is allowed; the number of exposure levels allowed is also arbitrary. Consequently, Fieller's method can be employed to obtain asymptotically exact confidence intervals.

I. INTRODUCTION

After much work and many long retrospective and prospective studies, epidemiologists believe that they now know some of the major, "causative" risk factors associated with some of the more prevalent chronic diseases. Thus they are now interested in conducting massive programs in an attempt to get people to alter their habits and living modes so as to reduce risk factor levels and therefore presumably the prevalence of the disease in question. In order to help determine, for a given chronic disease, which risk factors to attack epidemiologists have begun to study a "population parameter" they call the <u>etiologic fraction</u> EF. (Quotes have been used because, although the EF depends on some real parameters of the target population, it cannot be calculated from complete statistical knowledge of the population— the EF is an abstraction.) By definition

$$\text{EF} = \text{that proportion of the disease "caused" by the risk factors in question.} \tag{1}$$

We shall be more precise later. If the EF is large, then a successful intervention program would significantly reduce the prevalence of the disease. If it is small then probably an intervention program would not be "worthwhile". The validity of these statements depends upon two assumptions: one for the current generations and one for future generations. It is assumed that, for the current generations, a reduction in exposure levels prior to the onset of the disease will reduce the incidence to something close to the non-exposed group; for future generations, it is assumed that those who would be exposed, if no intervention occurred, are not genetically different (as a group) vis-á-vis the disease in question from those who would not be exposed anyway. Seemingly, the former assumption is more tenuous.

EF is a ratio and both numerator and denominator must be estimated. Unbiased estimates or at least approximately unbiased estimates are available. When what are termed confounding variables are considered and various levels of the risk factors are also considered (thus when a more detailed categorical approach is employed) a confidence interval (CI) has not been known for prospective studies.

In this paper we present asymptotically exact CI's for EF and two other population parameters, the (directly) standardized risk ratio, SRR, and the (indirectly) standardized mortality (morbidity) ratio, SMR. We assume that the data have been obtained from a prospective study in which individuals have been obtained by simple random sampling, or its equivalent, and only the overall sample size is fixed.

The basic method is due to Fieller [1]. In a later paper we plan to extend the estimation of EF, SRR, SMR to non-categorical approaches, ie. multiple regression and discriminant analysis.

Many authors have contributed to the literature on EF, SRR and SMR. For a rather extensive bibliography see Gart [2]; see also Miettinen [3,4].

II. NOTATION, DEFINITIONS AND THE STUDY DESIGN

a) Notation

We are concerned with the occurrence, D, or non-occurrence, $\bar{D}$, of some disease. We are also interested in one or more risk factors which, for purposes of analysis, have been categorized according to "levels of exposure". By combining levels for each risk factor we have

K = number of levels of exposure to risk factors.

Non-exposure will be denoted by 0, while k, $k = 1, \cdots, K$, will denote a given exposure level. There may also be confounding variables (categorical) such that

i) the event disease given exposure level k, $D|k$, is not independent of the confounding variables, $k = 0, \cdots, K$,

and

ii) the joint distribution of the confounding variables is not the same from exposure level to exposure level.

Consequently these confounding variables ought to be considered in any estimation of relative risk or EF. Let

C = total number of confounding variable categories.

For example, being overweight and increasing one's weight are believed to be risk factors for normotensives developing hypertension at a future time. We can dichotomize each: over- or not overweight and weight change, $\geq$ 10 lbs. or < 10 lbs. for a given number of years. Thus K = 3 and not overweight and weight change < 10 lbs. would constitute non-exposure. But it is true that $D|k$ is dependent upon age and current diastolic blood pressure, DBP. By dichotomizing age and using three levels of normal DBP we get C = 6.

Let p with subscripts be the probability of the event described by the subscripts: for example,

Dki = event of becoming diseased and being at exposure level k and in confounding category i

and

$D|ki$ = event of becoming diseased given exposure level k and confounding category i.

A dot, • , in a subscript means that the variable corresponding to that position has been summed out. Thus

$\cdot k|i$ = event of being at exposure level k given confounding category i

and

$D\cdot|k$ = event of becoming diseased given exposure level k.

We assume none of the probabilities are 0.

b) <u>Definitions</u>

The three parameters of interest can now be written

$$SRR_k = \frac{\sum_{i=1}^{C} p_{\cdot i|0}\, p_{D|ki}}{p_{D\cdot|0}} = \frac{\sum_{i=1}^{C} p_{\cdot 0i}\, p_{D|ki}}{p_{D0\cdot}} \qquad (2)$$

$$SMR_k = \frac{p_{D\cdot|k}}{\sum_{i=1}^{C} p_{\cdot i|k}\, p_{D|0i}} = \frac{p_{Dk\cdot}}{\sum_{i=1}^{C} p_{\cdot ki}\, p_{D|0i}} \qquad (3)$$

for $k = 1, \cdots, K$, and

$$EF_e = \frac{\sum_{k\epsilon e} p_{\cdot k\cdot}\, p_{D\cdot|k} - \sum_{k\epsilon e} \sum_{i=1}^{C} p_{\cdot ki}\, p_{D|0i}}{p_{D\cdot\cdot}}$$

$$\equiv \frac{p_{De\cdot} - \overset{o}{p}_{De\cdot}}{p_{D\cdot\cdot}} \qquad (4)$$

where e is any subset of exposure levels not including non-exposure.

In words,

SRR_k = ratio of the probability of becoming diseased given exposure level k if the confounding distribution were that in the non-exposed group to the probability of becoming diseased given non-exposure,

and

SMR_k = ratio of the probability of becoming diseased given exposure level k to the probability of becoming diseased given non-exposure if the confounding variable distribution among the non-exposed group were that of the group at exposure level k.

$\mathcal{e}$ is the set of exposure levels towards which an intervention campaign is contemplated, while

$$\overset{o}{p}_{D\mathcal{e}\cdot} = \text{probability of becoming diseased and being in the group at exposure levels in } \mathcal{e} \text{ if they were non-exposed.}$$

Thus we arrive at the "definition" given in (1).

c) Study Design

As mentioned we are concerned with a prospective study in which participants have been randomly selected. All participants are disease-free at initiation of this study. Let n = overall sample size. Subscripts on n will have the same meaning as before. Thus at study completion we will have C 2 x (K+1) tables:

Confounding Category i

	Risk Level					Totals
	0	1	2	...	K	
Diseased	n_{D0i}	n_{D1i}	n_{D2i}		n_{DKi}	$n_{D\cdot i}$
Not Diseased	$n_{\bar{D}0i}$	$n_{\bar{D}1i}$	$n_{\bar{D}2i}$		$n_{\bar{D}ki}$	$n_{\bar{D}\cdot i}$
Totals	$n_{\cdot 0i}$	$n_{\cdot 1i}$	$n_{\cdot 2i}$		$n_{\cdot ki}$	$n_{\cdot\cdot i}$

Various real life problems, such as change of exposure level and loss to the study, will be ignored. Because of the study design, all entries in in the table are random variables. We ignore the fact that the sample space is finite.

III. JOINT ASYMPTOTIC DISTRIBUTION OF ESTIMATORS OF NUMERATORS AND DENOMINATORS

In this section we prove that the usual, obvious estimators of numerator and denominator of SRR_k, SMR_k and EF_e are jointly asymptotically normal. We give the means and variances and covariances.

Since only n is fixed, there is a non-zero probability that an estimator presented below will be 0/0. We define such a ratio as 1. This means that, for example,

$$E\left(\frac{n_{D0i}}{n_{\cdot 0i}} \mid n_{\cdot 0i}\right) = \begin{cases} 1 & , n_{\cdot 0i} = 0 \\ p_{D|0i} & , n_{\cdot 0i} > 0 \end{cases}$$

$$\mathrm{Var}\left(\frac{n_{D0i}}{n_{\cdot 0i}} \mid n_{\cdot 0i}\right) = \begin{cases} 0 & , n_{0i} = 0 \\ p_{D|0i}(1 - p_{D|0i})/n_{\cdot 0i} & , n_{\cdot 0i} > 0. \end{cases}$$

Because the writing of the remainder of the paper will be easier and more concise if we just ignore complications caused by events whose probability goes to 0 as $n \longrightarrow \infty$ we shall, for example, write

$$E\left(\frac{n_{D0i}}{n_{\cdot 0i}}\right) = p_{D|0i}$$

instead of

$$E\left(\frac{n_{D0i}}{n_{\cdot 0i}}\right) = p_{D|0i}\, P(n_{\cdot 0i} > 0) + P(n_{\cdot 0i} = 0) \longrightarrow p_{D|0i}\,.$$

Thus we shall refer to estimators as being unbiased when in fact they are only, but rapidly, asymptotically unbiased.

Definitions

i) $$\widehat{SRR}_k = \frac{\sum_{i=1}^{C} \frac{n_{\cdot 0i}\, n_{Dki}}{n\, n_{\cdot ki\cdot}}}{\frac{n_{D0\cdot}}{n}} = \frac{\sum_{i=1}^{C} \hat{p}_{\cdot 0i}\, \hat{p}_{D|ki}}{\hat{p}_{D0\cdot}} \qquad (5)$$

ii) $$\widehat{SMR}_k \frac{\frac{n_{Dk\cdot}}{n}}{\sum_{i=1}^{C} \frac{n_{\cdot ki}\, n_{D0i}}{n\quad n_{\cdot 0i}}} = \frac{\hat{p}_{Dk\cdot}}{\sum_{i=1}^{C} \hat{p}_{\cdot ki}\, \hat{p}_{D|0i}} \qquad (6)$$

iii) $$\widehat{EF}_e = \frac{\frac{n_{De\cdot}}{n} - \sum_{i=1}^{C} \frac{n_{\cdot ei}\, n_{D0i}}{n\quad n_{\cdot 0i}}}{\frac{n_{D\cdot\cdot}}{n}} = \frac{\hat{p}_{De\cdot} - \hat{p}^{o}_{De\cdot}}{\hat{p}_{D\cdot\cdot}} \,. \qquad (7)$$

where

$$n_{De\cdot} = \sum_{k\epsilon e} n_{Dk\cdot}$$

$$n_{\cdot ei} = \sum_{k\epsilon e} n_{\cdot ki} \,.$$

It is of course clear that

$$\frac{\hat{p}_{DA\cdot} - p_{DA\cdot}}{\sqrt{p_{DA\cdot}(1 - p_{DA\cdot})/n}} \longrightarrow N(0, 1)$$

where A is any subset of {0, 1, ⋯ , K}. We, in particular, refer to A = {k}, k = 0, ⋯ , K, A = $\mathcal{e}$ and A = {0, 1, ⋯ , K}.

Lemma 1

$$\frac{\hat{p}_{D|ki} - p_{D|ki}}{\sqrt{p_{\cdot ki}^{-1}\, p_{D|ki}\left(1 - p_{D|ki}\right) / n}} \longrightarrow N(0, 1).$$

Proof

Let

$$Y_n = \frac{p_{Dki}\left(\hat{p}_{\cdot ki}^{-1} - p_{\cdot ki}^{-1}\right)}{\sigma(\hat{p}_{Dki})}$$

and

$$X_n = \frac{\hat{p}_{D|ki} - p_{D|ki}}{\sigma(\hat{p}_{Dki})}$$

where

$$\sigma^2(\hat{p}_{Dki}) = p_{Dki}(1 - p_{Dki}) / n.$$

Then $Y_n = y$ is equivalent to

$$n/n_{\cdot ki} = p_{Dki}^{-1}\,\sigma(\hat{p}_{Dki})\,y + p_{\cdot ki}^{-1} = a(\sqrt{n})y + b \longrightarrow p_{\cdot ki}^{-1} \quad \text{as } n \longrightarrow \infty$$

for fixed y. Thus

$$P(X_n \le x \mid Y_n = y) = P\left(\frac{\frac{n_{Dki}}{n_{\cdot ki}} - p_{D|ki}}{\sqrt{p_{D|ki}(1 - p_{D|ki})/n_{\cdot ki}}} \le \frac{x\sigma(\hat{p}_{Dki})}{\sqrt{p_{D|ki}(1 - p_{D|ki})/n_{\cdot ki}}} \;\middle|\; \frac{n}{n_{\cdot ki}} = a(\sqrt{n})y + b,\ y\right)$$

$$\longrightarrow P\left(Z \le \frac{x\sqrt{p_{Dki}(1 - p_{Dki})}}{\sqrt{p_{\cdot ki}^{-1}\,p_{D|ki}(1 - p_{D|ki})}}\right)$$

where $Z \sim N(0, 1)$. Thus X_n is independent of Y_n and

$$X_n \longrightarrow N\left(0,\ \frac{\sqrt{p_{\cdot ki}^{-1}\,p_{D|ki}(1 - p_{D|ki})}}{\sqrt{p_{Dki}(1 - p_{Dki})}}\right)$$

and consequently

$$\frac{\hat{p}_{D|ki} - p_{D|ki}}{\sqrt{p_{\cdot ki}^{-1}\,p_{D|ki}(1 - p_{D|ki})/n}} \longrightarrow N(0, 1).$$

QED.

Lemma 2

$$X_n = \frac{\hat{p}_{\cdot Ai}\,\hat{p}_{D|ki} - p_{\cdot Ai}\,p_{D|ki}}{\sqrt{p_{D|ki}^{-1}\,p_{\cdot Ai}(1 - p_{\cdot Ai})/n}} \longrightarrow N(0, 1)$$

where A is a subset of $\{0,\ldots,K\}$, $k \notin A$ and $\hat{p}_{\cdot Ai} = n_{\cdot Ai}/n$.

Proof

From Lemma 1 we have

$$Y_n = \frac{\hat{p}_{D|ki} - p_{D|ki}}{\sqrt{p_{\cdot ki}^{-1}\,p_{D|ki}(1 - p_{D|ki})/n}} \longrightarrow N(0, 1).$$

Now $Y_n = y$ is equivalent to

$$\hat{p}_{D|ki} = n_{Dki}\Big/n_{\cdot ki} = \sigma(Y)y/\sqrt{n} + p_{D|ki}$$

where $\sigma^2(Y) = p_{\cdot ki}^{-1}\,p_{D|ki}(1 - p_{D|ki})$. Thus

$$P(X_n \le x \mid Y_n = y) \longrightarrow P\left(\sqrt{p_{D|ki}^{3}}\,Z + ay \le x\right)$$

where $Z \sim N(0, 1)$ and

$$a = \frac{p_{D|ki}\sqrt{p_{\cdot Ai}\left(1 - p_{D|ki}\right)}}{\sqrt{p_{\cdot ki}\left(1 - p_{\cdot Ai}\right)}} .$$

Therefore

$$P(X_n \le x) = \int_{-\infty}^{\infty} P(X_n \le x \mid Y_n = y)dF_{Y_n}(y) \longrightarrow \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{x - ay}{\sqrt{p_{D|ki}}}} \phi(u)\phi(y)dudy = \int_{-\infty}^{x} \phi(z)dz$$

The last equality is obtained by making the transformation

$$y = y$$

$$z = \sqrt{p_{D|ki}^3}\; u + ay$$

and intergrating out y. QED.

The following lemma is presented without proof since a proof is just a straight forward application of conditional expectations.

Lemma 3

i) $$\mathrm{Var} \sum_{i=1}^{c} \hat{p}_{\cdot 0i}\, \hat{p}_{D|ki} = \frac{1}{n}\left(\sum_i p_{D|ki}^2\, p_{\cdot 0i} + \sum_i p_{\cdot 0i}^2\, p_{\cdot ki}^{-1}\, p_{D|ki}(1 - p_{D|ki}) - \left(\sum_i p_{\cdot 0i}\, p_{D|ki}\right)^2\right)$$

$$\mathrm{Cov}\left(\sum_{i=1}^{c} \hat{p}_{\cdot 0i}\, \hat{p}_{D|ki},\ \hat{p}_{D0\cdot}\right) = \frac{1}{n} \sum_i p_{D|ki}(p_{D0i} - p_{\cdot 0i}p_{D0\cdot})$$

ii) $$\mathrm{Var} \sum_{i=1}^{c} \hat{p}_{\cdot ki}\, \hat{p}_{D|0i} = \frac{1}{n}\left(\sum_i p_{D|0i}^2\, p_{\cdot ki} + \sum_i p_{\cdot ki}^2\, p_{\cdot 0i}^{-1}\, p_{D|0i}(1 - p_{D|0i}) - \left(\sum_i p_{\cdot ki}\, p_{D|0i}\right)^2\right)$$

$$\mathrm{Cov}\left(\sum_{i=1}^{c} \hat{p}_{\cdot ki}\, \hat{p}_{D|0i},\ \hat{p}_{Dk\cdot}\right) = \frac{1}{n} \sum_i p_{D|0i}(p_{Dki} - p_{\cdot ki}\, p_{Dk\cdot})$$

iii) $\text{Var}\ \hat{p}^{o}_{De\cdot} = \frac{1}{n}\left(\sum_i p^2_{D|0i}\, p_{\cdot ei} + \sum_i p^2_{\cdot ei}\, p^{-1}_{\cdot 0i}\, p_{D|0i}(1 - p_{D|0i}) - \left(\sum_i p_{\cdot ei}\, p_{D|0i}\right)^2\right)$

$$\text{Cov}\left(\hat{p}_{De\cdot}, \hat{p}^{o}_{De\cdot}\right) = \frac{1}{n}\left(\sum_i p_{Dei}\, p_{D|0i} - p_{De\cdot}\, p^{o}_{De\cdot}\right)$$

$$\text{Cov}\left(\hat{p}_{D\cdot\cdot}, \hat{p}_{De\cdot} - \hat{p}^{o}_{De\cdot}\right) = \frac{1}{n}\left(p_{De\cdot}(1 - p_{D\cdot\cdot}) - \sum_i p_{D|0i}\left(p_{\cdot ei}(1 - p_{D|0i} - p_{D\cdot\cdot}) + p_{Dei}\right)\right)$$

The last requirement before application of Fieller's method is the joint asymptotic normality of the appropriate random variables, for example, $\left\{\hat{p}_{\cdot 0i}\ \hat{p}_{D|ki},\ i = 1, \cdots, C\right\}$. We shall not write out the proofs here. A proof would be accomplished by showing that the variables ar asymptotically jointly normal given $Y_{ni} = y_i$, where the Y_{ni} are defined as in the proof of Lemma 2. This conditional distribution will have the proper mean and covariance structure so that upon integrating out the condition we arrive at joint normality.

IV. ASYMPTOTIC CONFIDENCE INTERVALS

Now suppose X and Y are jointly normal with means μ_X and μ_Y and covariance matrix (σ_{ij}). Then

$$P\left(|X - \ \ Y| < z_{\alpha/2}\sqrt{\sigma_{11} - 2\gamma\sigma_{12} + \gamma^2\sigma_{22}}\right) = 1 - \alpha$$

Upon squaring and rearranging we have

$$P\left(0 < (z^2\sigma_{22} - Y^2)\gamma^2 + 2(XY - z^2\sigma_{12})\gamma + (z^2\sigma_{11} - X^2)\right) = P(0 < A\gamma^2 + B\gamma + C) = 1 - \alpha$$

where $z = z_{\alpha/2}$. Then, as long as $A < 0$ and the zeros of the quadratic are real and distinct, they constitute the lower and upper endpoints of a $(1 - \alpha)$ confidence interval (CI). This procedure is due to Fieller.

Using Lemma 3, construction of CI's for SRR_k, SMR_k and EF_ℓ is straight forward. The CI's are given in Theorem 1 without further proof. We note however, that as $n \longrightarrow \infty$, in each CI presented, $A < 0$, $B < 0$, $C > 0$ and the zeros are real and distinct.

<u>Theorem 1</u>

Let

$$\hat{S} = \frac{-\hat{B} + \sqrt{\hat{B}^2 - 4\hat{A}\hat{C}}}{2\hat{A}}, \quad \hat{L} = \frac{-\hat{B} - \sqrt{\hat{B}^2 - 4\hat{A}\hat{C}}}{2\hat{A}}$$

Then $(\hat{S}, \hat{L})$ constitute asymptotic $(1 - \alpha)$ CI for

i) SRR_k where

$$\hat{A} = z^2 \widehat{Var}\ \hat{p}_{DO\cdot} - \hat{p}^2_{DO\cdot}$$

$$\hat{B} = 2\left(\hat{p}^2_{DO\cdot} \sum_i \hat{p}_{\cdot Oi}\ \hat{p}_{D|ki} - z^2 \widehat{Cov}\left(\hat{p}_{DO\cdot}, \sum_i \hat{p}_{\cdot Oi}\ \hat{p}_{D|ki}\right)\right)$$

$$\hat{C} = z^2\widehat{\text{Var}}\left(\sum_i \hat{p}_{\cdot 0i}\hat{p}_{D|ki}\right) - \left(\sum_i \hat{p}_{\cdot 0i}\ \hat{p}_{D|ki}\right)^2$$

ii) SMR_k where

$$\hat{A} = z^2\widehat{\text{Var}}\left(\sum_i \hat{p}_{\cdot ki}\hat{p}_{D|0i}\right) - \left(\sum_i \hat{p}_{\cdot ki}p_{D|0i}\right)^2$$

$$\hat{B} = 2\left(\hat{p}_{Dk\cdot}\sum_i \hat{p}_{\cdot ki}\ \hat{p}_{D|0i} - z^2\widehat{\text{Cov}}\left(\hat{p}_{Dk\cdot}, \sum_i \hat{p}_{\cdot 0i}\ \hat{p}_{D|ki}\right)\right)$$

$$\hat{C} = z^2\widehat{\text{Var}}\ \hat{p}_{Dk\cdot} - \hat{p}^2_{Dk\cdot}$$

iii) EF_e where

$$\hat{A} = z^2\widehat{\text{Var}}\ \hat{p}_{D\cdot\cdot} - \hat{p}^2_{D\cdot\cdot}$$

$$\hat{B} = 2\left(\hat{p}_{D\cdot\cdot}\left(\hat{p}_{De\cdot} - \hat{p}^o_{De\cdot}\right) - z^2\ \widehat{\text{Cov}}\left(\hat{p}_{D\cdot\cdot}, \hat{p}_{De\cdot} - \hat{p}^o_{De\cdot}\right)\right)$$

$$\hat{C} = z^2\left(\widehat{\text{Var}}\ \hat{p}_{De\cdot} - 2\widehat{\text{Cov}}\left(\hat{p}_{De\cdot}, \hat{p}^o_{De\cdot}\right) + \widehat{\text{Var}}\ \hat{p}^o_{De\cdot}\right) - \left(\hat{p}_{De\cdot} - \hat{p}^o_{De\cdot}\right)^2$$

The estimators of the variances and covariances need to be asymptotically consistent. This is satisfied if the probabilities in these quantities are replaced by their (consistent) estimates.

We note that the estimators and CI for SRR_k and SMR_k may be considered conditioned on $n_{\cdot k\cdot}$, $k = 0, \cdots, K$, being fixed. Asymptotically the results are identical to the unconditioned situation. The expected values of numerator and denominator of EF_e, given the $n_{\cdot k\cdot}$, are functions of these $n_{\cdot k\cdot}$ and will only be asymptotically unbiased since

$n_{\cdot k \cdot}/n \longrightarrow p_{\cdot k \cdot}$. Therefore there seems to be no reason to consider conditional estimators and CI's.

V. STRATIFIED SAMPLING STUDY DESIGN

If stratified random sampling is employed, so that $n_{\cdot k \cdot}$, $k = 0, \cdots, K$, are pre-determined, then an argument similar to that in the preceding section leads to CI's for SRR_k and SMR_k. It is necessary to have $n_{\cdot k \cdot}/n_{\cdot 0 \cdot}$ converge to a limit. If that limit is $p_{\cdot k \cdot}/p_{\cdot 0 i}$ then the stratified design is asymptotically equivalent to the simple random sample design, as stated above.

EF_{ℓ} presents a different problem. In order to use a stratified design we must have prior knowledge of the strata relative sizes: the $p_{\cdot k \cdot}$. If the $n_{\cdot k \cdot}$, $k = 0, \ldots, K$, are chosen so that

$$n_{\cdot k \cdot} \longrightarrow p_{\cdot k \cdot}$$

then stratification is asymptotically more efficient. Specifically, with $\gamma = EF_{\ell}$ and X and Y estimators of numerator and denominator,

$$\mathrm{Var}\,\sqrt{n}\,(X - \gamma Y) - \mathrm{Var}\,\sqrt{n}\,(X - \gamma Y)_{st} \longrightarrow \sum_{k \in \ell} p_{\cdot k \cdot}^{-1} \left(p_{Dk \cdot}(1 - \gamma) - \sum_{i=1}^{C} p_{\cdot ki}\, p_{D|0i} \right)^2 + \gamma^2 \sum_{k \notin \ell} p_{\cdot k \cdot}^{-1}\, p_{DK \cdot}^{2} > 0$$

The specific formulas for the CI are given in Sobel, Part II, [5]. We do not have any specific results concerning optimal allocation.

REFERENCES

[1] Fieller, E.C., "The Biological Standardization of Insulin," _Journal of the Royal Statistical Society_, Supplement, 7 (1940), 1-64.

[2] Gart, John J., "The Comparizon of Proportions: A Review of Significance Tests, Confidence Intervals and Adjustments for Stratification," _Review of the International Statistical Institute_, 39, No. 2 (1971), 148-169.

[3] Miettinen, Olli S., "Proportion of Disease Caused or Prevented by a Given Exposure, Trait or Intervention," _American Journal of Epidemiology_ 99 (May 1974), 325-332.

[4] Miettinen, Olli S., "Estimability and Estimation in Case-Referent Studies," _American Journal of Epidemiology_, 101 (February 1976), 226-235.

[5] Sobel, Eugene, "Confidence Limits, Based on Prospective Studies, for Risk Ratios and Etiologic Fractions Standardized for Confounding Variables, II. Equations and Examples," _The Institute of Statistics Mimeo Series_, Department of Biostatistics, University of North Carolina, Chapel Hill, N.C. 27514.