A STATISTICAL PROBLEM ARISING FROM RETROSPECTIVE STUDIES

JEROME CORNFIELD
OFFICE OF BIOMETRY, NATIONAL INSTITUTES OF HEALTH

1. Introduction

A recurring problem in medical statistics is the determination of the relative risks of developing a disease, say lung cancer, among two or more subclasses of a population, say, smokers and nonsmokers. Ordinarily, the risk for any subclass is estimated as the ratio of the number of cases of the disease developing in that subclass to the total number of persons in it, while an estimate of the risk for one subclass relative to another is provided by the ratio of the estimated absolute risks. Studies which start in this fashion with populations classified into subgroups, for each of which one counts the number of new cases of a disease which develop during some subsequent period of time are ordinarily referred to as "forward-looking" or "prospective" studies.

One may also be concerned with other types of relative risk, for example, the relative risk of dying from a disease or of having a disease. These different relative risks need not be the same for any one disease, and in cases where they are not it is customary to attempt to estimate all three. The relative risk of developing a disease is usually referred to as the relative incidence, the relative risk of dying from it as the relative mortality, and the relative risk of having it during some specified interval of time as the relative prevalence. In diseases, such as lung cancer, where the outcome is usually fatal and the interval between detection and death is relatively constant, the difference between these three different measures of relative risk will be small. In such cases it is common to choose that relative risk which can be estimated most easily. Thus, in prospective studies of lung cancer an estimate of relative mortality is usually preferred to one of relative incidence or prevalence because (a) the death registration system provides a complete enumeration which is lacking for newly developed or for existing cases and (b) diagnosis of cause of death is usually more accurate.

The risk of developing, having, or dying from any one disease in any one year is small. For this reason prospective studies designed to supply estimates of any one of the three relative risks must cover large numbers of persons, usually kept under observation for several years. An alternative method of gathering data, which avoids the necessity of observing large numbers of persons without the disease, but which, as usually done, supplies only estimates of relative prevalence is now commonly referred to as a retrospective study. In such a study one starts with a population (or a sample of it) classified into groups having and not having the disease, and determines for each group the proportion belonging to some subclass. Thus, one might classify a population into those having and not having lung cancer and

then determine the proportion of smokers in each group. The studies are termed retrospective because the determination of the subclass into which an individual falls requires looking back at his past behavior. If we denote by P the proportion of the population falling into the diseased group (that is, the prevalence rate), by p_1 the proportion of the diseased group falling into a subclass, and by p_2 the proportion of the nondiseased group falling into that subclass then the prevalence rate for that disease for members of the subclass is

$$p_1P/[p_1P + p_2(1 - P)] \tag{1.1}$$

for nonmembers of the subclass

$$(1 - p_1)P/[(1 - p_1)P + (1 - p_2)(1 - P)] \tag{1.2}$$

and the prevalence among members of the subclass relative to nonmembers is

$$\frac{p_1}{(1 - p_1)} \frac{(1 - p_1)P + (1 - p_2)(1 - P)}{p_1P + p_2(1 - P)} . \tag{1.3}$$

For most investigations P is sufficiently small relative to p_2 and $(1 - p_2)$ to write for (1.3)

$$\frac{p_1}{(1 - p_1)} \frac{(1 - p_2)}{p_2} . \tag{1.4}$$

We shall henceforth refer to (1.4) as the relative risk of having the disease. In a retrospective study in which a sample of n_1 individuals with the disease and n_2 normals are studied to yield X_1 diseased and X_2 normals falling into the subclass, it seems natural [1] to estimate the relative risk (1.4) with the statistic

$$\frac{X_1}{n_1 - X_1} \frac{n_2 - X_2}{X_2} . \tag{1.5}$$

It is the purpose of the present paper to consider the problem of obtaining single and simultaneous interval estimates of the relative risk, (1.4). This problem leads to the familiar ground of chi-square and the contingency table, but because the point of departure is interval estimation, and not hypothesis testing, the route may be new.

2. Exact confidence limits

The unconditional probability that samples of n_1 and n_2 individuals from populations in which proportions p_1 and p_2 have some characteristic will yield X_1 and X_2 individuals with that characteristic is

$$\binom{n_1}{X_1} p_1^{X_1} q_1^{n_1 - X_1} \binom{n_2}{X_2} p_2^{X_2} q_2^{n_2 - X_2} , \tag{2.1}$$

where $q = 1 - p$. The conditional probability of the observations for the subset of samples in which all marginal totals are fixed by the condition

$$X_1 + X_2 = m \tag{2.2}$$

is then, as can easily be verified,

$$C\binom{n_1}{X_1}\binom{n_2}{m - X_1}z^{X_1} \tag{2.3}$$

where $z = p_1q_2/p_2q_1$, the relative risk, and

$$\frac{1}{C} = \sum_{X_1=0}^{n_1}\binom{n_1}{X_1}\binom{n_2}{m - X_1}z^{X_1},$$

a distribution which has also been considered by Patnaik [2] and Stevens [3]. The conditional probability of the sample observation X_1, given the marginal totals thus depends only on z, the unknown parameter which we wish to estimate. An interval estimate now follows immediately by a well-known argument [4, p. 234].

Denote by z_2 the solution for z of the equation

$$\sum_{y=0}^{X_1} C\binom{n_1}{y}\binom{n_2}{m - y}z^y = \frac{\alpha}{2} \tag{2.4}$$

and by z_1 the solution for z of

$$\sum_{y=X_1}^{n_1} C\binom{n_1}{y}\binom{n_2}{m - y}z^y = \frac{\alpha}{2}. \tag{2.5}$$

Then the probability that the statement

$$z_1 \leqq z \leqq z_2 \tag{2.6}$$

is correct is equal to or greater than $1 - \alpha$.[1] This result is of course closely related to Fisher's exact test of independence for the 2×2 table (see section 21.02 in [5]), the relation being that the test of significance will reject the null hypothesis for a given set of data when and only when the confidence limits (2.6) computed from the same set of data fail to include unity (both computations being of course at the same significance level). Using tables of the binomial coefficients, such as Fry's appendix table III [6] it is possible to solve equations (2.4) and (2.5) by numerical methods and obtain the desired confidence limits. Except when min (n, m) is quite small, however, this is a difficult computation, the attempt to avoid which leads to the investigation of asymptotic approximations.

3. Large-sample confidence limits

In seeking the limiting distribution for (2.3), which we now denote by $P(X)$, we are faced with an initial difficulty arising from our inability to evaluate the constant C, in consequence of which we are unable even to derive exact expressions for the expectation and variance of X. We avoid this difficulty by seeking instead the limiting distribution of the ratio

$$\frac{P(X)}{P(\tilde{X})} \tag{3.1}$$

[1] This defines equal-tailed confidence limits, which in view of the skewness of the distribution (2.3) may be biased. There is no difficulty in defining unbiased limits, if they are desired.

where $\tilde{X}$ is the mode of the distribution (2.3) and is defined by the inequality

$$(3.2)\qquad \frac{(\tilde{X}+1)(n_2-m+\tilde{X}+1)}{(n_1-\tilde{X})(m-\tilde{X})} \geqq z \geqq \frac{\tilde{X}(n_2-m+\tilde{X})}{(n_1-\tilde{X}+1)(m-\tilde{X}+1)}.$$

For large samples it is sufficient to write

$$(3.3)\qquad \frac{\tilde{X}(n_2-m+\tilde{X})}{(n_1-\tilde{X})(m-\tilde{X})} = z.$$

If we substitute Stirling's formula for factorial n in (2.3), expand all terms of the form $\log(1+X)$ to the quadratic and set terms of the form

$$(3.4)\qquad \frac{\tilde{X}+\frac{1}{2}}{\tilde{X}},\quad \frac{n_1-\tilde{X}+\frac{1}{2}}{n_1-\tilde{X}},\quad \frac{m-\tilde{X}+\frac{1}{2}}{m-\tilde{X}},\quad \frac{n_2-m+\tilde{X}+\frac{1}{2}}{n_2-m+\tilde{X}}$$

equal to unity, we have as a limiting expression

$$(3.5)\qquad -2\log\frac{P(X)}{P(\tilde{X})} = (X-\tilde{X})^2\left[\frac{1}{\tilde{X}}+\frac{1}{n_1-\tilde{X}}+\frac{1}{m-\tilde{X}}+\frac{1}{n_2-m+\tilde{X}}\right].$$

To obtain the value of the maximum ordinate we use (2.3) to write

$$(3.6)\qquad \frac{1}{P(\tilde{X})} = \sum_{X=0}^{n_1}\frac{P(X)}{P(\tilde{X})},$$

and using (3.5), and approximating the summation in (3.6) with an integration from $-\infty$ to $+\infty$ we obtain

$$(3.7)\qquad P(\tilde{X}) = \frac{1}{\sqrt{2\pi}}\left[\frac{1}{\tilde{X}}+\frac{1}{n_1-\tilde{X}}+\frac{1}{m-\tilde{X}}+\frac{1}{n_2-m+\tilde{X}}\right]^{1/2}.$$

We thus conclude that the limiting distribution for (2.3) is normal with mean $\tilde{X}$, defined by (3.3) and variance

$$(3.8)\qquad \frac{1}{\left[\frac{1}{\tilde{X}}+\frac{1}{n_1-\tilde{X}}+\frac{1}{m-\tilde{X}}+\frac{1}{n_2-m+\tilde{X}}\right]}.$$

Denote by $\tilde{X}_2$ the largest real root of the quartic in $\tilde{X}$

$$(3.9)\qquad (\tilde{X}-X-\tfrac{1}{2})^2\left[\frac{1}{\tilde{X}}+\frac{1}{n_1-\tilde{X}}+\frac{1}{m-\tilde{X}}+\frac{1}{n_2-m+\tilde{X}}\right] = \chi_\alpha^2,$$

and by $\tilde{X}_1$ the smallest real root of

$$(3.10)\qquad (\tilde{X}-X+\tfrac{1}{2})^2\left[\frac{1}{\tilde{X}}+\frac{1}{n_1-\tilde{X}}+\frac{1}{m-\tilde{X}}+\frac{1}{n_2-m+\tilde{X}}\right] = \chi_\alpha^2$$

where χ_α^2 is the upper α per cent point of the chi-square distribution with one degree of freedom, the 1/2 arises as the usual approximate correction for the discontinuity

of the distribution (2.3), and X is the sample observation. Denote by z_2' and z_1' the values of z obtained by substituting $\tilde{X}_2$ and $\tilde{X}_1$ in (3.3). Then the probability that the statement

$$\tilde{X}_1 \leqq \tilde{X} \leqq \tilde{X}_2 \tag{3.11}$$

is correct is asymptotically equal to $1 - \alpha$, and since z is a monotonic function of X, the asymptotic probability that the statement

$$z_1' \leqq z \leqq z_2' \tag{3.12}$$

is correct is also $1 - \alpha$. The conceptual similarity between the limits (3.11) and those derived by E. B. Wilson [7] in 1927 for a single binomial variate should be noted.

4. Application of approximate limits to small samples

We now consider a single example designed to throw some light on the numerical adequacy of the approximate limits (3.12) when applied to small samples. The data are given in table I. It will be noted that the smallest marginal total is 14 and the

TABLE I

DISTRIBUTION OF PHYSICIANS WITH AND WITHOUT LUNG CANCER BY SMOKING STATUS

Smoking Status	Lung Cancer Patients	Controls*
Nonsmoker	3	11
Smoker	60	32
Total	63	43

(SOURCE: Wynder and Cornfield [8]).
* To keep the numbers small we have used only one of the three control groups given in the original article.

observed value of X is 3, so that the sample is small and the distribution skew. The test of the approximate procedure would consequently seem to be a stringent one.

To obtain 95 per cent confidence limits about the sample result we set $\chi_\alpha^2 = 3.841$ and find[2] as the solution of equations (3.9) and (3.10)

$$\tilde{X}_2 = 6.905 \qquad \tilde{X}_1 = 0.815 \tag{4.1}$$

so that

$$z_2' = 0.6229 \qquad z_1' = 0.0296\,. \tag{4.2}$$

[2] The most convenient numerical method that we have found for solving the quartic is iterative One starts with an initial approximation to the root and uses it to estimate the variance. This gives an improved estimate of the root. When one of the cell entries is small, relative to others as in table I, a good initial approximation is given by using only the two smallest cell entries, thus reducing the quartic to a quadratic, which is easily handled. When all cell entries are of the same magnitude, substituting X for $\tilde{X}$ in the variance gives a good initial approximation.

Thus at the 95 per cent level of confidence the risk for nonsmokers is at most 62 per cent of that of smokers and may be as little as 3 per cent. The adequacy of this approximation may be determined by substituting z_2' in equation (2.4) and z_1' in equation (2.5) and noting the extent to which the left-hand side of each equation differs from .025. This calculation is shown in detail in table II, from which it may

TABLE II

ESTIMATION OF TAIL AREAS CORRESPONDING TO APPROXIMATE 95 PER CENT CONFIDENCE INTERVALS $z_1' = 0.0296$, $z_2' = 0.6229$ AND APPROXIMATE 99 PER CENT CONFIDENCE LIMITS $z_1' = 0.0209$, $z_2' = 0.8790$.

X (1)	$\binom{63}{X}\binom{43}{14-X}$ (2)	(2) $(.6229)^X$ (3)	(2) $(.0296)^X$ (4)	(2) $(.8790)^X$ (5)	(2) $(.0209)^X$ (6)
0	7.838^{10}	7.838^{10}	7.838^{10}	7.838^{10}	7.838^{10}
1	2.305^{12}	1.436^{12}	6.823	2.026^{12}	4.817
2	2.996^{13}	1.162^{13}	2.624	2.315^{13}	1.309
3	2.284^{14}	5.520	5.916^{9}	1.551^{14}	2.085^{9}
4	1.142^{15}	1.719^{14}	0.877	6.818	0.218
5	3.964	3.717	0.090	2.080^{15}	0.016
6	9.853	5.755	0.007	4.544	0.001
7	1.783^{16}	6.488	.....	7.228	.....
8	2.361	5.350	.....	8.415	.....
9	2.278	3.217	.....	7.137	.....
10	1.577	1.387	.....	4.343	.....
11	7.599^{16}	4.163^{13}	.....	1.839	.....
12	2.409	8.220^{12}	.....	5.126^{14}	.....
13	4.502^{14}	9.567^{11}	.....	8.419^{13}	.....
14	3.739^{13}	4.950^{10}	.....	6.147^{12}	.....
Total	0–14	28.824^{14}	17.975^{10}	37.051^{15}	14.197^{10}
	0–3	6.834^{13}	.....	1.804^{14}	.....
	3–14	.....	6.890^{9}	.....	2.320^{9}
Per cent in tail	.....	2.37	3.83	0.49	1.63

Each entry should be multiplied by 10 raised to the power given in the upper right hand corner.

be noted that one tail area is 2.4 per cent, the other 3.8 per cent. For 99 per cent confidence limits we set $\chi_\alpha^2 = 6.635$ and find as solutions of equations (3.9) and (3.10)

$$\tilde{X}_2 = 7.94$$

(4.3)

$$\tilde{X}_1 = 0.59$$

so that

$$z_2' = 0.8790$$

(4.4)

$$z_1' = 0.0209.$$

As shown in table II one tail area is at the 0.5 per cent level, the other at the 1.6 per cent level. For this set of data therefore the effect of the approximation is to place us at the 6.2 per cent level when we wish to operate at the 5 per cent level and at the 2.1 per cent level when we wish to operate at the 1 per cent level.

The agreement between the actual and estimated tail areas would appear to be closer than one might expect in view of the considerable departure of the exact distributions from normality. Thus, the distribution yielding the 3.8 per cent tail is violently skewed and has the bulk of the probability concentrated at 0 and 1, in view of which the disagreement between 3.8 and 2.5, while perhaps larger than one might wish, is smaller than one has any right to expect. Wilson's quadratic approximation to confidence limits around a single binomial proportion [7] has this same property. Thus, with a sample of size 1 yielding none with the characteristic the exact 95 per cent confidence limit is, of course, $0 \leqq P \leqq .95$, while the approximate one (with a continuity correction) is $0 \leqq P \leqq .76$.

5. The r × s contingency table

In practice one is rarely satisfied with a single interval estimate of relative risk, but wishes instead to have several simultaneous estimates. The problem is illustrated by the data in table III, which shows the distribution of lung cancer and

TABLE III

DISTRIBUTION OF PERSONS WITH AND WITHOUT LUNG CANCER BY SMOKING STATUS

Smoking Status	With Lung Cancer		
	Adenocarcinoma	Epidermoid Carcinoma	Controls
Nonsmokers	4	15	56
Smokers:			
Pipe and cigar only	2	13	68
Cigarettes only	31	298	240
Cigarettes plus pipes and cigars	9	146	154
Total	46	472	518

(SOURCE: Breslow et al. [9]).

control patients by smoking status and a simultaneous breakdown of lung cancer patients as between two histologic types. It has been suggested, first that smokers have an excess risk of developing epidermoid, but not adenocarcinoma, and secondly that cigarette smokers have a greater excess risk of developing lung cancer than do pipe or cigar smokers. We propose to consider, from the point of view of interval estimation, what evidence the data in table III contain on these points.

Perhaps the most common way of attacking questions like this with data like those in table III is a chi-square test using independent single degree of freedom contrasts, each at the same predetermined level of significance. Such a procedure does not seem appropriate here because (a) we do not know which of the many possible orthogonal breakdowns to use, (b) we wish to have interval estimates of relative risk, and not tests of hypotheses, (c) we are not concerned with six independent questions, each to be separately tested, but rather with several facets of the single question of tobacco as a possible factor in the etiology of lung cancer.

We propose instead to consider the question from the point of view of confidence regions. We shall seek a closed region in the parameter space which we may, with a known probability of error, assert encloses all the parameters that determine the

distribution of the observations in table III. We may then, in a fashion indicated by Scheffé in connection with the analysis of variance [10], enumerate as many of the particular sets of parameter points falling inside this region as we please, and the probability that the totality of these enumerations is incorrect in any respect is the probability that the region itself does not include the true parameter point.

To obtain this region we note first that the unconditional probability that samples of n_j individuals ($j = 1, 2, \cdots, s$) from populations in which proportions p_{ij} ($i = 1, 2, \cdots, r$) have some characteristic will yield X_{ij} individuals with that characteristic is

$$\prod_{j=1}^{s} n_j! \prod_{i=1}^{r} \frac{p_{ij}^{X_{ij}}}{X_{ij}!} \tag{5.1}$$

where

$$\sum_{i=1}^{r} X_{ij} = n_j ,$$

$$\sum_{i=1}^{r} p_{ij} = 1 .$$

The conditional probability of the observations for the subset of samples in which all marginal totals are fixed by the conditions

$$\sum_{j=1}^{s} X_{ij} = m_i \tag{5.2}$$

is, as can easily be verified,

$$C \prod_{j=1}^{s} \frac{n_j!}{\prod_{i=1}^{r} X_{ij}!} \prod_{j=1}^{s-1} \prod_{i=1}^{r-1} z_{ij}^{X_{ij}} \tag{5.3}$$

where

$$z_{ij} = \frac{P_{ij}P_{rs}}{P_{is}P_{rj}}$$

and C is determined by the condition that the sum of (5.3) over its range is unity, and the rs variables X_{ij} are subject to the $r + s - 1$ linear restraints imposed by the fixed margins. The physical interpretation of z_{ij} is as follows. Let the sample of size n_s be control patients and the rth category in each sample be nonsmokers. Then z_{ij} is the risk that smokers of the ith category will have the jth disease relative to that for nonsmokers.

We find the limiting distribution for (5.3), which we denote by $P(X_{ij})$, exactly as in section 3. We denote the values of X_{ij} at the point of maximum density of (5.3) by $\tilde{X}_{ij}$, where, in large samples

$$\frac{\tilde{X}_{ij}\left[n_s - \sum_{1}^{r-1} m_i + \sum_{1}^{(r-1)}\sum_{1}^{(s-1)} \tilde{X}_{ij}\right]}{\left[m_i - \sum_{j=1}^{s-1} \tilde{X}_{ij}\right]\left[n_j - \sum_{i=1}^{r-1} \tilde{X}_{ij}\right]} = z_{ij} . \tag{5.4}$$

Then, by substituting Stirling's formula and making the other approximations of section 3, we obtain

$$-2 \log \frac{P(X_{ij})}{P(\tilde{X}_{ij})} = \sum_1^r \sum_1^s \frac{(X_{ij} - \tilde{X}_{ij})^2}{\tilde{X}_{ij}} . \tag{5.5}$$

We conclude that the limiting distribution of (5.3) is multivariate normal and that in consequence the positive definite quadratic form given by the right-hand side of (5.5) is distributed as chi-square with $(r - 1)(s - 1)$ degrees of freedom. In that case the required confidence region in the $\tilde{X}_{ij}$ is defined by

$$\sum_1^r \sum_1^s \frac{(X_{ij} - \tilde{X}_{ij})^2}{\tilde{X}_{ij}} \leqq \chi_\alpha^2 [(r - 1)(s - 1)] \tag{5.6}$$

where X_{ij} are observed values, $\tilde{X}_{ij}$ are the variables of the parameter space and the right-hand side is the upper α per cent point of the chi-square distribution with $(r - 1)(s - 1)$ degrees of freedom. A corresponding region about the z_{ij} is obtained from (5.4) in view of the fact z_{ij} is monotonic in X_{lk} for all i, j, l and k.

Thus, if we return to table III and set $\alpha = .05$, the confidence region becomes

$$\frac{(4 - \tilde{X}_{11})^2}{\tilde{X}_{11}} + \cdots + \frac{[9 - (46 - \tilde{X}_{11} - \tilde{X}_{21} - \tilde{X}_{31})]^2}{46 - \tilde{X}_{11} - \tilde{X}_{21} - \tilde{X}_{31}} + \cdots + \frac{[56 - (75 - \tilde{X}_{11} - \tilde{X}_{12})]^2}{75 - \tilde{X}_{11} - \tilde{X}_{12}} + \cdots \leqq 12.59 . \tag{5.7}$$

At the 95 per cent level of confidence therefore we shall reject any hypothesis specifying values of $\tilde{X}_{11}, \cdots, \tilde{X}_{32}$ for which the expression set out above exceeds 12.59 and will accept all hypotheses for which it has a lower value.[3] One such set is obtained by setting $X_{ij} = \tilde{X}_{ij}$ for $ij \neq 12$ and solving the quartic

$$(15 - \tilde{X}_{12})^2 \left[\frac{1}{\tilde{X}_{12}} + \frac{1}{71 - \tilde{X}_{12}} + \frac{1}{161 - \tilde{X}_{12}} + \frac{1}{139 + \tilde{X}_{12}} \right] = 12.59 \tag{5.8}$$

for X_{12}.

The smallest and largest root of this are $X_{12} = 6.49, 29.2$, leading to corresponding values of z_{12} of 10.2 to 1.33. If we assert that the risk that a smoker will have epidermoid carcinoma of the lung exceeds that for a nonsmoker by at least one-third and by no more than tenfold, the chance that this statement is wrong is less than .05. These limits are of course broad, but we may continue to investigate other relations in table III and enumerate as many sets of X_{ij} satisfying equation (5.7) as are of scientific interest. We have assembled 11 such statements in table IV. The chance that there is any error in these 11 statements is still less than .05. We now see as far as epidermoid carcinoma is concerned we can assert that smokers have a higher risk than nonsmokers (line 1), that cigarette smokers have a higher

[3] This sentence is an almost verbatim quotation from Fisher [11, p. 210]. I am indebted to Fairfield Smith for calling this passage to my attention.

TABLE IV

SIMULTANEOUS 95 PER CENT CONFIDENCE LIMITS ON

	the risk that	will develop	relative to the risk that	will develop	are:
1	any smoker	epidermoid carcinoma	a nonsmoker	epidermoid carcinoma	1.33 to 10.2
2	any smoker	adenocarcinoma	a nonsmoker	adenocarcinoma	0.22 to 7.1
3	any smoker	adenocarcinoma	any smoker	epidermoid carcinoma	0.05 to 2.1
4	a cigar or pipe smoker	epidermoid carcinoma	a nonsmoker	epidermoid carcinoma	0.17 to 2.9
5	a cigarette only smoker	epidermoid carcinoma	a nonsmoker	epidermoid carcinoma	1.64 to 13.0
6	a cigarette only smoker	epidermoid carcinoma	a pipe or cigar smoker	epidermoid carcinoma	2.19 to 18.8
7	a cigarette only smoker	epidermoid carcinoma	a noncigarette smoker	epidermoid carcinoma	2.49 to 11.9
8	a cigarette only smoker	adenocarcinoma	a noncigarette smoker	adenocarcinoma	0.59 to 12.1
9	a cigarette only smoker	adenocarcinoma	a non (cigarette only) smoker	adenocarcinoma	0.79 to 7.2
10	a cigarette plus smoker	epidermoid carcinoma	a cigarette only smoker	epidermoid carcinoma	0.45 to 1.3
11	a cigarette only smoker	epidermoid carcinoma or adenocarcinoma	a noncigarette smoker	epidermoid carcinoma or adenocarcinoma	2.39 to 10.4

risk than pipe and cigar smokers (line 6), that there is no evidence that pipe and cigar smokers have an excess risk, but that if there is one it is less than threefold (line 4), and that the cigarette smoker has a risk at least 2.5 times and perhaps as much as 11.9 times that of the noncigarette smoker (line 7). As far as adenocarcinoma of the lung is concerned, the confidence limits are all too broad to support any kind of useful statement (lines 2, 3, 8, 9, 10).

A numerical investigation of the adequacy of the approximate regions for small samples like that of the preceding section for the 2×2 table would be useful but it is a difficult calculation and we have not been able to undertake it.

6. The $2 \times 2 \times N$ contingency table

Another aspect of this problem is illustrated by the data in table V, which summarizes the findings of 14 retrospective studies. All studies agree in showing a

TABLE V

SUMMARY OF FINDINGS OF 14 RESTROSPECTIVE STUDIES ON THE ASSOCIATION BETWEEN SMOKING AND LUNG CANCER

Study Number	Lung Cancer Patients		Control Patients		Relative Risk
	Total	Nonsmokers	Total	Nonsmokers	
1	86	3	86	14	5.4
2	93	3	270	43	5.7
3	136	7	100	19	4.5
4	82	12	522	125	1.8
5	444	32	430	131	5.6
6	605	8	780	114	13.0
7	93	5	186	12	1.2
8	1357	7	1357	61	9.4
9	63	3	133	27	6.1
10	477	18	615	81	3.8
11	728	4	300	54	36.4
12	518	19	518	56	3.3
13	490	39	2365	636	4.2
14	265	5	287	28	5.5
Total	5437	165	7949	1401	7.5

(SOURCE: Dorn [12])

greater excess risk for smokers, but do not agree as to the magnitude of the difference. While methods exist for deciding whether the differences among the studies are significant, this is not a question of any great interest. Rather we should like an interval estimate of the extent to which they do differ, and a way of combining the results for those which do not appear to differ.

The unconditional probability of N studies each yielding a 2×2 table with fixed marginal totals and X_i nonsmokers is

$$\prod_{i=1}^{N} \binom{n_{i1}}{X_i} \binom{n_{i2}}{m_i - X_i} z_i^{X_i} \tag{6.1}$$

where z_i is the "true" relative risk implied by the definitions and procedures of the

ith study. The conditional probability given the additional restraint

$$\sum_{i=1}^{N} X_i = r \tag{6.2}$$

is

$$K \binom{n_{N1}}{r - \sum_{i=1}^{N-1} X_i} \binom{n_{N2}}{m_N - r + \sum_{i=1}^{N-1} X_i} \prod_{i=1}^{N-1} \binom{n_{i1}}{X_i} \binom{n_{i2}}{m_i - X_i} \theta_i^{X_i} \tag{6.3}$$

where

$$\theta_i = \frac{z_i}{z_N}$$

and

$$\frac{1}{K} = \sum_{X_1} \sum_{X_2} \cdots \sum_{X_{N-1}} \binom{n_{N1}}{r - \sum X_i} \binom{n_{N2}}{m_N - r + \sum X_i} \prod_{i=1}^{N-1} \binom{n_{i1}}{X_i} \binom{n_{i2}}{m_i - X_i} \theta_i^{X_i} .$$

The parameters of the distribution (6.3), the θ_i, are ratios of relative risks, so that we may proceed to the derivation of a large sample confidence region which will permit simultaneous statements about the θ_i. We denote the values of X_i at the point of maximum density of (6.3) by $\tilde{X}_i$, where, in large samples

$$\frac{\tilde{X}_i(n_{i2} - m_i + \tilde{X}_i)(n_{N1} - r + \sum \tilde{X}_i)(m_N - r + \sum \tilde{X}_i)}{(n_{i1} - \tilde{X}_i)(m_i - \tilde{X}_i)(r - \sum \tilde{X}_i)(n_{N2} - m_N + r - \sum \tilde{X}_i)} = \theta_i . \tag{6.4}$$

The large sample confidence region in the $\tilde{X}_i$ is then obtained as before as

$$\frac{\sum_{i=1}^{N} (X_i - \tilde{X}_i)^2}{\tilde{X}_i} = \chi_\alpha^2 [(N - 1)] \tag{6.5}$$

where the X_i are the observed values, the $\tilde{X}_i$ are the variables of the parameter space and the right-hand side is the upper α per cent point of the chi-square distribution with $N - 1$ degrees of freedom.

If we return now to table V we note that the least relative risk shown is 1.2, the largest 36.4, so that the two most extreme studies differ in their estimates by thirtyfold. To compute 95 per cent confidence limits we set $X_i = \tilde{X}_i$ for $i \neq 7$, and set

$$(\tilde{X}_i - 5)^2 \left[\frac{1}{\tilde{X}_i} + \frac{1}{17 - \tilde{X}_i} + \cdots + \frac{1}{719 + \tilde{X}_i} + \frac{1}{251 - \tilde{X}_i} \right] = 22.36 \tag{6.6}$$

for $i = 7$. The smallest root of this octic is $\tilde{X}_7 = 0.85$, so that from (6.4) we estimate the lower 95 per cent confidence limit on θ_7 as 1.71. Although the two studies appear to differ in their estimates of relative risk by thirtyfold the most we can claim at the 95 per cent confidence level is that the procedures and definitions adopted in the two studies differed sufficiently to lead to differences in computed relative risk of at least 70 per cent.

To compare the next two most extreme studies, 4 and 6, we set

$$\tilde{X}_i = X_i \quad \text{for} \quad i \neq 4$$

and

$$(6.7)\quad (\tilde{X}_i - 12)^2 \left[\frac{1}{\tilde{X}_i} + \frac{1}{137 - \tilde{X}_i} + \cdots + \frac{1}{585 + \tilde{X}_i} + \frac{1}{557 - \tilde{X}_i} \right] = 22.36$$

for $i = 4$. The smallest root is $\tilde{X}_4 = 4.05$, so that from (6.4) we compute the lower 95 per cent confidence limit on θ_4 as 1.08. Thus, although studies 4 and 6 appear to differ by sevenfold in their estimate of relative risk the most we can assert is that they differ by more than 8 per cent. For all remaining comparisons confidence limits on the θ_i include unity, so that as far as the evidence of table V goes, 10 of the 14 retrospective studies could be supplying the same estimates of relative risk, even though the lowest and highest differ by threefold.

If now we combine these 10 studies we obtain 136 nonsmokers among the 3,929 lung cancer patients and 1,096 nonsmokers among the 6,161 control patients. The 95 per cent interval estimate of relative risk can now be obtained from equations (3.9) and (3.10). This calculation gives

$$(6.8)\quad \begin{aligned} \tilde{X}_2 &= 158.3 \\ \tilde{X}_1 &= 116.4 \end{aligned}$$

so that

$$(6.9)\quad \begin{aligned} \frac{1}{z_2'} &= 5.03 \\ \frac{1}{z_1'} &= 7.24\ . \end{aligned}$$

We thus on the basis of table V make the composite assertion that (a) studies 7 and 11 are not samples from the same populations, (b) studies 4 and 6 are not samples from the same population, (c) all the remaining studies could be samples from the same population, and (d) these remaining studies indicate a risk of having lung cancer for smokers relative to nonsmokers of not less than 5.0 and not more than 7.2. The chance that this composite assertion is wrong in any respect is at most $.95 \times .95$.

We remark parenthetically that if the published data for all ten studies had permitted that estimation of relative risk for cigarette smokers, and particularly heavy cigarette smokers, these risks as calculated would be considerably larger.

REFERENCES

[1] J. CORNFIELD, "A method of estimating comparative rates from clinical data, applications to cancer of the lung, breast and cervix," *Jour. Nat. Cancer Inst.*, Vol. 11 (1951), pp. 1269–1275.

[2] P. B. PATNAIK, "The power function of the test for the difference between two proportions in a 2 × 2 table," *Biometrika*, Vol. 35 (1948), pp. 157–173.

[3] W. L. STEVENS, "Mean and variance of an entry in a contingency table," *Biometrika*, Vol. 38 (1951), pp. 468–470.
[4] A. M. MOOD, *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950.
[5] R. A. FISHER, "The logic of inductive inference," *Jour. Roy. Stat. Soc.*, Vol. 98, Part I (1935), pp. 39–54.
[6] T. C. FRY, *Probability and its Engineering Uses*, New York, D. Van Nostrand Company, 1928.
[7] E. B. WILSON, "Probable inference, the law of succession and statistical inference," *Jour. Amer. Stat. Assoc.*, Vol. 22 (1927), pp. 209–212.
[8] E. L. WYNDER and J. CORNFIELD, "Cancer of the lung in physicians," *New England Jour. of Medicine*, Vol. 248 (1953), pp. 441–444.
[9] L. BRESLOW, L. HOAGLIN, G. RASMUSSEN, and H. K. ABRAMS, "Occupations and cigarette smoking as factors in lung cancer," *Amer. Jour. Public Health*, Vol. 44 (1954), pp. 171–181.
[10] H. SCHEFFÉ, "A method for judging all contrasts in the analysis of variance," *Biometrika*, Vol. 40 (1953), pp. 87–104.
[11] R. A. FISHER, *The Design of Experiments*, Edinburgh, Oliver and Boyd, 1935.
[12] H. F. DORN, "The relationship of cancer of the lung and the use of tobacco," *Amer. Statistician*, Vol. 8 (1954), pp. 7–13.