19 Individually matched case-control studies

Analyses which preserve the matching of individual cases to their controls follow similar principles to those of Chapter 18. The strata are now the sets made up of each case and its matched controls. Studies designed to have a fixed number of controls, m say, drawn for each case, will be referred to as 1:m matched studies.

19.1 Mantel–Haenszel analysis of the 1:1 matched study

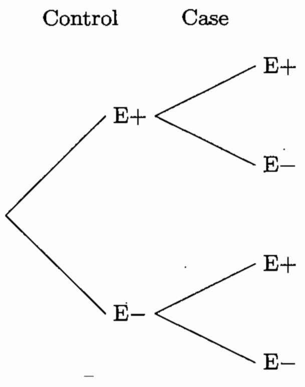
For reasons discussed in Chapter 18, the use of profile likelihood gives misleading estimates of odds ratios when there are a large number of strata with little data in each stratum. However, the Mantel–Haenszel method works perfectly well in these circumstances. The calculations are particularly easy in the 1:1 case, and illustrate ideas which are important for our later discussion of the likelihood approach.

The results of 1:1 matched studies are usually presented in 2×2 tables such as Table 19.1.* These data were drawn from the same study as reported in Chapter 17, and concern the relationship between tonsillectomy history and the incidence of Hodgkin's disease. The total study included 174 cases and 472 controls, but the controls were siblings of the cases, and the authors felt that the matching of cases and sibling controls should be preserved. They also wished to control for age and sex and therefore restricted their analysis to 85 matched case-control pairs in which the case and sibling control were of the same sex and matched for age within a specified margin. Note that, in the construction of matched sets, the original 174 cases and 472 controls have been reduced to only 85 cases and 85 controls.

Tables such as Table 19.1 can be confusing because we are used to seeing tables that count subjects, while this table counts case-control sets. The four cells of the table correspond to the four possible exposure configurations of a case-control set. These are illustrated in terms of a tree in Fig. 19.1. The first branching point is according to whether or not the control was exposed (denoted E+ and E- respectively), while the second branching is according to exposure of the case. The frequencies in Table 19.1 refer to counts of these four configurations.

Table 19.1. Tonsillectomy history in 85 matched pairs

History	History of control	
of case	Positive	Negative
Positive	26	15
Negative	7	37



Control	Case	H_1	H_0	D_1	D_0
E+	E+	1	0	1	0
	E−	1	0	0	1
E−	E+	0	1	1	0
	E−	0	1	0	1

Fig. 19.1. Exposure configurations for 1:1 sets.

Exercise 19.1. How often did each of the exposure configurations of Fig. 19.1 occur?

In the analysis of individually matched studies the strata are case-control sets so that, in the notation of Chapter 18, t indexes sets. The number of subjects in each stratum is $N^t = 2$, and since each stratum contains one case and one control, D^t and H^t are always 1. The values of D_1^t, D_0^t, H_1^t, and H_0^t for each exposure configuration are shown in Fig. 19.1. In this figure and henceforth we will omit the superscript t for clarity, and remember that the symbols refer to values in a single case-control set.

Exercise 19.2. What are the contributions of each configuration to Q and R in the Mantel–Haenszel estimate of the odds ratio? Similarly what are the contributions to the score and score variance, U and V? Which configurations contribute to estimation and testing?

It can be seen that only two exposure configurations make any contribution to estimation and testing of the odds ratio. These are the sets in which the exposure status of case and controls differ and are called *discordant* sets. The remaining sets are called *concordant sets.* In our current example, 63

*From Cole, P. *et al.* (1973) *New England Journal of Medicine*, **288**, 634.

of the case-control sets are concordant and are ignored.

Exercise 19.3. For the tonsillectomy data, what are the values for Q, R, U, V? Using the methods of Chapter 18, estimate the odds ratio, its 90% confidence interval, and a p-value for $\theta = 1$.

The odds ratio estimate is very close to that obtained in the analysis of Chapter 17, but so much data has been lost in this analysis that the result is no longer statistically significant. It is easy to criticize an analysis which discards so much data, but when it is necessary to preserve the matching of controls to cases it is not easy to see how one can adjust for the effects of additional variables by stratification, since the case and its control may fall within different strata. At the time this study was reported there would have been no alternative but to discard such sets. Nowadays, this problem is easily overcome by use of the regression methods to be described in Part II.

Before leaving this example, it is interesting to note that the above analysis is not the one originally reported. In their first report, the researchers subscribed to the misconception discussed in Chapter 18 — that the matching for age, sex, and family was sufficient to control for these variables and that subsequently the matching could be ignored in the analysis.

Exercise 19.4. Show that the odds ratio estimate obtained by ignoring the matching is less than that obtained by the correct analysis.

19.2 The hypergeometric likelihood for 1:1 matched studies

The hypergeometric likelihood is obtained by arguing conditionally upon *both* margins of the 2 × 2 table, and depends only upon the odds ratio parameter. It is usually difficult to compute, but its use is only necessary when the data within strata are few. This is the case for individually matched studies and the hypergeometric likelihood *must* be used. Luckily in this case the computations are quite easy — particularly in the 1:1 case.

Fig. 19.2 derives the probability of each exposure configuration by multiplying along branches of the tree in the usual way and also lists the total number of *subjects* in the set who were exposed, N_1. The odds that the control in the set was exposed is denoted by Ω_0 and the odds that the case was exposed by Ω_1, and we have written K for the expression

$$\frac{1}{(1+\Omega_0)(1+\Omega_1)}$$

which occurs in all four probabilities. To obtain the hypergeometric likelihood we argue conditionally on the number of subjects exposed, N_1. It is clear from the figure that, when $N_1 = 2$, there is only one possible exposure configuration; the *conditional* probability of the observation is 1 and there is no contribution to the log likelihood. Similarly, there is no

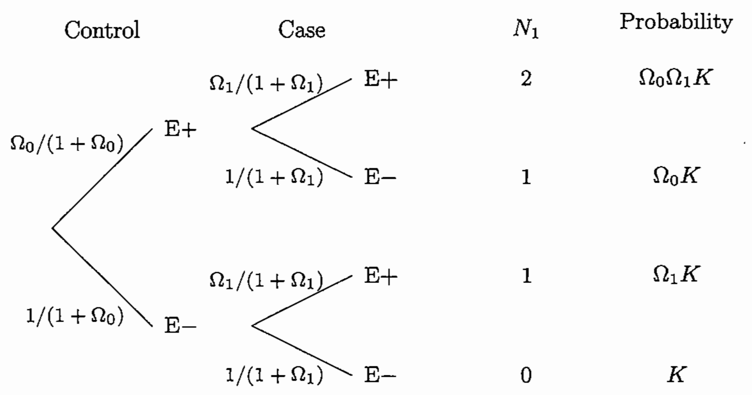


Fig. 19.2. Probabilities for a case-control set.

contribution to the log likelihood from sets in which $N_1 = 0$. These configurations correspond to the concordant sets which were also ignored in our previous analysis. However, when $N_1 = 1$ the exposure configuration could be either the second or third. These are the possible configurations of discordant sets. The observed split of discordant sets between the second and third configurations determines the log likelihood.

The conditional probabilities that a discordant set is of the third type (case exposed, control unexposed) and the second type (case unexposed, control exposed) are

$$\frac{\Omega_1 K}{\Omega_0 K + \Omega_1 K} \quad \text{and} \quad \frac{\Omega_0 K}{\Omega_0 K + \Omega_1 K}$$

respectively, and the conditional odds that the case was exposed is the ratio of these, Ω_1/Ω_0. This is the odds ratio parameter θ, assumed in our model to be constant for all the case control sets. The conditional argument therefore leads to a Bernoulli log likelihood based on splits of discordant sets into those in which the case is exposed and those in which the case is unexposed, the odds for such splits being θ. In our data, such sets split 15:7 and the log likelihood is

$$15\log(\theta) - 22\log(1+\theta).$$

Exercise 19.5. Calculate the most likely value of θ, a 90% confidence interval and the score test for the null hypothesis $\theta = 1$. These results of this exercise should agree precisely with those obtained using the Mantel–Haenszel method.

Table 19.2. Screening history in breast cancer deaths and matched controls

Status of the case	Number of controls screened			
	0	1	2	3
Screened	1	4	3	1
Unscreened	11	10	12	4

19.3 Several controls per case

The arguments outlined above may be extended to the situation in which there are several controls for each case. As before, we start with the Mantel–Haenszel approach.

Table 19.2 shows the results of a case-control study of breast cancer screening. Cases are deaths from breast cancer and each case is matched with three control women.[†] The exposure of interest is attendance for breast cancer screening. If screening is effective in prolonging life, screened women should have lower mortality rates and the odds ratio estimate from the case-control study should be less than 1. Note that as in Table 19.1, the table counts case-control sets and not women.

This study illustrates one of the reasons for matching discussed in Chapter 18. Women who die from breast cancer usually do so some years after initial diagnosis and during the period between diagnosis and death they would not be screened. Thus, controls would have a greater opportunity to be screened than cases. This difficulty was overcome by determining the relevant *exposure window*; the screening history of the controls was assessed over the period up to the time of diagnosis of the case, so that the screening histories of cases and controls are comparable. It was only possible to deal with this problem in this way because the study matched controls to individual cases.

Table 19.2 demonstrates the usual way such data are presented. However, it is very difficult to perceive any pattern — even as to whether or not screening appears to be a protective. To understand the analysis, we shall start by reordering the data as a tree. Fig. 19.3 illustrates the possible exposure configurations. The first three branches represent the exposure status of the three controls, the upper branch representing exposed (E+) and the lower unexposed (E−). Because we do not wish to differentiate between individual controls, this section of the tree may be abbreviated. For the first two controls, we do not need to differentiate between the configurations (E+, E−) and (E−, E+). These are simply grouped together as having 1 control exposed and we write the figure 2 at this point to remind us that branches emanating from this point are *double* branches. Similarly, after consideration of the third control we group together the 3 configurations with 2 exposed controls and the 3 configurations with 1 exposed control. The final branching represents the exposure status of the case.

[†] From Collette, H.J.A. *et al.* (1984) *The Lancet*, June 2, 1984, 1224–1226.

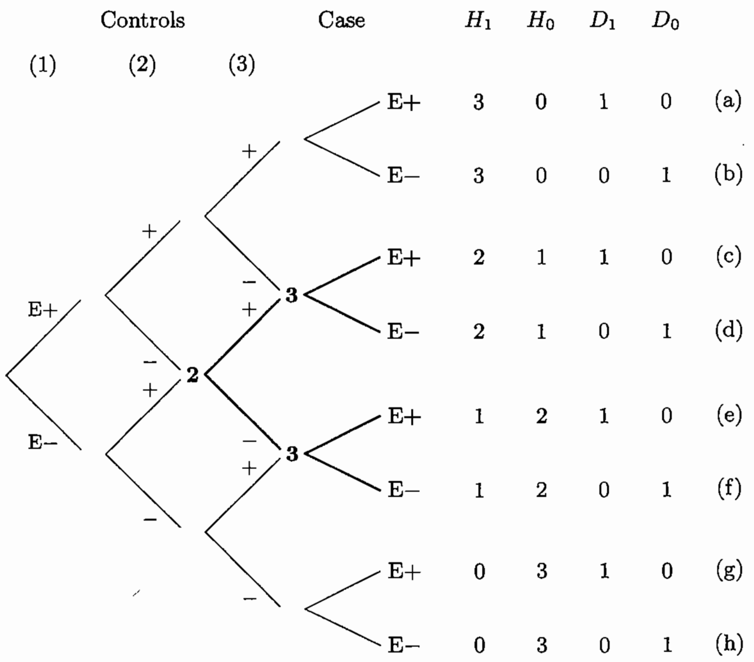


Fig. 19.3. Exposure configurations for 1:3 sets.

Exercise 19.6. In the screening data, how frequently do each of the eight types of exposure configuration occur?

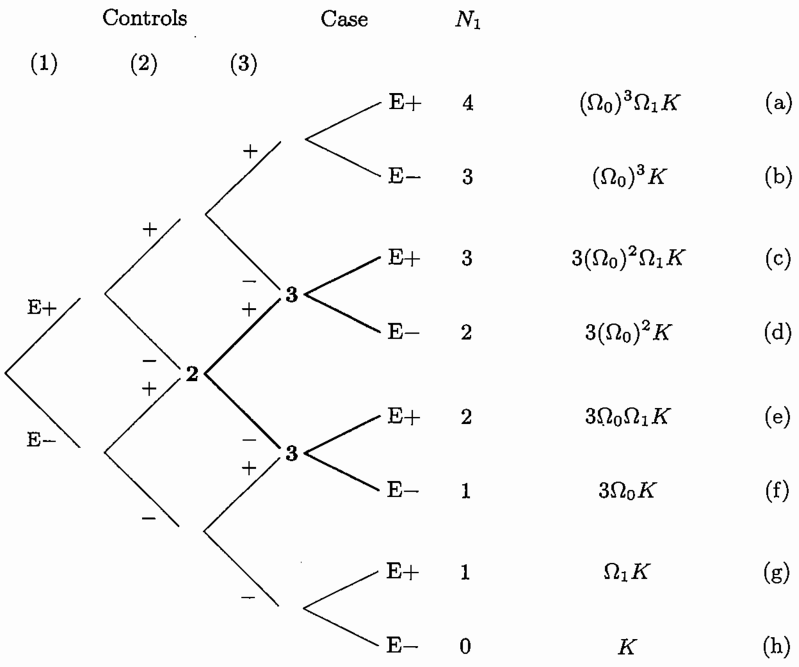
We shall first analyse these data by the Mantel–Haenszel method. In the next section, we shall discuss the likelihood approach and show how it suggests a more useful arrangement of the table.

Exercise 19.7. Tabulate the values of Q, R, U, and V for these eight tables and hence calculate the Mantel–Haenszel significance test, odds ratio estimate and an approximate 90% confidence interval.

This analysis shows that the study finds a substantial and statistically significant reduction in mortality as a result of breast cancer screening.

19.4 The likelihood ⋆

The analysis of these data by use of the hypergeometric likelihood method is also quite straightforward. As before we argue conditionally upon the margins. Fig. 19.4 shows the total number of *subjects* exposed, N_1, and the



Controls
Case
N_1
(1)
(2)
(3)
E+
E−
+
−
2
3
E+ 4 $(\Omega_0)^3\Omega_1 K$ (a)
E− 3 $(\Omega_0)^3 K$ (b)
E+ 3 $3(\Omega_0)^2\Omega_1 K$ (c)
E− 2 $3(\Omega_0)^2 K$ (d)
E+ 2 $3\Omega_0\Omega_1 K$ (e)
E− 1 $3\Omega_0 K$ (f)
E+ 1 $\Omega_1 K$ (g)
E− 0 K (h)

Fig. 19.4. Probabilities for 1:3 sets.

probability of each configuration, again writing K for the common factor, in this case

$$K = \frac{1}{(1+\Omega_0)^3(1+\Omega_1)}.$$

Note that the probabilities for configurations (c) to (f) are multiplied by 3 because each of these represents three paths in the complete tree. Now there are 5 possible values for the total number of subjects exposed. Again there are two *concordant* configurations in which the number of subjects exposed uniquely determines the configuration. $N_1 = 4$ ensures configuration (a) and $N_1 = 0$ ensures configuration (h). These make no contribution to the log likelihood. Each of the other three values of N_1 allows for two possible configurations, one in which the case is exposed and the other in which the case is unexposed. It is the splits of the observed data between these that yield the likelihood.

If the total number of exposed subjects in the set, N_1, is fixed at 3, then the exposure configuration must be either (b) or (c) and the conditional

Table 19.3. Splits of case-control sets

N_1	Split	Odds	Observed
3	(c):(b)	3θ	3:4
2	(e):(d)	θ	4:12
1	(g):(f)	$\theta/3$	1:10

odds for the split (c):(b) is

$$\frac{3(\Omega_0)^2\Omega_1 K}{(\Omega_0)^3 K} = \frac{3\Omega_1}{\Omega_0} = 3\theta.$$

Similarly, $N_1 = 2$ implies (d) or (e) and $N_1 = 0$ implies (f) or (g). The odds predicted by the model for these splits are set out in Table 19.3, together with the observed frequencies. By eye we can see that a value of θ of about 0.3 predicts the observed splits very well indeed. More formally, the log likelihood is

$$\begin{aligned} & 1\log\left(\frac{\theta}{3}\right) - 11\log\left(1+\frac{\theta}{3}\right) \\ + \quad & 4\log(\theta) - 16\log(1+\theta) \\ + \quad & 3\log(3\theta) - 7\log(1+3\theta). \end{aligned}$$

There is no simple expression for the maximum likelihood estimate and it is necessary to use a computer program to search for the maximum. This occurs at $\theta = 0.31$ ($\log(\theta) = -1.18$). The plot of the log likelihood ratio against $\log(\theta)$ is shown in Fig. 19.5. A Gaussian approximation with $S = 0.404$ fits quite closely.

The generalization of this argument to any number of controls per case may be carried out algebraically or by extending our tree. For sets of N_1 exposed subjects and N_0 unexposed subjects, the constant odds ratio model predicts that sets will split between those with an exposed case and those with an unexposed case with odds

$$N_1\theta/N_0.$$

A similar generalization is possible for several *cases* in each set. We will not give the details here, but computer software is readily available. Such analyses do not arise frequently in practice. An exception is family studies in which more than one sibling may be affected by a disease and unaffected siblings are used as controls.

In the examples discussed in this chapter, the Mantel–Haenszel and likelihood methods agree closely. The calculations for the former are rather easier, but the advantage of the likelihood approach lies in its greater generality and possibilities for extension. For example, when there are more

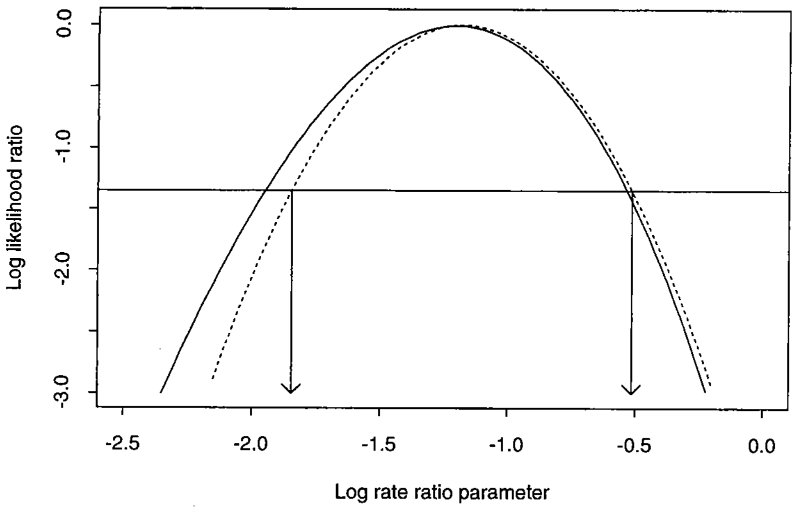


Fig. 19.5. Log likelihood ratio for $\log(\theta)$.

than two exposure categories, there is no simple method analogous to the Mantel–Haenszel approach. We shall defer discussion of such extensions to Part II of the book.

Solutions to the exercises

19.1 In the order in which the exposure configurations are listed in the figure, their frequencies are 26, 7, 15, and 37.

19.2 In the same order as listed,

Q	R	U	V
0	0	0	0
0	1/2	-1/2	1/4
1/2	0	1/2	1/4
0	0	0	0

Only the second and third configurations contribute to Q, R, U, and V.

19.3

$$\begin{aligned} Q &= 15 \times (1/2) \\ R &= 7 \times (1/2) \\ U &= 15 \times (1/2) - 7 \times (1/2) = 4 \end{aligned}$$

$$V = 15 \times (1/4) + 7 \times (1/4) = 5.5$$

The odds ratio estimate is $15/7 = 2.14$. This estimates the underlying rate ratio, so that the suggestion is that tonsillectomy doubles the rate of Hodgkin's disease. Using the expression

$$S = \sqrt{\frac{V}{QR}} = 0.4577,$$

the 90% error factor for the odds ratio is $\exp(1.645 \times 0.4577) = 2.12$. The 90% confidence limits are, therefore, $2.14/2.12 = 1.01$ (lower limit) and $2.14 \times 2.12 = 4.54$ (upper limit). Referring the value $(U)^2/V = 2.91$ to the chi-squared distribution gives $p \approx 0.09$.

19.4 If the matching is ignored, the following 2×2 table is obtained:

History:	Positive	Negative
Cases	41	44
Controls	33	52

The odds ratio in this table is $(41 \times 52)/(33 \times 44) = 1.47$, as compared to the value of 2.14 obtained by the correct analysis.

19.5 The most likely value is $15/7 = 2.14$. To calculate the approximate 90% interval using Gaussian approximation of the log likelihood for $\log(\theta)$ we use

$$S = \sqrt{\frac{1}{15} + \frac{1}{7}} = 0.4577,$$

the same as we obtained with the Mantel–Haenszel method. Under the null hypothesis, the probability for the split is 0.5 so that the expected number of sets with an exposed case is $22 \times 0.5 = 11$. The score and score variance are

$$\begin{aligned} U &= 15 - 11 = 4, \\ V &= 22 \times 0.5 \times 0.5 = 5.5. \end{aligned}$$

Again these are the values we obtained using the Mantel–Haenszel method.

19.6 In the order listed in the figure, the 8 exposure configurations have frequencies 1, 4, 3, 12, 4, 10, 1, 11.

19.7 The contributions to Q, R, U and V are shown below:

	Number				
	of sets	Q	R	U	V
(a)	1	0	0	0	0
(b)	4	0	3/4	−3/4	9/48
(c)	3	1/4	0	1/4	9/48
(d)	12	0	2/4	−2/4	12/48
(e)	4	2/4	0	2/4	12/48
(f)	10	0	1/4	−1/4	9/48
(g)	1	3/4	0	3/4	9/48
(h)	11	0	0	0	0
Total		14/4	46/4	-32/4	354/48

Note that each contribution has to be multiplied by the number of times it occurred so that, for example, the total value of Q is

$$(3 \times 1/4) + (4 \times 2/4) + (1 \times 3/4) = 14/4.$$

The Mantel–Haenszel estimate of θ is $14/46 = 0.30$ and the chi-squared test is $(U)^2/V = 8.68$ ($p < 0.01$). An approximate error factor can be calculated from

$$\exp\left(1.645 \times \sqrt{\frac{V}{QR}}\right) = 2.02$$

so that the 90% confidence interval lies from $\theta = 0.15$ to $\theta = 0.60$.

0 Preamble to Ch 18 & 19: The non-central hypergeometric distribution:

The **null (central)** hypergeometric distribution arises when

i. making inference about N_1 and N_0 from a sample of size $n < N$ from a finite population of (say) $N_1 + N_0 = N$ elements, with N_1 of them having the values $Y = 1$ and N_0 having $Y = 0$. The resulting random variable, $\sum_1^n y_i$, is the number, out of the n sampled, in which the sampled Y takes the value 1. Its minimum value is the larger of 0 and $N_1 - (N - n)$, and its maximum value is the smaller of n and N_1.

   Examples include sampling from populations such as elected politicians, or university presidents, or G20 leaders, and the urn sampling used in lotteries (e.g. 6/49) and casinos (e.g., Keno).

ii. *testing* for equality of 2 binomial parameters π_1 and π_0 using independent samples of sizes n_1 and n_0, but conditioning on the overall numbers of 'positives' (m_1) and 'negatives' (m_0) in the combined samples, i.e., on all 4 margins of the 2×2 table that cross-classifies the sampled elements by their Y value and whether they arose in the reference or index categories (0 and 1) of the contrast of interest.

   With this conditioning, introduced by Fisher, the parameter space is reduced from 2 to 1 dimension, the sample space from $(n_1+1) \times (n_0+1)$ points in the 2-D grid, to $\min\{n_1, n_0, m_1, m_1\} + 1$ points along a single diagonal, and the test statistic from 2-dimensional to 1-dimensional.

   Moreover, it is the same test, no matter whether the parameter of interest is the simple difference, $\pi_1 - \pi_0$; their ratio, $\pi_1 \div \pi_0$; or the ratio of the corresponding odds, $\frac{\pi_1}{1-\pi_1} \div \frac{\pi_0}{1-\pi_0}$. The same holds true for unconditional tests, provided one is consistent about dis-continuity corrections, etc.

   The 5 tables from Fisher's **famous tea-tasting experiment**[1] with the 2×2 tables with all marginal totals = 4 are another example of this hypergeometric distribution.

The unity in (ii) is lost when we move to interval estimation, with separate approaches for the different comparative parameters. Since case-base series ('case-control' studies) lead to a Rate Ratio estimator that is a numerical cross-product (i.e., the statistic *looks like* an empirical odds ratio) that can be seen as arising from 2 independent binomials with different parameter values, we will focus for now on the **odds-ratio parameter**. If we use the same conditioning as in (ii) above, and keep our focus on the single parameter $\frac{\pi_1}{1-\pi_1} \div \frac{\pi_0}{1-\pi_0}$, we arive at the **non-central hypergeometric distribution**.

We will use our own notation, but Fisher's example 1, next, to introduce it. It is also described in section 4.2 'Exact statistical inference for a single 2×2 table' in Chapter 4 of Volume I of Breslow and Day.

<u>**SETUP**</u>: Let $Y_i \sim \text{Binomial}(n_i, \pi_i)$, $i = 0, 1$, be 2 independent binomial random variables. We wish to make inference regarding the parameter

$$\psi = \{\pi_1/(1-\pi_1\}/\{\pi_0/(1-\pi_0\}.$$

We can do so by considering only those data configurations which have the same total number of 'positives', $y_1 + y_0 = y$, say, as were observed in the actual study, and then considering the distribution of $Y_1 \mid y$.

$$Prob[Y_1 = y_1 \,;\, Y_0 = y_0] = \,^{n_1}\mathrm{C}_{y_1}\, \pi_1^{y_1}(1-\pi_1)^{n_1-y_1} \times \,^{n_0}\mathrm{C}_{y_0}\, \pi_0^{y_0}(1-\pi_0)^{n_0-y_0}.$$

If we condition on $Y_1 + Y_0 = y$, then

$$Prob[Y_1 = y_1 \mid Y_1 + Y_0 = y] = Prob[Y_1 = y_1 \,;\, Y_0 = y - y_1]/Prob[Y_1 + Y_1 = y].$$

If we rewrite the quantity

$$\pi_1^{y_1}(1-\pi_1)^{n_1-y_1} \times \pi_0^{y_0}(1-\pi_0)^{n_0-y_0}$$

as

$$\pi_1^{y_1}(1-\pi_1)^{-y_1}\pi_0^{-y_0}(1-\pi_0)^{y_1} \times (1-\pi_1)^{n_1}\pi_0^{y}(1-\pi_0)^{n-y}$$

we see that it simplifies to

$$\psi^{y_1} \times (1-\pi_1)^{n_1}\, \pi_0^{y}\, (1-\pi_0)^{n-y}$$

and that the last three terms do not involve ψ and do not involve the random variable y_1. Since they appear in both the numerator and the denominator of the conditional probability, they cancel out.

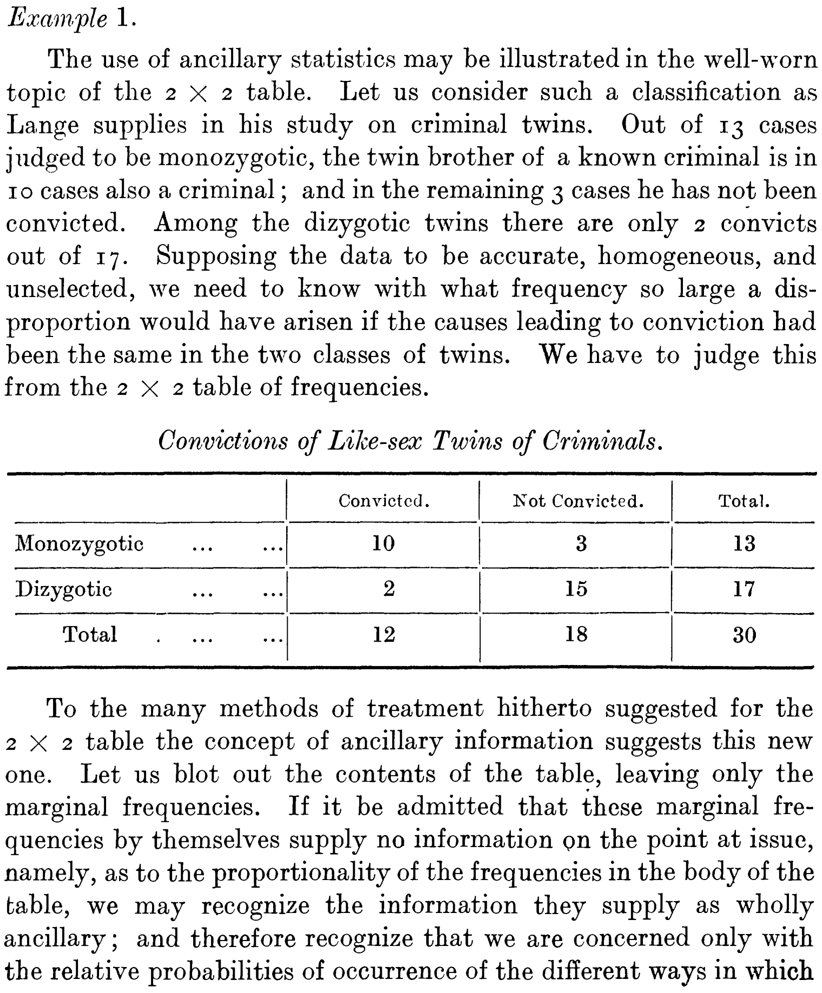
Thus we can write the conditional probability $Prob[Y_1 = y_1 \mid Y_1 + Y_0 = y]$ as

$$Prob[\, y_1 \mid y\,] = \,^{n_1}\mathrm{C}_{y_1}\, ^{n_0}\mathrm{C}_{y-y_1}\, \psi^{\,y_1} \,/\, \Sigma\, ^{n_1}\mathrm{C}_{y_1'}\, ^{n_0}\mathrm{C}_{n-y_1'}\, \psi^{\,y_1'},$$

where the summation is over those y_1' values that are compatible with the 4 marginal frequencies.

[1] See the slides 'What the P-Value IS and IS NOT' in JH's material for the Bionano Workshop.

Aside: note that if we set $\psi = 1$, the probabilities are the same as those in the central hypergeometric distribution, used for Fisher's exact test of two binomial proportions, Indeed, Fisher, in page 48-49, first computes the null probabilities for the 2 × 2 table. The combinatorials are only computed once.



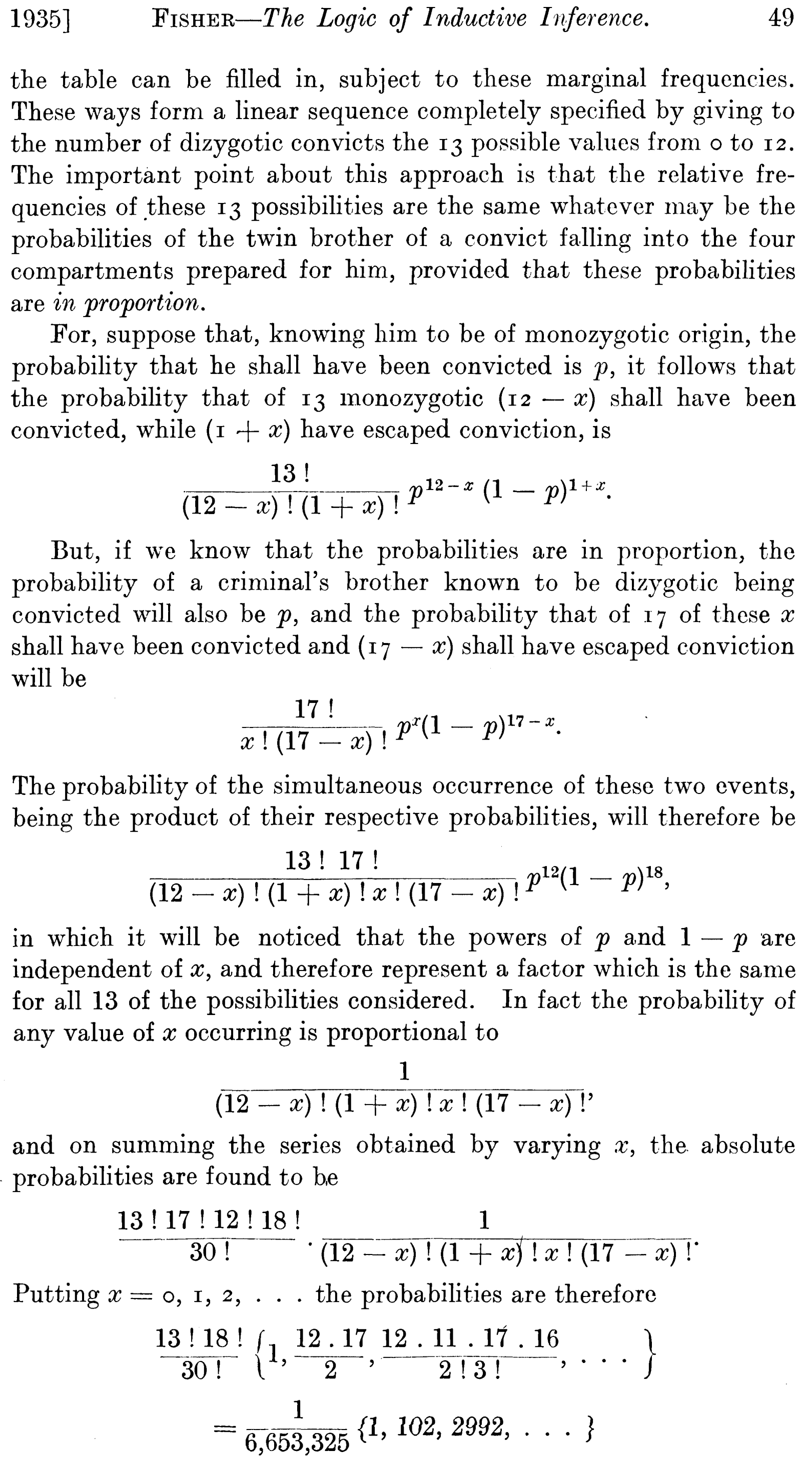
Example 1.

The use of ancillary statistics may be illustrated in the well-worn topic of the 2 × 2 table. Let us consider such a classification as Lange supplies in his study on criminal twins. Out of 13 cases judged to be monozygotic, the twin brother of a known criminal is in 10 cases also a criminal; and in the remaining 3 cases he has not been convicted. Among the dizygotic twins there are only 2 convicts out of 17. Supposing the data to be accurate, homogeneous, and unselected, we need to know with what frequency so large a disproportion would have arisen if the causes leading to conviction had been the same in the two classes of twins. We have to judge this from the 2 × 2 table of frequencies.

Convictions of Like-sex Twins of Criminals.

	Convicted.	Not Convicted.	Total.
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

To the many methods of treatment hitherto suggested for the 2 × 2 table the concept of ancillary information suggests this new one. Let us blot out the contents of the table, leaving only the marginal frequencies. If it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely, as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which



1935] FISHER—*The Logic of Inductive Inference.* 49

the table can be filled in, subject to these marginal frequencies. These ways form a linear sequence completely specified by giving to the number of dizygotic convicts the 13 possible values from 0 to 12. The important point about this approach is that the relative frequencies of these 13 possibilities are the same whatever may be the probabilities of the twin brother of a convict falling into the four compartments prepared for him, provided that these probabilities are *in proportion*.

For, suppose that, knowing him to be of monozygotic origin, the probability that he shall have been convicted is p, it follows that the probability that of 13 monozygotic $(12 - x)$ shall have been convicted, while $(1 + x)$ have escaped conviction, is

$$\frac{13\,!}{(12 - x)\,!\,(1 + x)\,!}\,p^{12-x}\,(1 - p)^{1+x}.$$

But, if we know that the probabilities are in proportion, the probability of a criminal's brother known to be dizygotic being convicted will also be p, and the probability that of 17 of these x shall have been convicted and $(17 - x)$ shall have escaped conviction will be

$$\frac{17\,!}{x\,!\,(17 - x)\,!}\,p^{x}(1 - p)^{17-x}.$$

The probability of the simultaneous occurrence of these two events, being the product of their respective probabilities, will therefore be

$$\frac{13\,!\;17\,!}{(12 - x)\,!\,(1 + x)\,!\,x\,!\,(17 - x)\,!}\,p^{12}(1 - p)^{18},$$

in which it will be noticed that the powers of p and $1 - p$ are independent of x, and therefore represent a factor which is the same for all 13 of the possibilities considered. In fact the probability of any value of x occurring is proportional to

$$\frac{1}{(12 - x)\,!\,(1 + x)\,!\,x\,!\,(17 - x)\,!},$$

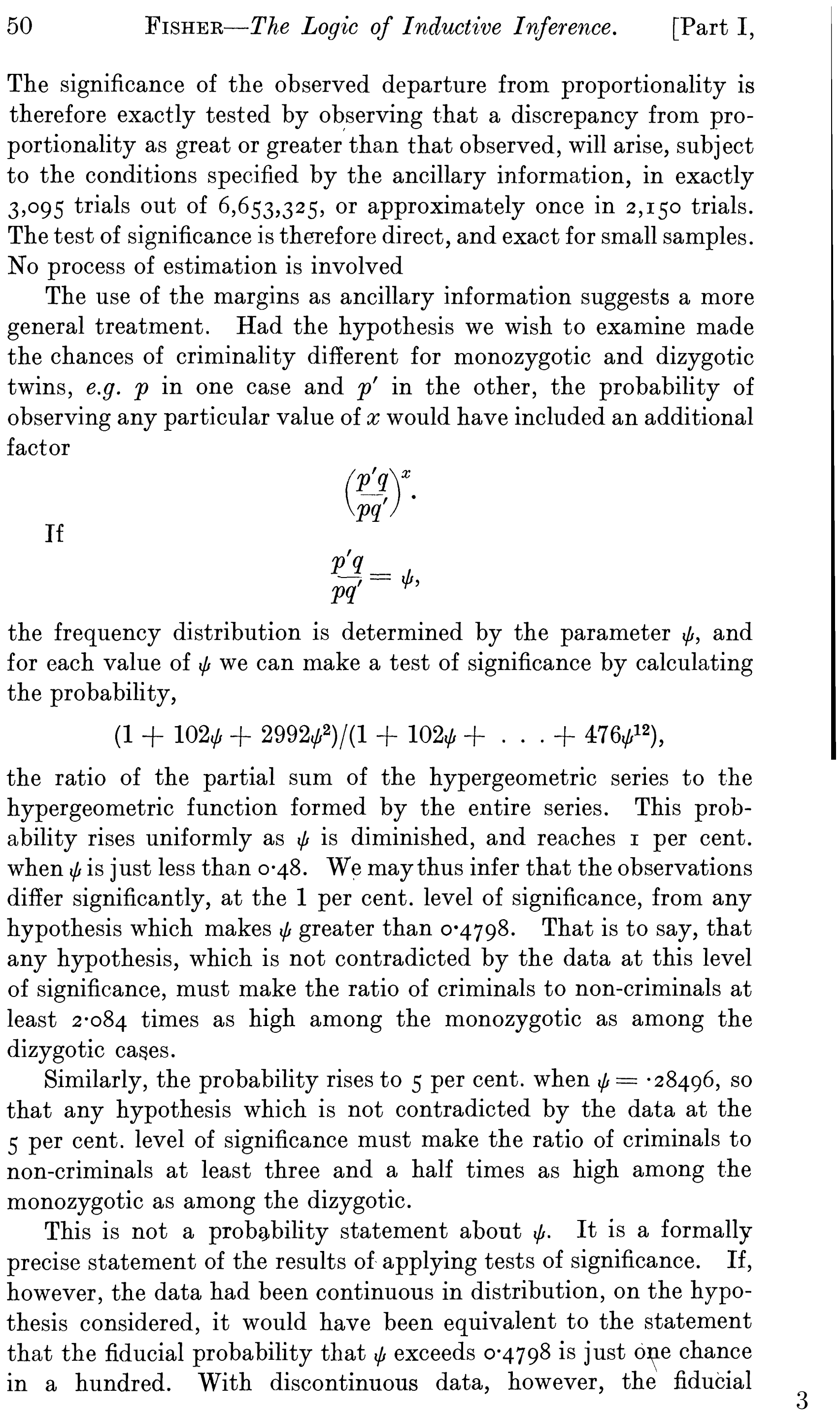
and on summing the series obtained by varying x, the absolute probabilities are found to be

$$\frac{13\,!\,17\,!\,12\,!\,18\,!}{30\,!}\cdot\frac{1}{(12 - x)\,!\,(1 + x)\,!\,x\,!\,(17 - x)\,!}.$$

Putting $x = 0, 1, 2, \ldots$ the probabilities are therefore

$$\frac{13\,!\,18\,!}{30\,!}\left\{1, \frac{12\,.\,17}{2}, \frac{12\,.\,11\,.\,17\,.\,16}{2\,!\,3\,!}, \ldots\right\}$$

$$= \frac{1}{6{,}653{,}325}\{1, 102, 2992, \ldots\}$$



The significance of the observed departure from proportionality is therefore exactly tested by observing that a discrepancy from proportionality as great or greater than that observed, will arise, subject to the conditions specified by the ancillary information, in exactly 3,095 trials out of 6,653,325, or approximately once in 2,150 trials. The test of significance is therefore direct, and exact for small samples. No process of estimation is involved

The use of the margins as ancillary information suggests a more general treatment. Had the hypothesis we wish to examine made the chances of criminality different for monozygotic and dizygotic twins, *e.g.* p in one case and p' in the other, the probability of observing any particular value of x would have included an additional factor

$$\left(\frac{p'q}{pq'}\right)^x.$$

If

$$\frac{p'q}{pq'} = \psi,$$

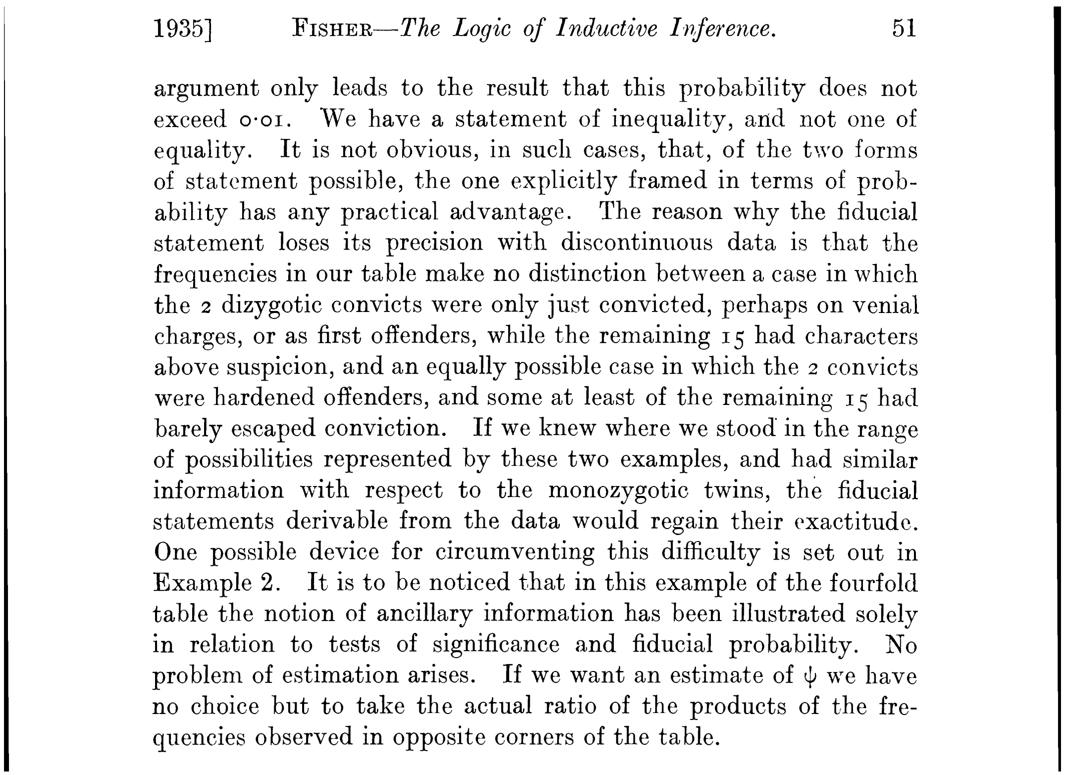
the frequency distribution is determined by the parameter ψ, and for each value of ψ we can make a test of significance by calculating the probability,

$$(1 + 102\psi + 2992\psi^2)/(1 + 102\psi + \ . \ . \ . + 476\psi^{12}),$$

the ratio of the partial sum of the hypergeometric series to the hypergeometric function formed by the entire series. This probability rises uniformly as ψ is diminished, and reaches 1 per cent. when ψ is just less than 0·48. We may thus infer that the observations differ significantly, at the 1 per cent. level of significance, from any hypothesis which makes ψ greater than 0·4798. That is to say, that any hypothesis, which is not contradicted by the data at this level of significance, must make the ratio of criminals to non-criminals at least 2·084 times as high among the monozygotic as among the dizygotic cases.

Similarly, the probability rises to 5 per cent. when $\psi = {\cdot}28496$, so that any hypothesis which is not contradicted by the data at the 5 per cent. level of significance must make the ratio of criminals to non-criminals at least three and a half times as high among the monozygotic as among the dizygotic.

This is not a probability statement about ψ. It is a formally precise statement of the results of applying tests of significance. If, however, the data had been continuous in distribution, on the hypothesis considered, it would have been equivalent to the statement that the fiducial probability that ψ exceeds 0·4798 is just one chance in a hundred. With discontinuous data, however, the fiducial



argument only leads to the result that this probability does not exceed 0·01. We have a statement of inequality, and not one of equality. It is not obvious, in such cases, that, of the two forms of statement possible, the one explicitly framed in terms of probability has any practical advantage. The reason why the fiducial statement loses its precision with discontinuous data is that the frequencies in our table make no distinction between a case in which the 2 dizygotic convicts were only just convicted, perhaps on venial charges, or as first offenders, while the remaining 15 had characters above suspicion, and an equally possible case in which the 2 convicts were hardened offenders, and some at least of the remaining 15 had barely escaped conviction. If we knew where we stood in the range of possibilities represented by these two examples, and had similar information with respect to the monozygotic twins, the fiducial statements derivable from the data would regain their exactitude. One possible device for circumventing this difficulty is set out in Example 2. It is to be noticed that in this example of the fourfold table the notion of ancillary information has been illustrated solely in relation to tests of significance and fiducial probability. No problem of estimation arises. If we want an estimate of ψ we have no choice but to take the actual ratio of the products of the frequencies observed in opposite corners of the table.

Fisher calculated that the probability that 1, 2, 3, . . . monozygotic twins would escape conviction[2] was $(1/6\ 652\ 325)\times\{1, 102, 2992, ...\}$. Thus, "a discrepancy from proportionality as great or greater than that observed, will arise, subject to the conditions specified by the ancillary information, in exactly 3,095 trials out of 6,652,325 or approximately once in 2,150 trials."

He then went on to work out the lower limit of the 90% 2-sided CI (or a 95% 1-sided CI), for the odds ratio: i.e. for the odds, $\pi_{mono-z}/(1-\pi_{mono-z})$, of criminals to non-criminals in twins of monozygotic criminals divided by the corresponding odds $\pi_{di-z}/(1-\pi_{di-z})$, in twins of dizygotic criminals.

Let Y_{mono} be the number of MZ twins convicted. Fisher finds the value ψ_L such that

$$Prob[\,Y_{mono} \geq 10 \mid \psi_L\ ,\ y = 12\,] = 0.05.$$

He reports that this value is $1/0.28496 \approx 3.509$. In the Excel spreadsheet for Fisher's exact test and exact CI for OR (on website), you can verify that

[2] the range is 1 to 13; 0 cannot escape, since then there would be 13 convicted in the first row, but there are only 12 convicted in all.

indeed, with $\psi_L = 3.509$, $\;Prob[\,Y_{mono} \geq 10 \mid \psi = 3.509\ ,\ y = 12\,] = 0.05$.

One has to admire Fisher's ability, in 1935, to solve a polynomial equation of order 12, namely

$$\frac{1 + 102\psi + 2992\psi^2}{1 + 102\psi + 2992\psi^2 + \cdots + 476\psi^{12}} = 0.05.$$

It is ironic that while Fisher introduced the idea of conditioning to simplify significance tests of null and non-null ψ values, and through them, produced 'fiducial' limits that look like confidence limits, he did not give a **conditional** MLE for ψ: instead he gave the **unconditional** one:

> It is to be noticed that in this example of the fourfold table the notion of ancillary Information has been illustrated solely in relation to tests of significance and fiducial probability. No problem of estimation arises. ***If we want an estimate of ψ we have no choice but to take the actual ratio of the products of the frequencies observed in opposite corners of the table*** (i.e., ad/bc)

0.0.1 Point estimation of ψ under Hypergeometric Model

See also section 4.2 of Breslow and Day, Volume I. And see sections 7.3 & 7.4, and exercise 9.9 in McCullagh and Nelder's *Generalized Linear Models*, 2nd Edition.

It will come as a surprise to many that *there are 2 point estimators of* ψ:

one, the familiar – *unconditional* – based on the "2 independent Binomials" model, with two random variables y_1 and y_2, and

the other – *conditional* – based on the *single* random variable $y_1 \mid y$ with a Non-Central Hypergeometric distribution.

While the two estimators yield similar estimates when sample sizes are large, the estimates can be quite different from each other in small sample situations.

Estimator, based on Unconditional Approach:

The estimator derives from the principle that if there are two parameters θ_1 and θ_0, with Maximum Likelihood Estimators $\hat{\theta_1}$ and $\hat{\theta_0}$, then the Maximum Likelihood Estimator of θ_1/θ_0 is $\hat{\theta_1}/\hat{\theta_0}$.

Thus, since $\hat{\pi}_1 = 10/13$, and $\hat{\pi}_0 = 2/17$, we have

$$\hat{\psi}_{UMLE} = \frac{(10/13)/(2/13)}{(2/17)/(15/17)} = \frac{10 \times 15}{3 \times 2} = 25 = \frac{a \times d}{b \times c}.$$

Estimator, based on Conditional Approach:

We can find the Maximum Likelihood Estimate $\hat{\psi}_{CMLE}$ by inspecting the plot of the log L function, or using the Newton-Raphson approach, or the `optim` software, or trial and error, to find the solution of $d\log L/d\psi = 0$.

If we use Σ as shorthand for the denominator of $prob[\,y_1 \mid y\,]$, then $\hat{\psi}_{CMLE}$ is the solution of

$$\frac{y_1}{\psi} = \frac{d\log\Sigma}{d\psi} = \frac{d\Sigma}{d\psi} \times \frac{1}{\Sigma}.$$

Re-arranging, we find that $\hat{\psi}_{CMLE}$ is the solution of

$$y_1 = \mathrm{E}[\,Y_1 \mid \psi\,].$$

In this case the CMLE of ψ is the same as the estimate obtained by equating the observed and expected moment (the "Method of Moments").

Using trial and error in a spreadsheet (Excel has the central Hypergeometric probability) or R, or by a numerical search, we find that the value of ψ that satisfies this estimating equation is

$$\hat{\psi}_{CMLE} = 21.3.$$

It can be shown that $\hat{\psi}_{CMLE}$ is always *closer to null* ($\psi = 1$) than $\hat{\psi}_{MLE}$ is. The **CMLE is like a UMLE shrunk towards the null**. (Cf. Hanley JA, Miettinen OS. An Unconditional-like Structure for the Conditional Estimator of Odds Ratio from 2 × 2 Tables. Biometrical Journal 48 (2006) 1, 2334)

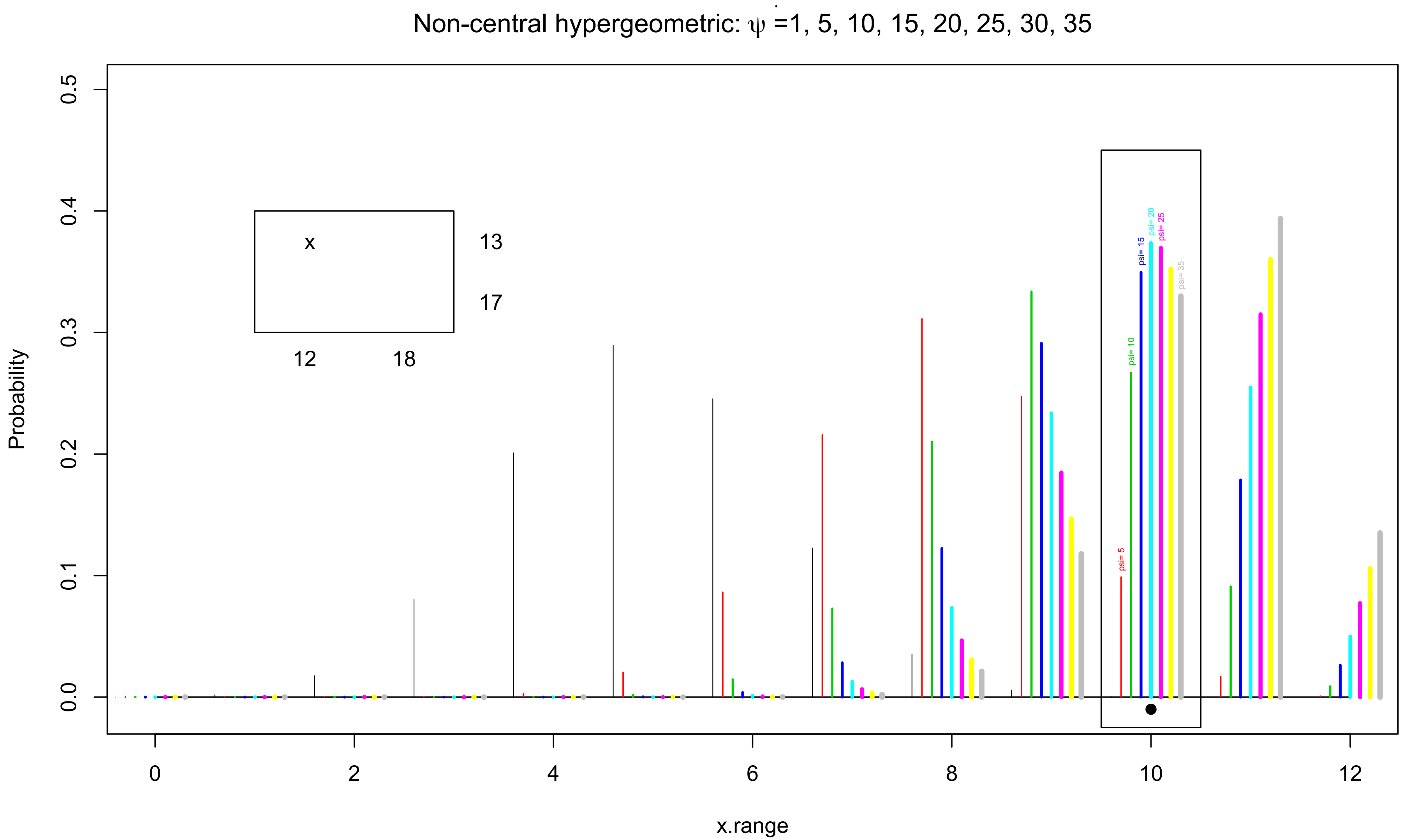
SPECIAL CASE: When **smallest of the marginal totals is 1**, i.e.,

	1	0	Total
1	a	·	1
0	·	·	$K \geq 1$
Total	$N_1 \geq 1$	$N_0 \geq 1$	

$\text{Info}[\log\psi] = \pi_\psi(1-\pi_\psi);\ \pi_\psi = \frac{N_1\psi}{N_1\psi + N_0}$

the Non-Central Hypergeometric Distribution of a is **Bernoulli**, with

$$Pr[a = 1] = \frac{N_1\psi}{N_1\psi + N_0};\ \text{logit} = \log\left(\frac{Pr[a=1]}{Pr[a=0]}\right) = \log\left(\frac{N_1}{N_0}\right) + \log\psi.$$



Effectively, by calculating $Prob_{non-central}(x|\psi)$ as a function of ψ, as is done inside the rectangle above, we are mapping out the likelihood function.

19 Individually matched case control studies

Examples: comparisons involving paired data (not all case-control)

- Response of **same** subject in each of 2 conditions (self-paired)
- Responses of **matched pair**, one in 1 condition, 1 in other
- Differences in **paired responses on interval scale, reduced to +/-**

		Result in Other Member		n
		+	−	
Result in One PAIR Member	+	a	b	
	-	c	d	
Total				n

		'Control' Exposed ?		n
		+	−	
'Case' Exposed ?	+	a	b	
	-	c	d	
Total				n

Extreme situations (1 or other / forced choice e.g. exercise 8.18, or who dies first among twin pairs discordant for handedness, or whether shorter/taller US presidential candidate won election)

		Shorter		n
		Won	Lost	
Taller	Won	-	b	
	Lost	c	-	
Total				n

Can also turn this table 'inside-out' and analyze using 'case-control' approach

		Loser		n
		Won	Lost	
Winner	Taller:	-	b	
	Shorter:	c	-	
Total				n

Example: HIV in twins in relation to order of delivery: mother to infant transmission of HIV-infection: n=66 sets of twins [Lancet, Dec. 14, 1991]

		2nd born		n
		HIV+	HIV-	
1st born	HIV+	10	18	
	HIV-	4	34	
				66

If we restrict to pairs with 1 HIV+ and 1 HIV- infant, we can display the data in case-control format

		HIV- infant born		n
		1st	2nd	
HIV+ infant born	1st		18	
	2nd	4		
				22

```
To: Goedert, James (NIH/NCI) [E]
Sent: Thu Mar 24 21:41:23 2011
Subject: 1991 Lancet study on twins and hiv

Dear Dr Goedert

Just this week, in teaching a graduate biostatistics course, I dragged out what I think of as
one of the classic epi studies that changed practice.. your and your colleagues' Lancet study
in 1991 on twins and hiv transmission.. High risk of HIV-1 infection for first-born twins

The students admired the design and analysis.. and I got to thinking that with the raw data it
would be a very valuable 'teaching dataset' for them to work on..

So I wonder if the raw data from the study are still available an if so whether you would be
willing to share them for teaching purposes.. I would be happy to receive them in any format,
electronic, paper, fax, whatever.. it would be great if on each twin we had all the covariates
as well as the HIV status and which twin pair it was..

There are a few classic paired-data datasets in epi and statistical  epi. I often use the one by
Miettinen and Trichopoulus from early 1970s. I think it was on  induced abortions and
subsequent ectopic pregnancy but it has very few covariates.. and we now know that the history
data in that study are suspect -- as the controls were probably quite coy about their histories
(The study was  done in Greece where induced abortions were illegal)

But yours would be a real nice one for them to 'get multivariate' about..

I would ensure the data were just for teaching.. and I would keep them in a locked teaching
website. I have had good luck with several generous authors and I hope it will be likewise here.

Sincerely

Jim Hanley
```

```
Dear Jim.

You bring back very fond memories - the origin of the idea (at a very small pediatric
AIDS meeting in California overlooking the Pacific ), assembling the collaborators and
contributors, monitoring the raw data as they arrived in my Fax machine (very modern
then), my amazement at the difference in risk by birth order, going back to contributors
to validate birth order (a few changed actually strengthening the difference, and
discussions with the Statistician on the analysis.

It will take some digging by a programmer or two who are still around, but the odds are
good of finding a clean data set. I will let you know.

Jim Goedert

==

From: Goedert, James (NIH/NCI) [E] [goedertj@mail.nih.gov]
Sent: April 7, 2011 2:18 PM
To: James Hanley, Dr.
Subject: Re: 1991 Lancet study on twins and hiv

Dear Dr Hanley.

One of our expert programmers, thinks she has found it on the NIH mainframe.
Please correspond with her regarding a format that would be useful for you.

Jim

========================
```

19.1 Mantel-Haenzel analysis of the 1:1 matched study

In the first paragraph, C&H motivate this chapter by noting that in individually matched case-control studies, one cannot add a **separate 'intercept' for each matched set or 'stratum'.** The best example of the **danger of doing this** (and those overfitting) is the example of matched pairs:

Below is how Breslow and Day Vol. I, section 7.1, illustrate why in this extreme situation, the 'unconditional' (and close to saturated) model yields a $\hat{\beta}$ value that is twice the value of the one obtained when the individual intercepts are conditioned out.

> **7.1 Bias arising from unconditional analysis of matched data**
>
> Use of the unconditional regression model (6.12) for estimation of relative risks entails explicit estimation of the α stratum parameters in addition to the β coefficients of primary interest. For matched or finely stratified data, the number of a parameters may be of the same order of magnitude as the number of observations and much greater than the number of β's. In such situations, involving a large number of nuisance parameters, it is well known that the usual techniques of likelihood inference can yield seriously biased estimates (Cox & Hinkley, 1974, p. 292). This phenomenon is perhaps best illustrated for the case of 1-1 pair matching with a single binary exposure variable x.
>
> Returning to the general set-up of section 6.2, suppose that each of the I strata consists of a matched case-control pair and that each subject has been classified as exposed ($x = 1$) or unexposed ($x = 0$). The outcome for each pair may be represented in the form of a 2×2 table, of which there are four possible configurations, as shown in (5.1). The model to be fitted is of the form
>
> $$pr(y = 1|x) = \frac{\exp(\alpha_1 + \beta x)}{1 + \exp(\alpha_1 + \beta x)},$$
>
> where $\beta = \log \psi$ is the logarithm of the relative risk, assumed constant across matched sets.
>
> According to a well-known theory developed for logistic or log-linear models (Fienberg, 1977), unconditional maximum likelihood estimates (MLEs) for the parameters α and β are found by fitting frequencies to all cells in the $2 \times 2 \times K$ dimensional con- figuration such that (i) the fitted frequencies satisfy the model and (ii) their totals agree with the observed totals for each of the two dimensional marginal tables. For the noo concordant pairs in which neither case nor control is exposed, and the nll concordant pairs in which both are exposed, the zeros in the margin require that the fitted frequencies be exactly as observed. Such tables provide no information about the relative risk since, whatever the value of p , the nuisance parameter a i may be chosen so that fitted and observed frequencies are identical ($\alpha_1 = 0$ for tables of the first type and $\alpha_1 = -\beta$ for tables of the latter to give probability 1/2 of being a case or control).
>
> The remaining $n_{10} + n_{01}$ discordant pairs have the same marginal configuration, and for these the fitted frequencies are of the form

	Exposure +	 -	
Case	μ	$1-\mu$	1
Control	$1-\mu$	μ	1
	1	1	2

> where
>
> $$\mu = pr_i(y = 1|x = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)},$$

> and
>
> $$1 - \mu = pr_i(y = 1|x = 0) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)},$$
>
> which can be expressed as
>
> $$\psi = \exp(\beta) = \left(\frac{\mu}{1 - \mu}\right)^2.$$
>
> The additional constraint satisfied by the fitted frequencies is that the total number of exposed cases, $n_{10} + n_{11}$, must equal the total of the fitted values, namely $(n_{10} + n_{01})\mu + n_{11}$. This implies $\hat{\mu} = n_{10}/(n_{10} + n_{01})$ and thus that the unconditional MLE of the relative risk is
>
> $$\hat{\psi} = \left(\frac{\hat{\mu}}{1 - \hat{\mu}}\right)^2 = \left(\frac{n_{10}}{n_{01}}\right)^2.$$
>
> the square of the ratio of discordant pairs (Andersen, 1973, p. 69).
>
> The estimate based on the more appropriate conditional model has already been presented in section 5.2. There we noted that the distribution of n_{10} given the total $n_{10} + n_{01}$ of discordant pairs was binomial with parameter $\pi = \psi/(1 + \psi)$. It followed that the conditional MLE was the simple ratio of discordant pairs
>
> $$\hat{\psi} = \frac{n_{10}}{n_{01}}.$$
>
> Thus the ***unconditional analysis of matched pair data results in an estimate of the odds ratio which is the square of the correct, conditional one***: a relative risk of 2 will tend to be estimated as 4 by this approach, and that of 1/2, by 1/4.
>
> While the disparity between conditional and unconditional analyses is particularly dramatic for matched pairs, it persists even with other types of fine stratification. Pike, Hill and Smith (1979) have investigated by numerical means the extent of the bias in unconditional estimates obtained from a large number of strata, each having a fixed number of cases and controls. Except for matched pairs, the bias depends slightly on the proportion of the control population which is exposed, as well as on the true odds ratio. Table 7.1 presents an extension of their results. For sets having 2 cases and 2 controls each, a true odds ratio of 2 tends to be estimated in the range from 2.51 to 2.53, depending upon whether the exposure probability for controls is 0.1 or 0.3. Even with 10 cases and 10 controls per set, an asymptotic bias of approximately 4% remains for estimating a true odds ratio of $\psi = 2$, and of about 15% for estimating $\psi = 10$.
>
> These calculations demonstrate the need for considerable caution in fitting unconditional logistic regression equations containing many strata or other nuisance parameters to limited sets of data.
>
> There are basically two choices: ***one should either use individual case-control matching in the design and the conditional likelihood for analysis; or else the stratum sizes for an unconditional analysis should be kept relatively large, whether the strata are formed at the design stage or post hoc***.

C&H begin with a case-control study concern the relationship between tonsillectomy history and the incidence of Hodgkin's disease, in which 'the controls were siblings of the cases, and the authors felt that the matching of cases and sibling controls should be preserved.' Since they also wished to control for age and sex, they 'therefore restricted their analysis to 85 matched case-control pairs in which the case and sibling control were of the same sex and matched for age within a specified margin, thereby reducing the original 174 cases and 472 controls to only 85 cases and 85 controls.

Just as you already found in getting your head around the tables for HIV transmission in twins, 'Tables such as Table 19.1 can be confusing because we are used to seeing tables that count subjects, while this **table counts case-control sets**.' '*In the analysis of individually matched studies the strata are case-control sets so that, in the notation of Chapter 18, t indexes sets.*'

'The four cells of the table corresponding to the four possible exposure configurations of a case-control set are illustrated in terms of a tree in Fig. 19.1.' These configurations can also be laid out as pair-specific 2 × 2 tables, of the same type as those used by B&D above. In the same order as those listed by C&H, these are

	Exposure			Exposure			Exposure			Exposure		
	+	-		+	-		+	-		+	-	
Case	1	0	**1**	0	1	**1**	1	0	**1**	0	0	**1**
Control	1	0	**1**	1	0	**1**	0	1	**1**	0	0	**1**
	1	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**2**
M-H												
$ad/2$			0			0			1/2			0
$bc/2$			0			1/2			0			0

Only two exposure configurations (the '<u>discordant</u>' sets in which the exposure status of case and controls differ) contribute to estimation and testing of the odds (or rate) ratio.

Supplementary Exercise 19.1

Refer to 'Occurrence of non-fatal myocardial infarction and the prevalence of obesity, smoking and vasectomy'. described in R file 'MI and Vasectomy Documentation, Data, R code' (data are at end of material).

These non-fatal myocardial infarctions arose out of a cohort of 4830 vasectomized/non vasectomized pairs of men matched from the membership files of a large group medical plan, on the basis of year of birth and calendar time of follow-up. For each pair, follow-up began when one of the pair members underwent vasectomy. There were no pairs of which both the vasectomized and non- vasectomized man suffered a myocardial infarction (MI).

For parts (i) to (vi) below, ignore the 2 variables obesity and smoking.

i. Taking a 'prospective' view, summarize the data in a 2×2 frequency table that (a) ignores (b) respects the matching.

ii. Taking a 'retrospective' view, summarize the data in a 2×2 frequency table that (a) ignores (b) (as C&H do in their Table 19.1) respects the matching.

iii. Repeat C&H's exercises 19.1, 19.2, 19.3 and 19.4, but using these vasectomy-MI data.

iv. Consider the data in the 2×2 table for set number 1 (the first of the 36 rows). What is the probability of observing this set-specific 2×2 table if the Rate ratio, θ [contrasting the MI rate in vasectomized men, λ_v, with the MI rate in non-vasectomized men, λ_n] equals 2? [3]

   Consider the data in the 2×2 table for set number 2 (the second of the 36 rows). What is the probability of observing this set-specific 2×2 table if, again, $\theta = 2$?

   Again, still with $\theta = 2$, what is the (joint) probability of observing the 36 2×2 tables that were observed? *You might find it more workable to calculate the log of this product, i.e., the sum of the logs of the 36 specific log-likelihoods.*

   Now, repeat the overall (i.e sum of 36 contributions) log-likelihood calculation for a range of θ values, from say $\theta = 1/10$ to $\theta = 10$, and plot this log-likelihood function.

v. Instead of this visual/numerical way of maximizing the log-likelihood, derive an analytic (closed form) ML estimator for θ (or $\log \theta$).

vi. Instead of analytically maximizing the log-likelihood, fit a `glm` in your favourite software package function to obtain the MLE of θ (or $\log \theta$).

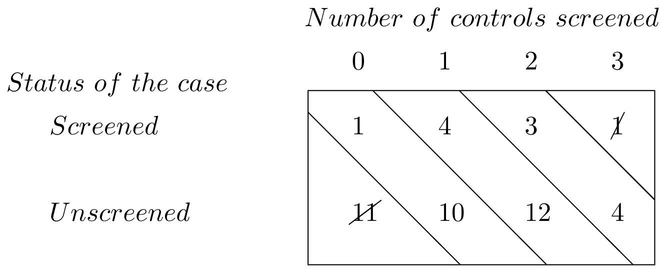
vii. How big would your dataset be if you restricted your analysis to sets where the two men in the same set were also matched on obesity and smoking? Can you suggest how we might recover some information from the sets where they were not fully matched on obesity and smoking?

[3] Clearly, λ_v and λ_n will be increasing functions of age, but for now we will assume that for every age, $\theta_{age} = \lambda_{age,v}/\lambda_{age,n}$ is the same, so that we can drop the *age* from the θ_{age} and refer to a single (proportional-hazards) hazard ratio (HR) θ.

19.2 Several controls per case

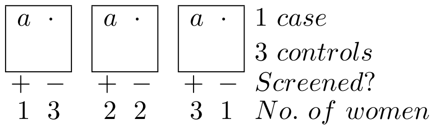
'Table 19.2 demonstrates the usual way such data are presented. However, it is very difficult to perceive any pattern even as to whether or not screening appears to be a protective. To understand the analysis, we shall start by reordering the data as a tree (Fig. 19.3).'

JH has added diagonal bands, used in Breslow and Day, to distinguish configurations with different marginal totals in a 2×2 table for a matched set of 4 women. Tables with different marginal totals (here, column totals, numbers screened and not) contribute different amounts of information regarding the log of the mortality rate ratio parameter θ. The amount of **expected information** concerning $\log \theta$ from the 3 table configurations is given below.



	Number of controls screened			
Status of the case	0	1	2	3
Screened	1	4	3	~~1~~
Unscreened	~~11~~	10	12	4

Informative configurations, single 1:3 set

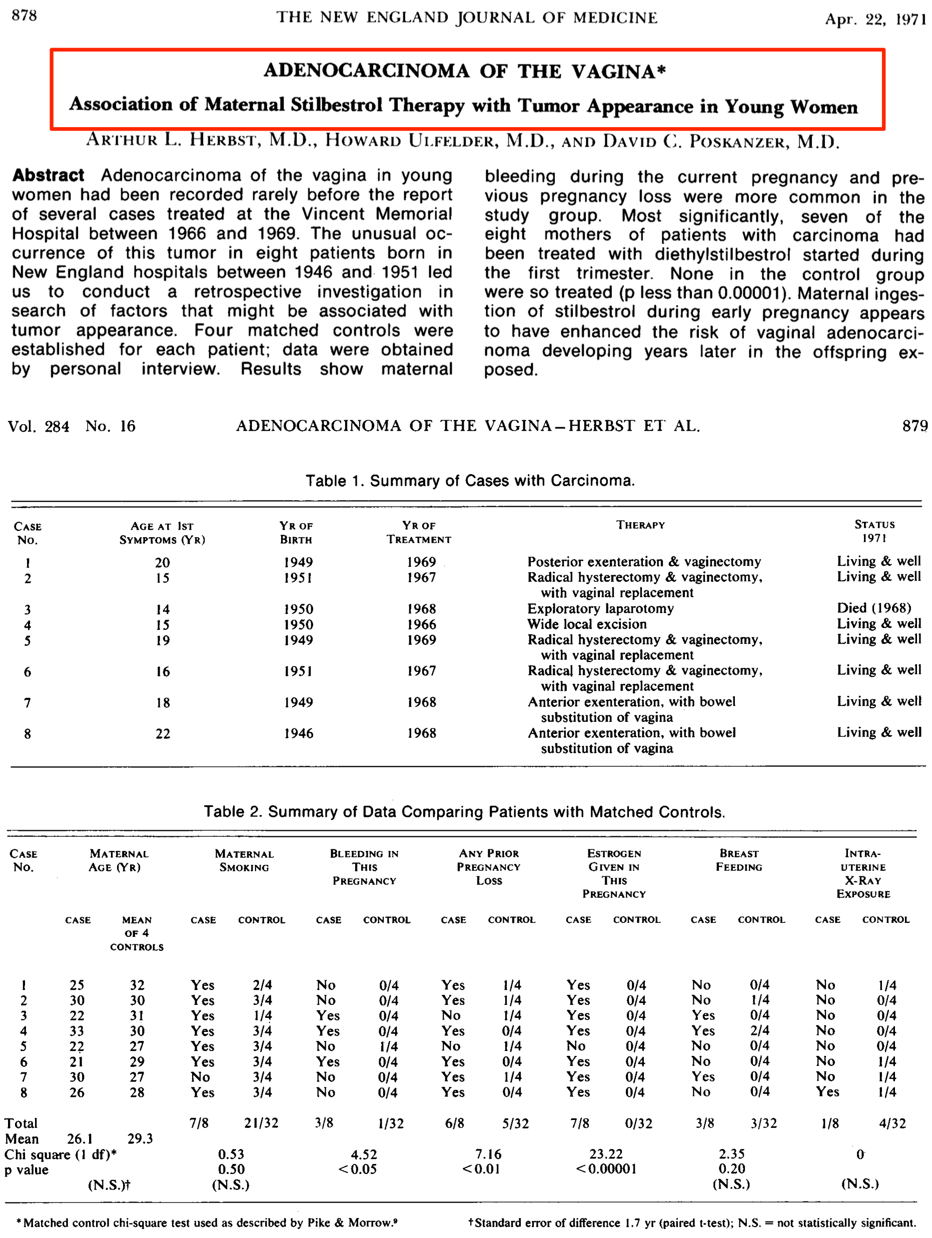


$$logit[Pr(a = 1)] = log(\psi) + \log \quad \left(\tfrac{1}{3}\right); \quad \left(\tfrac{2}{2}\right); \quad \left(\tfrac{3}{1}\right).$$

```
n.case.screened =c(1,4,3);n.case.not.screened =c(10,12,4) ;
o = log ( c(1/3, 2/2, 3/1) ) ; n.scr = 1:3; n=4
fit = glm( cbind(n.case.screened,n.case.not.screened) ~ 1+offset(o),family=binomial)
round(c(fit$coefficients,exp(fit$coefficients)),3) ; -1.205(SE:0.427)  0.300
mh=sum(n.case.screened*(1*(n-n.scr)/n))/sum(n.case.not.screened*(1*n.scr/n)); 0.304

Expected Info[log theta] per 1:3 set, if theta=1.0:  0.188    0.250   0.188
                                      if theta=0.3:  0.083    0.178   0.249
                                 sqrt(  1    /    (11*0.083+16*0.178+7*0.249) )  [1] 0.427
```

Supplementary Exercise 19.2



878 THE NEW ENGLAND JOURNAL OF MEDICINE Apr. 22, 1971

ADENOCARCINOMA OF THE VAGINA*

Association of Maternal Stilbestrol Therapy with Tumor Appearance in Young Women

ARTHUR L. HERBST, M.D., HOWARD ULFELDER, M.D., AND DAVID C. POSKANZER, M.D.

Abstract Adenocarcinoma of the vagina in young women had been recorded rarely before the report of several cases treated at the Vincent Memorial Hospital between 1966 and 1969. The unusual occurrence of this tumor in eight patients born in New England hospitals between 1946 and 1951 led us to conduct a retrospective investigation in search of factors that might be associated with tumor appearance. Four matched controls were established for each patient; data were obtained by personal interview. Results show maternal bleeding during the current pregnancy and previous pregnancy loss were more common in the study group. Most significantly, seven of the eight mothers of patients with carcinoma had been treated with diethylstilbestrol started during the first trimester. None in the control group were so treated (p less than 0.00001). Maternal ingestion of stilbestrol during early pregnancy appears to have enhanced the risk of vaginal adenocarcinoma developing years later in the offspring exposed.

Vol. 284 No. 16 ADENOCARCINOMA OF THE VAGINA – HERBST ET AL. 879

Table 1. Summary of Cases with Carcinoma.

Case No.	Age at 1st Symptoms (Yr)	Yr of Birth	Yr of Treatment	Therapy	Status 1971
1	20	1949	1969	Posterior exenteration & vaginectomy	Living & well
2	15	1951	1967	Radical hysterectomy & vaginectomy, with vaginal replacement	Living & well
3	14	1950	1968	Exploratory laparotomy	Died (1968)
4	15	1950	1966	Wide local excision	Living & well
5	19	1949	1969	Radical hysterectomy & vaginectomy, with vaginal replacement	Living & well
6	16	1951	1967	Radical hysterectomy & vaginectomy, with vaginal replacement	Living & well
7	18	1949	1968	Anterior exenteration, with bowel substitution of vagina	Living & well
8	22	1946	1968	Anterior exenteration, with bowel substitution of vagina	Living & well

Table 2. Summary of Data Comparing Patients with Matched Controls.

Case No.	Maternal Age (Yr)		Maternal Smoking		Bleeding in This Pregnancy		Any Prior Pregnancy Loss		Estrogen Given in This Pregnancy		Breast Feeding		Intra-uterine X-Ray Exposure	
	case	mean of 4 controls	case	control	case	control	case	control	case	control	case	control	case	control
1	25	32	Yes	2/4	No	0/4	Yes	1/4	Yes	0/4	No	0/4	No	1/4
2	30	30	Yes	3/4	No	0/4	Yes	1/4	Yes	0/4	No	1/4	No	0/4
3	22	31	Yes	1/4	Yes	0/4	No	1/4	Yes	0/4	Yes	0/4	No	0/4
4	33	30	Yes	3/4	Yes	0/4	Yes	0/4	Yes	0/4	Yes	2/4	No	0/4
5	22	27	Yes	3/4	No	1/4	No	1/4	No	0/4	No	0/4	No	0/4
6	21	29	Yes	3/4	Yes	0/4	Yes	0/4	Yes	0/4	No	0/4	No	1/4
7	30	27	No	3/4	No	0/4	Yes	1/4	Yes	0/4	Yes	0/4	No	1/4
8	26	28	Yes	3/4	No	0/4	Yes	0/4	Yes	0/4	No	0/4	Yes	1/4
Total			7/8	21/32	3/8	1/32	6/8	5/32	7/8	0/32	3/8	3/32	1/8	4/32
Mean	26.1	29.3												
Chi square (1 df)*			0.53		4.52		7.16		23.22		2.35		0	
p value			0.50		<0.05		<0.01		<0.00001		0.20			
	(N.S.)†		(N.S.)								(N.S.)		(N.S.)	

*Matched control chi-square test used as described by Pike & Morrow.[9]

†Standard error of difference 1.7 yr (paired t-test); N.S. = not statistically significant.

Focus on the data on maternal smoking and breast Feeding

i. Display the data in the same format as C&H's Table 19.2

ii. Repeat C&H's exercises 19.6 and 19.7, but using these matched data from Herbst et al.

iii. In the 2nd column of p193, after setting up the 3 (binomial-based) log-likelihoods from the 3 informative binomial configurations, C&H say that

> There is no simple expression for the maximum likelihood estimate and it is necessary to use a computer program to search for the maximum.

In fact, as is shown below the diagram at the bottom of the previous page of the Notes, these log-likelihoods arise within a generalized linear model, and so we can use any standard GLM software that allows for offsets. Apply this GLM approach to the data from Herbst et al., and compare with the M-H estimates. Do you think that, as C&H found with the cancer screening data, 'a Gaussian approximation fits quite closely'?

iv. Alter the wording of the title of Table 2 to reflect the modern way of thinking about what it is that is really compared in case-control studies.

Supplementary Exercise 19.3

Many textbooks on matched case-control studies, and some statistical packages, use the data linking induced abortions to ectopic pregnancies and to secondary Infertility (see Website). JH stopped using these data after he read the article from the 2 Dutch authors – on abortion and breast cancer.

i. Read (and summarize) the article by the 2 Dutch authors.

ii. Do you think the exquisitely matched case-control data on abortions, ectopic pregnancies and secondary Infertility, are valid? Why/why not?

iii. How about the results from the Danish study (also on website)?

Supplementary Exercise 19.4

Inside a 2 × 2 table, show the (central) hypergeometric r.v. representing:

(i) (a) 6/49 lottery result [numbers categorized into (rows) picked/not by player & (cols) drawn by Lottery]; (b) Keno, 2 picks (c) 10 picks.

(ii) result: Oscar for best actress (a) the day after (b) the day before the results became known.