16
Case-control studies

In a cohort study, the relationship between exposure and disease incidence is investigated by following the entire cohort and measuring the rate of occurrence of new cases in the different exposure groups. The follow-up allows the investigator to register those subjects who develop the disease during the study period and to identify those who remain free of the disease. In a *case-control* study the subjects who develop the disease (the cases) are registered by some other mechanism than follow-up, and a group of healthy subjects (the controls) is used to represent the subjects who do not develop the disease. In this way the need for follow-up is eliminated. If there is no relationship between exposure and disease incidence the distribution of exposure among the cases should be the same as the distribution among the controls.

Historically the aim of case-control studies was limited to testing for association between exposure and disease. Often little thought went into the selection of control groups, or even of cases to be studied. Frequently, studies were carried out using whatever cases could be traced from medical records at a given centre. In this rather careless climate, case-control studies fell into disrepute. However, it is now understood that properly conducted case-control studies allow *quantitative* estimates of exposure effects and this discovery has clarified the fundamental assumptions of the method. It has also contributed to a clearer understanding of the design of case-control studies issues and to a considerable improvement in the quality of studies.

We shall look first at estimating exposure effects and then consider how best to select controls. In the last section of the chapter there is a brief account of some of the difficulties which arise when case-control studies are based on prevalent rather than incident cases.

16.1 The probability model in the study base

Every case-control study of incidence can be seen within the context of an underlying cohort which supplies the cases on which the case-control study depends. A useful terminology refers to this underlying cohort, observed for the duration of the study, as the study *base*.

To estimate the quantitative relationship between exposure and disease

incidence we need to look more closely at what is happening in the study base. Consider the simple situation where the study base is divided into two groups, unexposed and exposed, and let π_0, π_1 be the probabilities that a member of the unexposed or the exposed group will fail over the period of the study and become a case.

The branches in the probability tree shown in Fig. 16.1 refer to the different possibilities for a randomly chosen member of the study base, and the events are taken in order of occurrence. The first branching of the tree refers to exposure. The subject may have been exposed (E+), or not (E−); we have taken the probability that a subject was exposed as 0.1, for illustration. The next branching refers to failure. The subject may fail (F), or survive (S); these are the probabilities already referred to as π_1 for the exposed group and π_0 for the unexposed group. The final branching refers to whether the subject is selected into the study or not; for illustration we have chosen a probability of 0.97 that a failure is registered and therefore included as a case, and a probability of 0.01 that a surviving subject is selected as one of the sample of controls. Note that the probability that a failure is registered is assumed to be the same for both exposure groups, and the probability that a healthy subject is chosen as a control is assumed to be the same for both exposure groups.

There are 8 possible outcomes for a member of the study base, corresponding to the 8 tips of the tree, but only 4 of these appear in the study. The four outcomes corresponding to the case-control study are: exposed cases, exposed controls, unexposed cases and unexposed controls. The numbers of subjects in these categories are referred to as D_1, H_1, D_0, H_0, respectively, where D refers to cases, H to healthy controls, and the suffices 1 and 0 refer to exposed and unexposed. The probabilities of the four outcomes appearing in the case-control study are calculated by multiplying conditional probabilities along the branches, and are shown to the right of the figure.

The estimation of the disease exposure relationship in the study base from the results of the case-control study may be approached using either a *retrospective* conditional argument or a *prospective* conditional argument. These correspond to two different ways of reorganizing the probability tree.

16.2 The retrospective probability model

In this argument we re-express our model as a model for the conditional probabilities of exposure given that the subject was a case (F) or a control (S). The reordering of the probability tree to reflect this argument is shown in Fig. 16.2. We define the parameter Ω_1 as the odds of having been exposed for a case. From Fig. 16.2, Ω_1 is related to the odds of failure in the study

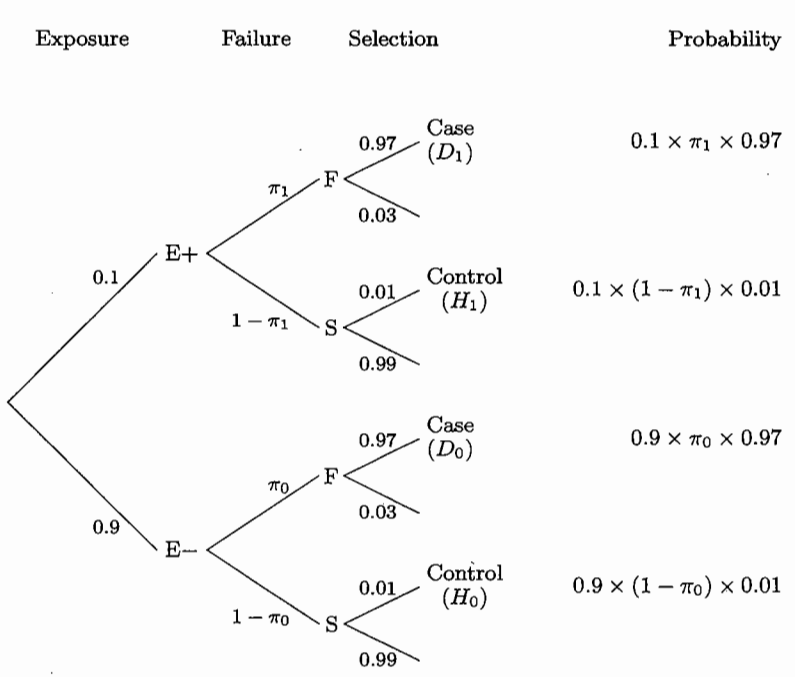


Fig. 16.1. The probability model in the study base.

base by the equations

$$\Omega_1 = \frac{0.1 \times \pi_1 \times 0.97}{0.9 \times \pi_0 \times 0.97} = \frac{0.1}{0.9} \times \frac{\pi_1}{\pi_0}.$$

The value of Ω_1 can be estimated by D_1/D_0, the ratio of exposed to unexposed cases. Similarly, we define Ω_0 as the odds of a having been exposed for a control. From Fig. 16.2,

$$\Omega_0 = \frac{0.1 \times (1-\pi_1) \times 0.01}{0.9 \times (1-\pi_0) \times 0.01} = \frac{0.1}{0.9} \times \frac{1-\pi_1}{1-\pi_0},$$

and the value of Ω_0 can be estimated by H_1/H_0, the ratio of exposed to unexposed controls. Finally the *odds ratio*

$$\frac{\Omega_1}{\Omega_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

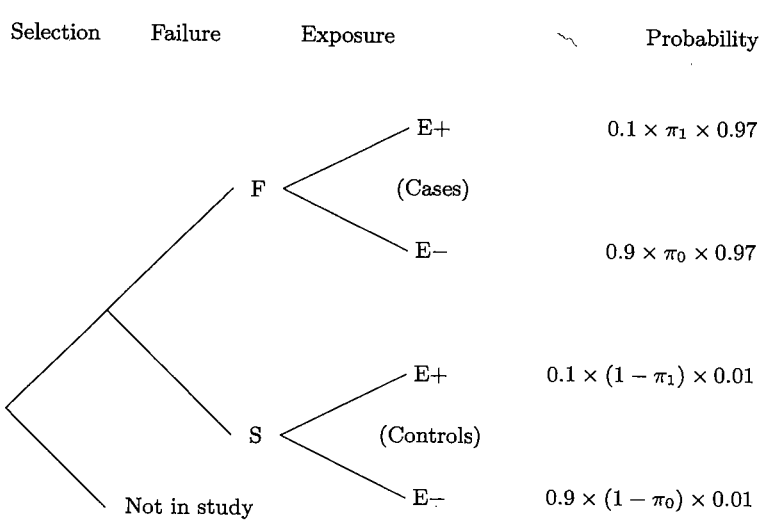


Fig. 16.2. The probability tree for the retrospective argument.

can be estimated by

$$\frac{D_1/D_0}{H_1/H_0}.$$

Thus although it is not possible to estimate π_0 and π_1 separately from a case-control study it *is* possible to estimate the odds ratio.

EXAMPLE: BCG VACCINATION AND LEPROSY

The data in Table 16.1 are from a rather unusual example of a case-control study in which the controls were obtained from a 100% cross-sectional survey of the study base.* The aim of the study was to investigate whether BCG vaccination in early childhood, whose purpose is to protect against tuberculosis, confers any protection against leprosy, which is caused by a closely related bacillus. New cases of leprosy reported during a given period in a defined geographical area were examined for presence or absence of the characteristic scar left by BCG vaccination. During approximately the same period, a 100% survey of the population of this area had been carried out, and this survey included examination for BCG scar. The tabulated data refer only to subjects under 35, because persons over the age of 35 at the time of the study would have been children at a time when vaccination was not widely available.

*From Fine, P.E.M. *et al.* (1986) *The Lancet*, August 30 1986, 499–502.

Table 16.1. BCG scar status in new leprosy cases and in a healthy population survey

BCG scar	Leprosy cases	Population survey
Present	101	46 028
Absent	159	34 594

Table 16.2. A simulated study with 1000 controls

BCG scar	Leprosy cases	Population survey
Present	101	554
Absent	159	446

Exercise 16.1. Estimate the odds of BCG vaccination for leprosy cases and for the controls. Estimate the odds ratio and hence the extent of protection against leprosy afforded by vaccination.

This example provides a good illustration of the potential economy of the case-control approach. Here a population survey was available for control but had it not been there would have been no need to carry out such a large-scale exercise. The precision of the odds ratio estimate is dominated by the precision of the odds for BCG scar among the 260 leprosy cases. Perhaps 1000 suitably chosen controls would be enough to estimate the corresponding odds among healthy subjects— there is little gain in precision to be obtained by using 80 000!

Exercise 16.2. Table 16.2 shows the results of a computer-simulated study which picked 1000 controls at random. What is the odds ratio estimate in this study?

16.3 The prospective probability model

In this argument we re-express our model in terms of the conditional probabilities of failure given selection into the study and given exposure status. The re-ordering of the conditional probability tree to reflect this argument is shown in Fig. 16.3. Define the parameter ω_1 as the *odds of being a case* for exposed subjects. By the odds of being a case we mean

$$\frac{\text{Probability of failure given that the subject is in the study}}{\text{Probability of survival given that the subject is in the study}}$$

From Fig. 16.3

$$\omega_1 = \frac{0.1 \times \pi_1 \times 0.97}{0.1 \times (1-\pi_1) \times 0.01} = \frac{0.97}{0.01} \times \frac{\pi_1}{1-\pi_1},$$

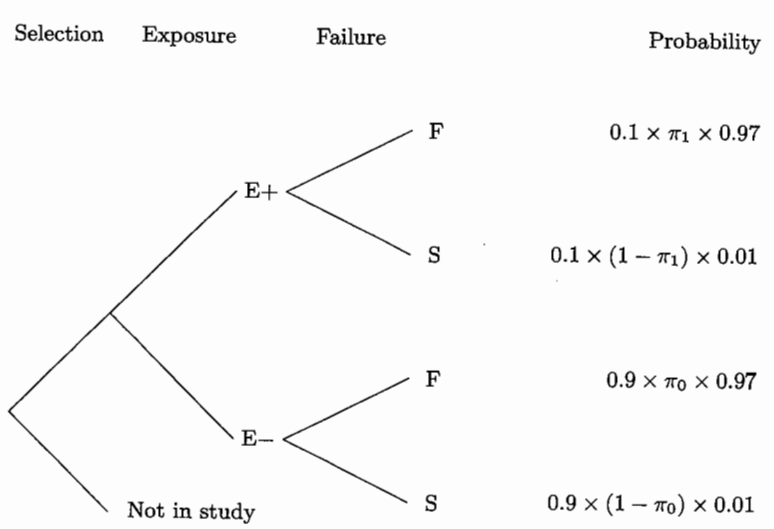


Fig. 16.3. The prospective probability model.

and this can be estimated by the case/control ratio among exposed subjects, D_1/H_1. Similarly the odds of being a case for unexposed subjects is

$$\omega_0 = \frac{0.9 \times \pi_0 \times 0.97}{0.9 \times (1-\pi_0) \times 0.01} = \frac{0.97}{0.01} \times \frac{\pi_0}{1-\pi_0},$$

which can be estimated by the case/control ratio among unexposed subjects, D_0/H_0. Finally, the odds ratio

$$\frac{\omega_1}{\omega_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)},$$

can be estimated by

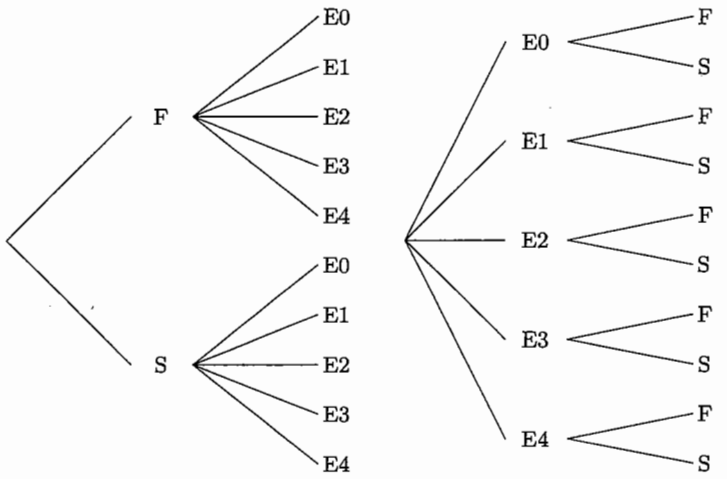
$$\frac{D_1/H_1}{D_0/H_0}.$$

This is the same estimate as that obtained from the retrospective approach since

$$\frac{D_1/D_0}{H_1/H_0} = \frac{D_1/H_1}{D_0/H_0} = \frac{D_1 H_0}{D_0 H_1}.$$

16.4 Many levels of exposure

In the retrospective argument it is the exposure status which is the response (outcome variable); in the prospective argument it is the disease status which is the response. The retrospective argument is more natural,



(a) Retrospective model (b) Prospective model

Fig. 16.4. Five exposure categories.

but the prospective argument leads to the same answers and is more convenient when studying exposures with many levels. This is illustrated by Fig. 16.4, which shows probability trees for both arguments when there are 5 exposure categories. Disease status is indicated by F (for cases) or S (for controls) and the 5 exposure categories are labelled E0 to E4. To construct a likelihood using the retrospective likelihood we must use a probability model for a response with 5 possible outcomes, but the prospective argument only requires the binary probability model. The odds of being a case for subjects in exposure category i is a constant multiple of the corresponding odds of failure in the study base; with the selection probabilities assumed in Fig. 16.1,

$$\omega_i = \frac{\pi_i}{1-\pi_i} \times \frac{0.97}{0.01}.$$

As the complexity of the exposure grouping increases, the retrospective probability model must become ever more complex, while the prospective model remains binary.

As an example of an exposure with more than two levels we shall look at a famous study carried out in the middle of the nineteenth century by William Guy.† This was possibly the first case-control study. The level of physical activity of the occupations of pulmonary tuberculosis outpatients (cases) was compared with that of other outpatients (controls). The data

†From Guy, W.A. (1843) *Journal of the Royal Statistical Society*, **6**, 197–211.

Table 16.3. Physical exertion at work of 1659 outpatients

Level of exertion in occupation	Pulmonary consumption (Cases)	Other diseases (Controls)	Case/ control ratio	Estimated odds ratio
Little	125	385	0.325	1.64
Varied	41	136	0.301	1.52
More	142	630	0.225	1.14
Great	33	167	0.198	1.00

Table 16.4. Alcohol and tobacco use by oral cancer cases and (controls)

Alcohol (oz/day)	Tobacco (cigarette equivalents per day)							
	0		1–19		20–39		40+	
0	10	(38)	11	(26)	13	(36)	9	(8)
0.1 – 0.3	7	(27)	16	(35)	50	(60)	16	(19)
0.4 – 1.5	4	(12)	18	(16)	60	(49)	27	(14)
1.6 +	5	(8)	21	(20)	125	(52)	91	(27)

are shown in Table 16.3. There are four levels of exposure corresponding to different levels of activity and the table shows the ratio of cases to controls. Each of these case-control ratios estimates some constant times the odds of failure conditional on exposure level. Since the constant depends on the probability of registration for cases and selection for controls it will be the same for all exposure levels and the case/control ratios can be compared as though they were the odds of failure.

Looking at the case/control ratios in this way, they suggest that there is a steady increase in the odds of failure (and hence the incidence rate) with decreasing level of physical activity. The table also shows odds ratio estimates with the 'great' activity category taken as reference. By definition, the odds ratio for this reference category is 1. The natural choice of reference category is the one with lowest exposure to adverse factor. In some cases, however, the natural reference category might contain very few cases and controls, leading to poor estimation of all the odds ratios; another reference category should then be chosen.

Exercise 16.3. Table 16.4 shows the distribution of 483 cases of oral cancer by level of alcohol consumption and level of tobacco consumption, together with the corresponding distribution for 447 controls.[‡] Calculate the case/control ratios, and describe the joint action of the two exposures.

[‡]From Rothman, K.J. and Keller, A.Z. (1972) *Journal of Chronic Diseases*, **23**, 711–716.

16.5 Incidence density sampling

We saw in Chapter 1 that, when the probabilities of failure are small, the risk and odds parameters are approximately equal. In these conditions, we showed in Chapter 5 that the risk parameter is also approximately equal to the cumulative rate, λt. It follows that

$$\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \approx \frac{\pi_1}{\pi_0} \approx \frac{\lambda_1}{\lambda_0}$$

for rare events. These ratios are known as the *odds ratio*, the *risk ratio*, and the *rate ratio*, and the condition for these to be approximately the same is usually described as the *rare disease assumption*. Taken together with the arguments developed in this chapter, we see that the odds ratio in a case-control study may be used to estimate the rate ratio in the underlying study base. There are two additional assumptions in this argument:

1. all subjects in the base are observed from the beginning of the study period, that is, there are no *late entries*;
2. all subjects who do not fail from the cause of interest will remain under observation until the end of the study period, that is, there is no *censoring*.

In practice, these assumptions are more likely to be violated than the rare disease assumption.

All of these assumptions can be guaranteed by the simple device of selecting a short enough study period. If insufficient cases would be obtained from such a study then the remedy is simple – carry out several *consecutive* short studies. The subjects remaining in the base at the end of one study immediately enter the next study. Each study then provides a separate estimate of the rate ratio, and provided this ratio remains constant over the whole study period, the information can be aggregated using methods very similar to those discussed in Chapter 15.

Taken to the limit, the total time available for the study may be divided into clicks which contain at most one case. Those clicks in which no case occurs are not informative so there is no purpose in drawing controls, but controls are drawn for all clicks in which a case occurs. Thus one or more controls are drawn from the study base immediately after the occurrence of each case. This design is termed *incidence density sampling*.

A study carried out in this way involves matching of controls to cases with respect to time. Methods for stratified case-control studies will be discussed in Chapter 18, but in the special case where the ratio of exposed to unexposed persons in the study base does not vary appreciably over the study period, it is legitimate to ignore the matching by time during the analysis.

One practical problem with this sampling method is that it is possible for the same individual to be included in the study more than once. For example, a control drawn at one point in time may later become a case or may be selected as a control a second time. Is it legitimate to carry out analyses which count the same person more than once? In Chapter 4 we saw that a single subject observed through several consecutive time bands can be treated as a series of different subjects, one for each band. In exactly the same way, in a case-control study it turns out to be correct to allow subjects to be sampled again in later time bands and treated as independent controls.

16.6 Nested case-control studies and case-cohort studies

An important use of incidence density sampling is in *nested* case-control studies, where case-control analysis is used in cohort studies. This is an attractive option whenever the assessment of exposure of any subject is, for some reason or other, expensive. For example, in dietary studies, individual diet may have been assessed by very detailed diary records of food intake, perhaps referring to several periods of time. The coding and transcription of such records for computer analysis is laborious and expensive. Much of this work is avoided in a nested case-control study by coding these records only for cases, as they occur, and for groups of controls drawn for each case. Since there is (usually) little to be gained by drawing more than five controls for each case, there are considerable savings to be made by such a strategy. We shall discuss the design and analysis of nested case-control studies in Chapter 33.

In recent years some authors have suggested that there are sometimes practical advantages in selecting controls by taking a single random sample of the cohort at the *beginning* of the study. This type of study has been termed a *case-cohort* or *case-base* study. If the disease is rare and there is little loss to follow-up, then the analysis may be carried out as usual, after first removing from the control sample any individuals who later became cases. However, if stratification by time becomes necessary the analysis is more difficult.

16.7 Selection bias

One important reason for obtaining wrong answers from case-control studies is incorrect sampling of controls (or cases) from the study base. This is called *selection bias*. It should be clear from this chapter that case-control studies will only yield unbiased estimates of

$$\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

if the selection probabilities for both cases and controls do not vary between exposure groups. Selection bias occurs when this is not true.

A study can only be truly convincing in this respect if its base is closely defined. The type of study with the best defined base is a nested case-control study, in which the study base consists of a documented and closely traced cohort. This method has proved particularly useful in occupational studies, where employment records identify an underlying cohort and pension schemes provide a mechanism for long term follow-up.

In a *geographically* based case-control study the base is defined by residence in a particular geographical area during the period of study. Although all such individuals are not specifically identified, it may nevertheless be possible to carry out a study in such a way that all cases are registered and controls drawn in a manner unrelated to exposure. Such studies require complete registration of disease in the study area, including capture of resident cases diagnosed and treated elsewhere. Control selection may also be difficult, since few countries have accurate and accessible population registers.

Another important base for case-control studies is the patient list of the family doctor. These lists offer good possibilities for representative control selection and for complete registration of cases particularly when, as in the United Kingdom, access to all medical services is channelled through the family practitioner.

For reasons of economy and convenience, a common choice is the *hospital-based* case-control study in which the case series is made up of all new cases presenting at one or more hospitals during the period of the study. Here the study base consists of the *catchment population* comprising all those persons who would have attended these hospitals if they had developed disease during this period. This is ill defined and it is difficult to demonstrate convincingly that the probability of control selection from the study base is independent of exposure. The device of using other patients, attending for unrelated conditions, has two clear difficulties:

1. catchment populations for different specialities in the same hospital do not necessarily coincide, and
2. patients who are sick with other diseases are not necessarily representative of the population of persons free of the disease of interest. In particular, factors associated with increased risk of these diseases may appear to be *protective* against the disease of interest simply because they are over represented in controls.

Against these difficulties must be set the claim that recall bias and other forms of differential exposure misclassification may be reduced when both case and control groups are hospital patients.

Two further points should be made briefly before concluding this section. First, matching is extremely useful in avoiding selection bias although

its use is more frequently advocated on the grounds of efficiency. We shall return to this discussion in Chapter 18. Second, it is important to draw attention to the fact that the best sampling scheme can be invalidated by poor subject compliance. If a substantial number of potential cases and controls refuse to participate there is considerable potential for bias as a result of differential compliance in different exposure groups. All too often case-control studies do not report compliance, and the potential for such bias is hard to assess.

16.8 Prevalent cases

If a case-control study is carried out using prevalent cases it is no longer a study of disease incidence and the odds ratio estimate cannot be interpreted as an estimate of a ratio of incidence rates. However, such studies can be used to study relationships of exposures to the prevalence of disease.

If the cases can be considered a random sample of those with disease in the population, and controls can be considered a random sample of the healthy section of the population, then the odds that a case was exposed divided by the odds that a control was exposed is an estimate of

$$\frac{\text{Prevalence odds in exposed population}}{\text{Prevalence odds in unexposed population}}.$$

When the prevalence in both groups is low this ratio is approximately equal to the prevalence in the exposed population divided by the prevalence in the unexposed population.

The remarks concerning sources of bias in incident case-control studies apply equally here. In particular, recall bias is a serious problem when interviewing prevalent cases who have been sick and in contact with medical professionals for some time. However, the main problems of interpretation are those of interpreting prevalence itself; the odds ratio is affected by factors which influence the *duration* for which a case, once diagnosed, remains in the sampling frame. These include not only factors related to survival, but factors relating to migration which may be complex and difficult to quantify.

Solutions to the exercises

16.1 The estimate of the odds *for* vaccination in leprosy cases is $101/159 = 0.635$ as compared with $46\,028/34\,594 = 1.331$ in the healthy subjects. The odds ratio estimate is $0.635/1.331 = 0.48$.

16.2 The odds ratio is

$$\frac{101/159}{554/446} = 0.51.$$

16.3 The case/control ratios are as follows:

Alcohol	Tobacco (cigs. per day)			
(oz/day)	0	1–19	20–39	40+
0	0.26	0.42	0.36	1.12
0.1–0.3	0.26	0.46	0.83	0.84
0.4–1.5	0.33	1.13	1.22	1.93
1.6 +	0.63	1.05	2.40	3.37

Because the frequencies in the table are small, there is much random variation, but there is an overall tendency for the ratios to increase both from left to right along rows, and from top to bottom down columns. This indicates that *both* variables have an effect on cancer incidence; there is an effect of tobacco when alcohol intake is held constant, and vice versa.