14 Confounding and standardization

14.1 Confounding

Epidemiological studies generally involve comparing the outcome over a period of time for groups of subjects experiencing different levels of exposure. Such studies are usually not controlled experiments but 'experiments of nature' of which the epidemiologist is a passive observer. In such investigations, there is always the possibility that an important influence on the outcome, which would have been fixed in a controlled experiment, differs systematically between the comparison groups. It is then possible that part of an apparent effect of exposure is due to these differences, and the comparison of the exposure groups is said to be *confounded.* Statistical approaches to dealing with the problem of confounding aim to correct, during analysis, for such deficiencies in the design of experiments of nature.

A particularly important potential confounding variable (or *confounder* in many epidemiological studies is the age of subjects. We shall consider an example in which subjects in a follow-up study are classified according to whether their age at the start of follow-up was less than 55 years or 55 years or more. Suppose that the breakdown between the two age groups is 0.8 : 0.2 and that the conditional probability of failure is 0.1 in the first age group and 0.3 in the second. When age is ignored the overall or *marginal* probability of failure is

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14.$$

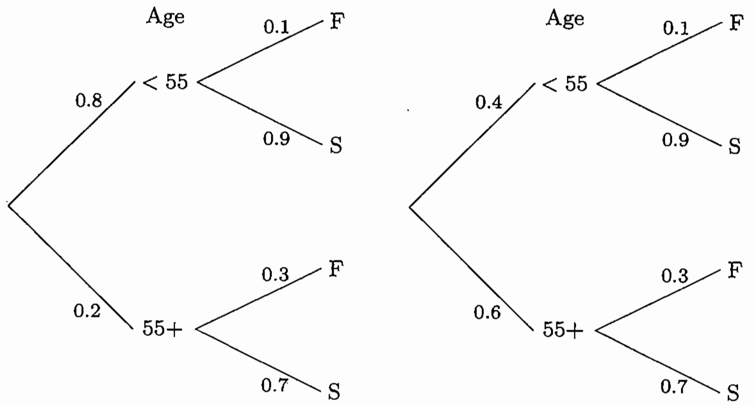
Now suppose that the age distribution differs between the two exposure groups, being 0.8 : 0.2 in the not exposed group but 0.4 : 0.6 in the exposed group (see Fig. 14.1). The marginal probability of failure for the unexposed group is still

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14,$$

but for the exposed group it is now

$$(0.4 \times 0.1) + (0.6 \times 0.3) = 0.22.$$

The marginal probabilities of failure now suggest an apparent effect of



Unexposed subjects

Exposed subjects

Fig. 14.1. Confounding by age.

exposure, but this is entirely due to the difference in age distributions between the exposed and unexposed subjects.

In this example the apparent effect of exposure is entirely due to age differences but confounding may also be partial, acting either to exaggerate or to dilute a real relationship. As an example of this, suppose the effect of exposure is to raise the probability of failure from 0.1 to 0.2 in the younger age group and from 0.3 to 0.5 for older subjects. When the age distribution is 0.8 : 0.2 in both exposure groups the overall effect of exposure is to increase the marginal probability of failure from

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14$$

in the unexposed group to

$$(0.8 \times 0.2) + (0.2 \times 0.5) = 0.26$$

in the exposed group. When the age distribution is 0.8 : 0.2 in the unexposed group and 0.4 : 0.6 in the exposed group the overall effect of exposure is to increase the marginal failure probability of failure from

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14$$

in the unexposed group to

$$(0.4 \times 0.2) + (0.6 \times 0.5) = 0.38$$

in the exposed group. Thus the overall effect of exposure appears greater when the age distributions differ than when they are the same.

These examples demonstrate that a third variable, such as age, can distort the relationship between an exposure and failure provided it is related to both exposure and failure. This dual relationship is often taken as the definition of a confounder. However, although it is a necessary condition for a variable to be a confounder, it is not sufficient: a confounder must also be a variable which would have been held constant in a controlled experiment. For example, in perinatal epidemiology, we might ask whether birthweight could be regarded as confounding the relationship between the receipt of proper antenatal care and the risk of perinatal death. Although birthweight is related to both antenatal care and perinatal risk, it cannot be regarded as a confounder since one of the *results* of successful antenatal care should be adequate birthweights. Since it would not make sense to envisage an experiment in which we varied the provision of antenatal care while maintaining the distribution of birthweight constant, differences in birthweight distribution cannot be regarded as a deficiency in the design of the experiment of nature. It is not, therefore, a confounder.

14.2 Correction for confounding

The linking of confounding to an imaginary experiment helps to clarify the ideas which lie behind statistical methods for dealing with the problem. There are two rather different approaches, and these closely mimic the ways in which extraneous influences are dealt with in experimental science.

The classical approach to experimentation is to hold constant all influences other than the experimental variable(s) of interest. For example, to avoid confounding by age, we would simply compare failure risks in exposed and unexposed subjects *of a fixed age* or, at least, falling within a narrow range of ages. The statistical comparison would then be of failure probabilities conditional upon age. The same comparison can be made in an non-experimental study by the analytical strategy called *stratification*. By dividing (or stratifying) the data according to age, the single experiment of nature in which age has not been adequately controlled is transformed into a series of smaller experiments within which age is closely controlled. The analysis then compares probabilities of failure between exposure groups within age bands. However, a consequence of this strategy is that individual strata may contain too little data to be informative on their own. The more finely we stratify the data, the more closely we control for confounding, but the sparser our data becomes within strata. This impasse may only be broken by making the further assumption that the comparisons estimate the same quantity within each stratum, and then combining the information from the separate strata. We shall defer further discussion of this approach to Chapter 15.

Holding extraneous variables constant is not the only model for good experimentation, although it is certainly the most familiar. In the twentieth century, experimentation has become a valuable tool in fields of study such as biology, in which such close control of experimental material and conditions is not possible. The idea of *randomization* has been central to this development; if we cannot ensure that experimental groups are identical in all important respects, then by assigning subjects to groups *at random*, we ensure that the probability distributions for extraneous variables do not differ between exposure groups. Comparisons between the groups can then be safely made.

Returning to the comparison of failure probabilities between exposure groups, it is rarely possible, in epidemiology, to use randomization to ensure that extraneous variables have equal distributions in the different exposure groups. However, it is possible to take account of differences in the distribution of a specific variable, such as age, by predicting the outcome for exposure groups which have the same age distribution. This is done by first estimating the age-specific probabilities of failure for each exposure group, and then using these to predict the marginal probabilities of failure for exposure groups which have a standard age distribution. This forms the basis of the second statistical approach to dealing with confounding, known in epidemiology as *direct standardization.*

14.3 Standardized rates

The remainder of this chapter concerns the use of direct standardization to compare *rates.* Since rates are probabilities per unit time they can be compared in the same way as failure probabilities. Age-specific failure rates are estimated for each of the groups being compared, and these are used to predict the marginal rates which would have been observed if the age distributions in the comparison groups had been the same as the standard age distribution. These estimates are called *standardized rates.*

The choice of the age distribution to use for standardization depends on the purpose of the analysis. It is quite common for the overall distribution of age, added over exposure groups, to be used as the standard, thus simulating the results of an experiment in which the total study group was randomly allocated between exposure categories. However, if one of our aims is to facilitate comparisons with other published studies, it is more useful to use an age distribution which is in general use. Several distributions are commonly used for this purpose. One is the age distribution of the world population, another is the age distribution for developed countries. Since there is no 'correct' standard there is much to be said in favour of using a *uniform* age distribution where the percentage falling in each age group is the same. One advantage of using a uniform age distribution is that the standardized rate is then directly proportional to the *cumulative rate* for a subject experiencing the age-specific rates from the study throughout life.

Table 14.1. IHD incidence rates per 1000 person-years

	Exposed			Unexposed		
	(< 2750 kcal)			(≥ 2750 kcal)		
Age	Cases	P-yrs	Rate	Cases	P-yrs	Rate
40–49	2	311.9	6.41	4	607.9	6.58
50–59	12	878.1	13.67	5	1272.1	3.93
60–69	14	667.5	20.97	8	888.9	9.00
Total	28	1857.5	15.07	17	2768.9	6.14

Direct standardization is most commonly used when comparing quite large groups, such as the populations of different countries or regions. When used with less extensive data it will yield statistically unreliable estimates if some of the age-specific rates, although based on very few cases, receive appreciable weight in the analysis.

To illustrate the technique of direct standardization we shall return to study of ischaemic heart disease and energy intake, discussed in Chapter 13. The incidence of ischaemic heart disease in the exposed group (low energy-intake) is 15.1 per 1000 person-years while the rate in the unexposed group is 6.1 per 1000 person-years. These rates, which take no account of any possible confounding effect of age, are often referred to as *crude* rates to distinguish them from standardized rates.

Table 14.1 shows the data stratified by 10-year age bands. The age distribution is different in the two exposure groups; this may be seen by converting the person-years to a proportion of the total person-years in each group giving 0.168, 0.472, and 0.359 in the three age bands for the exposed (low energy-intake) group and 0.210, 0.459, and 0.321 for the unexposed (high energy-intake) group. These age differences might explain some of the difference in the crude IHD incidence rates.

Using the uniform age distribution as standard, our estimate of the marginal rate for a group of exposed subjects with a uniform age distribution is

$$(0.333 \times 6.41) + (0.333 \times 13.67) + (0.333 \times 20.97) = 13.67$$

per 1000 person years and, for a group of unexposed subjects with a uniform age distribution, it is

$$(0.333 \times 6.58) + (0.333 \times 3.93) + (0.333 \times 9.00) = 6.50$$

per 1000 person-years. The standardized rates for the two groups are therefore 13.7 and 6.5 per 1000 person-years. These do not differ greatly from the crude rates of 15.1 and 6.1 per 1000 person-years, showing that the confounding effect of age is small in this case.

Exercise 14.1. Find the standardized rates for the exposed and not exposed groups using as standard the age distribution with probabilities of 0.2, 0.5, and 0.3 in the three age bands.

★ 14.4 Approximating the log likelihood

When there are three age bands, as in the IHD and energy example, the standardized rate parameter takes the form of a weighted sum of the age-specific rate parameters,

$$W^1\lambda^1 + W^2\lambda^2 + W^3\lambda^3,$$

where

$$\lambda^1, \lambda^2, \lambda^3$$

are the rate parameters for the age bands and

$$W^1, W^2, W^3$$

are the probabilities of the standard age distribution. Since λ^1, λ^2 and λ^3 have independent log likelihoods, we can use the ideas introduced in section 13.4 and Appendix C to derive a Gaussian approximation to the profile log likelihood for the standardized rate. The most likely value is

$$W^1M^1 + W^2M^2 + W^3M^3$$

where $M^1 = D^1/Y^1$ is the most likely value of the age-specific rate parameter in band 1, and similarly expressions hold for bands 2 and 3. The standard deviation of the Gaussian approximation is

$$\sqrt{(W^1S^1)^2 + (W^2S^2)^2 + (W^3S^3)^2}$$

where $S^1 = \sqrt{D^1}/Y^1$ is the standard deviation of the Gaussian approximation to the log likelihood for λ^1, again with similar expressions for bands 2 and 3.

For the IHD and energy example the proability weights are

$$W^1 = W^2 = W^3 = 0.333.$$

The age-specific rate for the first age band of the exposed group is 6.41 and the corresponding standard deviation is

$$\sqrt{2}/311.9 = 0.00453,$$

or 4.53 per 1000 person-years. The most likely values for the rates in the other two age bands are 13.67 and 20.97 with standard deviations 3.94 and

5.61 per 1000 person-years. The standard deviation of the standardized rate is therefore

$$\sqrt{(0.333 \times 4.53)^2 + (0.333 \times 3.94)^2 + (0.333 \times 5.61)^2} = 2.74$$

per 1000 person-years.

Exercise 14.2. Show that the standard deviation of the standardized rate for the unexposed group is 1.63 per 1000 person-years.

LOG TRANSFORMATION OF STANDARDIZED RATES

Just as for any other rate, Gaussian approximations to the log likelihood are more accurate when related to the *log* of the standardized rate. The most likely value on the log scale is, of course, just the log of the standardized rate, and the corresponding standard deviation can be calculated by using the rule described in Chapter 9. There we saw that the standard deviation of the Gaussian approximation to the likelihood for $\log(\lambda)$ is obtained from the standard deviation of the Gaussian approximation to the likelihood for λ by multiplying by $1/M$, where M is most likely value of λ. It follows that for the example of energy intake and IHD incidence, the standard deviations of the standardized rates on a log scale are $2.74/13.67 = 0.200$ and $1.63/6.50 = 0.251$.

A simple extension of the same ideas allows us to calculate estimates and confidence intervals for the ratio of two standardized rates. The log of this ratio is equal to the difference between the logarithms of the two standardized rates, and from section 13.4 and Appendix C the standard deviation of the log of the ratio of the standardized rates is

$$\sqrt{(0.200)^2 + (0.251)^2} = 0.321.$$

This can be used to obtain a confidence interval for the ratio of the standardized rates by using the error factor

$$\exp(1.645 \times 0.321) = 1.696.$$

Exercise 14.3. Use this error factor to find an approximate 90% confidence interval for the ratio of the two standardized rate parameters.

Solutions to the exercises

14.1 The estimated standardized rates are

$$(0.2 \times 6.41) + (0.5 \times 13.67) + (0.3 \times 20.97) = 14.41$$

for the exposed group, and

$$(0.2 \times 6.58) + (0.5 \times 3.93) + (0.3 \times 9.00) = 5.98$$

for the unexposed group.

14.2 The standard deviations of the age-specific rates are 3.29, 1.76, and 3.18 respectively. The standard deviation of the standardized rate is

$$\sqrt{(0.333 \times 3.29)^2 + (0.333 \times 1.76)^2 + (0.333 \times 3.18)^2} = 1.63.$$

14.3 The ratio of standardized rates is 13.67/6.50 = 2.10 and the 90% range for this is from 2.10/1.696 = 1.24 to 2.10 × 1.696 = 3.56 .