factor. The word "treatment" is generally used to describe an agent applied specifically to affect the outcome factor under consideration (as was true for all the examples in the first paragraph of this chapter). The term "risk factor," borrowed from epidemiology, is used when exposure to the agent is accidental or uncontrollable, or when the agent is applied for some purpose other than to affect the specific outcome factor under consideration. An example would be the study of the effect of smoking on the incidence of lung cancer. The use of the term "risk factor" does not in itself imply that the agent is "risky" or in fact, that risk enters the discussion at all. We use whichever term ("treatment" or "risk factor") appears more natural in context.

In later chapters we talk about quantities or labels that measure the presence, absence, level or amount of a risk factor, treatment, outcome factor, or confounding factor. Such quantities or labels will be termed *variables.* In studying the effect of seat belts on accident mortality (Example 1.1) we may define a risk variable taking the value 1 or 0, depending on whether or not the driver was wearing a seat belt at the time of the accident. The logical distinction between a factor and a variable which measures that factor is not always made in the literature, but it can be useful.

The term "comparison group" is used interchangeably with the more familiar "control group." When the important comparison is between a proposed new treatment and the present standard treatment, the standard treatment (rather than no treatment) should be given to the comparison group. In dealing with risk factors it is natural to speak of "risk groups" or of "exposed" and "nonexposed" groups. We may have several different "exposed" or "treatment" groups, corresponding to different levels of the risk factor or treatment.

CHAPTER 2

Confounding Factors

In the discussion of Example 1.1 (effect of wearing seat belts on auto accident fatality) we saw that a background factor (speed at impact) could seriously distort the estimate of the effect of the risk factor on the outcome. The distortion will arise whenever two conditions hold:

1. The risk groups differ on the background factor.
2. The background factor itself influences the outcome.

Background factors which satisfy conditions 1 and 2 are called confounding factors. If ignored in the design and analysis of a study, they may affect its conclusions, for part of the effect of the confounding factor on the outcome may appear to be due to the risk factor. Table 1.1 is misleading because the effect on accident fatality apparently due to wearing seat belts (the risk factor) is actually due to speed at impact (the confounding factor).

In Section 2.1 we show by another example how the effect of a risk factor can

sometimes be disentangled from that of a confounding factor. A useful measure of the likely influence of a confounding factor on the estimate of treatment effect is the *bias*. Section 2.2 quantifies the term "bias" and briefly introduces the concepts of precision and statistical significance. The qualitative discussion in Chapter 1 of the relation among the risk, outcome, and confounding factors is extended in Section 2.3. Formulas relating bias, standard error, and mean squared error are given in Appendix 2A.

2.1 ADJUSTMENT FOR A CONFOUNDING FACTOR

In Example 1.1 the risk factor had no actual effect on the outcome. Table 1.2 shows that its apparent effect was due entirely to the confounding factor. In most studies many factors will each have some effect on the outcome and the investigator will want to estimate the magnitude of the treatment effect after allowing for the effect of the other factors. An example will show that sometimes this can be done quite easily.

Example 2.1 Coffee drinking, obesity, and blood pressure: Suppose that a physician, Dr. A, wants to assess the effect on the diastolic blood pressure of his male patients of their regularly drinking coffee. We shall consider just two levels of the risk factor, coffee drinking, corresponding to patients who drink coffee regularly (the drinkers) and patients who do not drink coffee regularly (the nondrinkers). The outcome variable, diastolic blood pressure, is a numerical measurement. Dr. A is unwilling to instruct his patients to drink coffee or to stop drinking coffee, but he can rely (let us say) on truthful answers to questions on the subject in his medical records.

Because he knows that blood pressure is also influenced by weight—overweight patients tend to have higher blood pressures that those of normal weight—Dr. A classifies all his male patients by obesity (overweight or not overweight) as well as by coffee drinking. Dr. A calculates the average diastolic blood pressure in millimeters of mercury (mm Hg) of patients in the four categories. We shall suppose that the average diastolic blood pressure among the nondrinkers who are not overweight is 70 mm Hg, but that among the nondrinkers who are overweight the average is 90 mm Hg. Let us also suppose that the effect of drinking coffee regularly is to increase blood pressure by exactly 4 mm Hg, and that there are no other complicating factors. Then the average diastolic blood pressures among the drinkers who are and who are not overweight are 94 and 74 mm Hg, respectively. These assumptions are summarized in Table 2.1. Notice that we have not yet specified the numbers of patients in each category.

Suppose that Dr. A were to attempt to estimate the effect of drinking coffee on blood

Table 2.1 Average Diastolic Blood Pressures (mm Hg)

	Overweight	Not Overweight
Drinkers	94.0	74.0
Nondrinkers	90.0	70.0

Table 2.2 "Even" Distribution for Dr. A's Patients

	Overweight	Not Overweight	Total
Drinkers	100	300	400
Nondrinkers	50	150	200

pressure ignoring the effect of obesity. He would compare the average blood pressure of the drinkers with that of the nondrinkers. To calculate these averages Dr. A will need to know the numbers of his patients in each category of Table 2.1. We shall suppose that he has 600 male patients in all, and will consider two different distributions of their numbers, an "even" distribution (Table 2.2) and an "uneven" distribution (Table 2.3).

In the "even" distribution the proportion of overweight patients among the drinkers ($100/400 = 0.25$) is the same as that among the nondrinkers ($50/200 = 0.25$). In statistical language, Table 2.2 exhibits no association between coffee drinking and obesity. The average blood pressure among all the drinkers is the weighted mean of the averages on the top line of Table 2.1, weighted by the numbers of patients contributing to each average. From Table 2.1 and Table 2.2, this is

$$\frac{(94.0 \times 100) + (74.0 \times 300)}{100 + 300} = 79.0 \text{ mm Hg.}$$

From the second line of the same tables, the average blood pressure among the nondrinkers is

$$\frac{(90.0 \times 50) + (70.0 \times 150)}{50 + 100} = 75.0 \text{ mm Hg.}$$

Dr. A's estimate of the average increase in blood pressure due to coffee drinking would be

$$79.0 - 75.0 = 4.0 \text{ mm Hg.}$$

This is the correct answer because it agrees with the rise of 4.0 mm Hg that we assigned to coffee drinking. To summarize, if there is no association between the risk factor, coffee drinking, and the background factor, obesity, among Dr. A's patients, a straight comparison of average blood pressures among the drinkers and among the nondrinkers will be adequate. Here the background factor satisfies condition 2 of the definition of a confounding factor given at the beginning of this chapter, but it does not satisfy condition 1 and so is not a confounding factor.

If, instead, Dr. A's patients follow the "uneven" distribution of Table 2.3, then both parts of the definition will be satisfied, as Table 2.3 does indicate an association between coffee drinking and obesity. Obesity will now be a confounding factor. The average blood

Table 2.3 "Uneven" Distribution for Dr. A's Patients

	Overweight	Not Overweight	Total
Drinkers	300	100	400
Nondrinkers	50	150	200

pressure among the drinkers who visit Dr. A will be

$$\frac{(94.0 \times 300) + (74.0 \times 100)}{300 + 100} = 89.0 \text{ mm Hg.}$$

Among the nondrinkers, the average blood pressure will be

$$\frac{(90.0 \times 50) + (70.0 \times 150)}{50 + 150} = 75.0 \text{ mm Hg.}$$

The crude estimate of the average increase in blood pressure due to coffee drinking, namely

$$89.0 - 75.0 = 14.0 \text{ mm Hg,}$$

would be incorrect.

Of course, this problem does not arise if Dr. A assesses the effect of coffee drinking separately among his overweight patients and among his patients who are not overweight. He then uses the values given in Table 2.1 to arrive at the correct estimate of the effect of coffee drinking, namely that it increases average blood pressure by 4.0 mm Hg among both classes of patient.

However, Dr. A may prefer to calculate a single summary measure of the effect of coffee drinking on blood pressure among all his patients. He can do this by applying the average blood pressures in Table 2.1 to a single hypothetical standard population consisting, for example, of 50% patients of normal weight and 50% patients who are overweight. These calculations would tell him what the average blood pressures would be in this standard population (*a*) if they all drank coffee and (*b*) if none of them drank coffee. The calculations give

$$(94.0 \times 0.50) + (74.0 \times 0.50) = 84.0 \text{ mm Hg}$$

for the average blood pressure among the patients in the standard population if they were all to drink coffee, and

$$(90.0 \times 0.50) + (70.0 \times 0.50) = 80.0 \text{ mm Hg}$$

if none of them were to drink coffee. The comparison between these two averages gives the correct result.

This adjustment procedure is an example of standardization, to be described further in Chapter 7.

2.2 BIAS, PRECISION, AND STATISTICAL SIGNIFICANCE

For many reasons the estimated treatment effect will differ from the actual treatment effect. We may distinguish two types of error: random error, and systematic error or *bias*. This book is primarily concerned with the bias introduced into the estimate of treatment effect by confounding factors. However, to illustrate the distinction between random error and bias, we give a simple example not involving a treatment or confounding factor.

Example 2.2 The Speak-Your-Weight machines: An old Speak-Your-Weight machine is rather erratic but not discernibly off-center. A new machine gives perfectly reproducible results but may be off-center because of an incorrect setting at the factory.

A man of weight 170 lb who weighs himself five times on the old machine may hear weights of 167, 172, 169, 173, and 168 lb. The new machine might respond 167, 167, 167, 167, and 167 lb. The old machine is exhibiting random error, the new machine systematic error. Of course, a Speak-Your-Weight machine could easily be both erratic and off-center. Such a composite machine, exhibiting the defects described both for the old and the new machines, might give readings of 164, 169, 166, 170, and 165 lb, which are subject to random error and to bias.

There is clearly no way to distinguish between these types of error by a single measurement (e.g., of 167 lb). Implicit in the distinction between random error and systematic error is the notion of repetition: random error would approximately cancel out if repeated measurements were taken and averaged [in this example the average of the five weights spoken by the old machine, $(167 + 172 + 169 + 173 + 168)/5 = 169.8$ lb is quite close to the true weight, 170 lb], while systematic error is impervious to averaging (the average of the weights spoken by the new machine is still 167 lb).

Statistical techniques such as significance testing and the calculation of standard errors and confidence intervals are often helpful in gauging the likely effect of random error on the conclusions of a study. These techniques cannot in themselves assess the effect of systematic error.

2.2.1 Bias

We can now attempt a formal definition of the term "bias."

Definition: The *bias* of an estimator is the difference between the average value of the estimates obtained in many repetitions of the study and the true value of what it is estimating.

By thinking of an estimator as a procedure that produces estimates, we introduce the notion of repetition into the definition. Because of random error the estimate would change from repetition to repetition, although the estimator, the procedure used to derive the estimates, would not change. The definition emphasizes that the bias is a number, positive or negative. This contrasts with the common use of the term as an abstract noun, or even as a general insult to impugn any study that disagrees with one's own opinions: "This study is biased."

Confounding factors are the major source of bias in nonrandomized studies (in both the common and the technical usage of the term "bias") and it is with the bias due to confounding factors that this book is primarily concerned. Other possible sources of bias will be mentioned in Chapter 5.

Unfortunately, the definition we have just given rarely enables the bias to be calculated, even in terms of the unknown true treatment effect, since studies

are not repeated and we cannot say what would happen if they were. With the partial exceptions of matching and standardization, all the techniques described herein depend on assumed *statistical models,* which state what would happen in hypothetical repetitions. A simple statistical model for the weight X registered by the "old" Speak-Your-Weight machine of Example 2.2 has

$$X = \mu + \epsilon,$$

where μ is the true weight of the man and ϵ denotes a random error. The true weight μ would not change from repetition to repetition, but the random error ϵ would change, with an average value close to zero after many repetitions.

By contrast, the "new" Speak-Your-Weight machine has

$$X = \mu + b,$$

where μ is the true weight, as before, and b is a systematic error which does not change from repetition to repetition.

Rarely can the validity of an assumed statistical model be checked directly. The methodological chapters (Chapters 6 to 12) will discuss the statistical models demanded by each technique and such indirect checks of the validity of these models as are available. The equations are not usually as simple as those given above because they must relate the outcome variable to the treatment and confounding variables of interest and to the measure chosen to describe the effect of the treatment. The distribution of the random error must also be specified.

2.2.2 Precision and Statistical Significance

The precision of an unbiased estimator of a treatment effect is usually measured by the variance of the estimator or by the square root of this variance, the standard error. The smaller the variance or standard error, the more precise is the estimator. The standard error of a biased estimator still measures the influence of random error on the estimator, but it gives no clue as to the magnitude of systematic error. As systematic error is usually a more serious threat to the validity of observational studies than is random error, this book assesses techniques by their ability to reduce bias and places only a secondary emphasis on precision. However, most of the procedures we describe permit the calculation of standard errors of estimated treatment effects.

The *mean squared error* of an estimator is defined as the mean value, in hypothetical repetitions, of the square of the difference between the estimate and the true value. We show in Appendix 2A that the mean squared error can be calculated as the variance plus the square of the bias. It provides a useful criterion for the performance of estimators subject to both systematic and random error.

The function of a test of *statistical significance* is to determine whether an

apparent treatment effect could reasonably be attributed to chance alone. When applied to data from well-designed randomized studies, significance tests can effectively demonstrate the reality of the observed treatment effect. In nonrandomized studies, where systematic error will usually provide a more plausible explanation of an observed treatment effect than will random variation, significance tests are less crucial. Nevertheless, they can, if carried out after adjustment for confounding factors, be useful indicators of whether the observed treatment effect is real.

The concepts of precision and statistical significance are closely related. Whether an estimated treatment effect is statistically significant depends not only on the magnitude of the estimated effect but also on the precision of the estimator. A useful rule of thumb, based on an assumed normal distribution, holds an estimated treatment effect at least twice its standard error from the no-effect value to be on the borderline of statistical significance, and to be highly significant if away by at least three times its standard error.

The methodological chapters include some discussion of tests of statistical significance and of the precision of estimators.

2.3 SOME QUALITATIVE CONSIDERATIONS

For the two examples involving confounding factors discussed so far (seat belts to reduce accident fatalities, effect of coffee drinking on blood pressure), the assumed relations among the factors are summarized in Figures 2.1 and 2.2.

In these figures an arrow ($\rightarrow$) denotes a direct casual link. That is, $A \rightarrow B$ if a change in A would result in a change in B if all other factors listed in the figure do not change. A double arrow ($\leftrightarrow$) denotes a possible association between factors A and B which may not have a simple causal interpretation. The two factors may influence each other and may both be influenced by other factors not included in the figure. The relation of primary interest is, as always, that

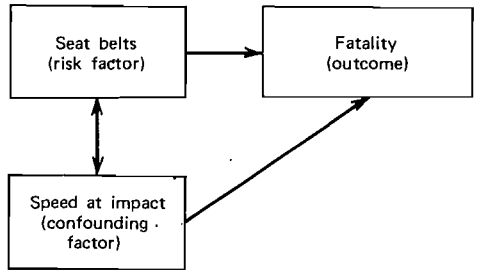

Figure 2.1 *Seat belts and fatalities.*

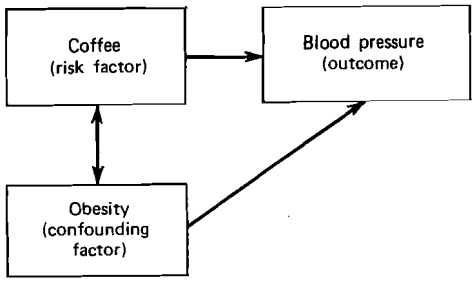

Figure 2.2 *Coffee drinking and blood pressure.*

between the risk factor and the outcome. The figures indicate the defining properties of a confounding factor: it is associated with the risk factor and it influences the outcome. As we have seen, the correct statistical analysis for both Examples 1.1 and 1.2 is to adjust for the effect of the confounding factor.

2.3.1 Unnecessary Adjustment

The following example, from MacMahon and Pugh (1970, p. 256), suggests that adjustment is not always called for.

Example 2.3 Oral contraceptives and thromboembolism: Consider an investigation of the effect of oral contraceptives on the risk of thromboembolism in women. A factor possibly associated with the risk factor (use of oral contraceptives) is religion. Catholic women may be less likely to use oral contraceptives than are other women. The relation between the three factors mentioned might be as shown in Figure 2.3. The cynic may add a second arrowhead to the arrow connecting "Religion" and "Oral contraceptive." As always, the relation between the risk factor (oral contraceptive use) and the outcome (thromboembolism) is of primary interest.

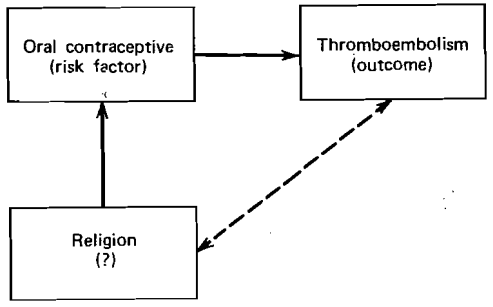

Figure 2.3 *Oral contraceptives and thromboembolisms.*

Table 2.4 Number of Thromboembolisms and Number of Women by Religion and Oral Contraceptive (OC) Use

	Catholic	Non-Catholic	Total
OC users	2000	5000	7000
(thromboembolisms)	(100)	(250)	(350)
Nonusers	8000	5000	13,000
(thromboembolisms)	(240)	(150)	(390)

To amplify the discussion, let us assume that the true lifetime risks of thromboembolism among users and nonusers of the contraceptive pill are 5% and 3%, respectively, irrespective of religion. Consider a study population consisting of 10,000 Catholic women and 10,000 non-Catholic women and suppose that 20% of the Catholics but 50% of the non-Catholics use oral contraceptives. Table 2.4 gives the number of women in each category of the study population and the number of these women who would suffer a thromboembolism if the rates of 5% and 3% were to apply.

In this example an analysis ignoring religion will give the correct risks ($350/7000 = 0.05$ and $390/13{,}000 = 0.03$), as should be clear from the construction of Table 2.4. However, the background factor of religion is apparently related not only to the risk factor—this we assumed at the start—but also to the outcome, as Table 2.5 demonstrates. The risk of thromboembolism is slightly higher among non-Catholics than among Catholics. Apparently, religion here satisfies the definition of a confounding factor, since it is a background factor associated with both the risk factor and the outcome.

Closer examination reveals that religion does not satisfy the definition. Although this background factor is associated with the outcome, it does not influence the outcome except through its effect on the risk factor. The dashed arrow in Figure 2.3 is a consequence of the other two arrows in the diagram.

If, nevertheless, the investigator does choose to correct for religion as a confounding factor using one of the techniques described in later chapters, he or she will not introduce bias into the study. Depending on the procedure chosen, there will be a slight or substantial loss of precision.

This last point applies more generally. Unnecessary adjustment—adjustment for a background factor that is not in fact confounding—will not introduce bias into a study except in some rather special circumstances, involving regression effects to be discussed in Section 5.3 (but note also Example 2.4). However, the precision of the estimated treatment effect may be reduced.

Table 2.5 Totals from Table 2.4

	Catholic	Non-Catholic
All women	10,000	10,000
Thromboembolisms	(340)	(400)

2.3.2 Proxy Variables

Before leaving Example 2.3, we should consider the possible effects of other important background variables. In fact, correction for the effect of religion will be useful if religion is associated with a confounding variable not measured in the study. Religion would then be called a *proxy variable*. This could happen, for example, in the following cases:

1. If risk of thromboembolism is affected by diet and the eating habits of Catholic and non-Catholic women differ. Diet would then be confounding, being related to both the risk factor (oral contraceptive use), through its relation to religion, and to the outcome (thromboembolism).
2. If risk of thromboembolism is affected by family size, and Catholic women had more children than did non-Catholic women. Here family size would be confounding for the same reason as diet in (1).

The investigator may choose to adjust for religion as a substitute for the unmeasured confounding factor. Unfortunately, the association between the proxy variable and the unmeasured confounding factor needs to be quite strong before the former can substitute effectively for the latter.

2.3.3 Defining the Factors

In some situations confusion over the definition of the risk factor can actually introduce bias into the study.

Example 2.4 Maternal age and infant mortality: Suppose that we want to determine the effect of maternal age on infant mortality. Birth weight might be considered as a confounding factor, as younger mothers have lower-weight babies and lower-weight babies have higher mortality. However, adjusting for birth weight in the analysis would be misleading, because we would be adjusting away the major difference we should be looking for. Birth weight in this example is a kind of intermediate outcome which leads to the final outcome of interest. Figure 2.4 summarizes the relationships among the three factors. If the effect of maternal age on infant mortality is entirely attributable to its effect on birth weight, an analysis adjusted for birth weight will indicate no association between maternal age and infant mortality.

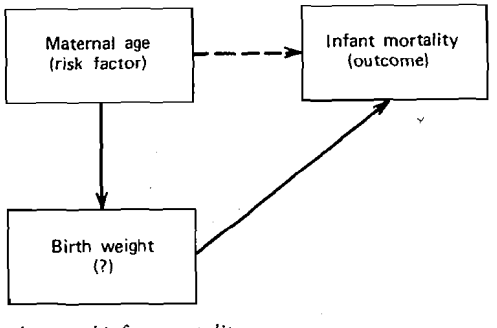

Figure 2.4 *Maternal age and infant mortality.*

Of course, it is possible that maternal age affects infant mortality through factors other than birth weight. Two infants of identical birth weight but whose mothers were of different ages would then be subject to different risks. An investigator interested in the effect of these other factors should adjust for birth weight. The new, adjusted estimate of the effect of the risk factor would differ from the unadjusted estimate, because the investigator's definition of the risk factor would be different.

Often the question of whether to adjust for a particular factor is not statistical but arises because the researcher has not defined with sufficient care the risk factor he or she wants to study.

APPENDIX 2A BIAS, PRECISION, AND MEAN SQUARED ERROR

Let θ denote the true value of the treatment effect and $\hat{\theta}$ the estimator of θ. The expectation symbol E denotes averaging with respect to the distribution of $\hat{\theta}$ in hypothetical repetitions. The bias, variance, and mean squared error (m.s.e.) of $\hat{\theta}$ are, respectively,

$$\text{bias}\,(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$$\text{var}\,(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 = E(\hat{\theta})^2 - [E(\hat{\theta})]^2$$

$$\text{m.s.e.}\,(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

On expanding the squared term in the last formula, we see that the cross-product term vanishes, and we obtain

$$\begin{aligned}\text{m.s.e.}\,(\hat{\theta}) &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \\ &= \text{var}\,(\hat{\theta}) + [\text{bias}\,(\hat{\theta})]^2.\end{aligned}$$

REFERENCE

MacMahon, B. and Pugh, T. F. (1970), *Epidemiology Principles and Methods*, Boston: Little, Brown.