CHAPTER 2

Statistical Introduction

2.1 DRAWING CONCLUSIONS FROM DATA

This chapter has two purposes. First, before considering difficulties in drawing inferences that are peculiar to observational studies, we will review some difficulties present to a greater or lesser degree, in all statistical studies, and will review standard techniques for handling them. This review is elementary and of the type given in introductory courses in statistics. Second, because of the limited amount of control that the investigator can exercise over data, a constant danger in observational studies is that estimates tend to be biased. Consequently, the last part of this chapter considers the effect of bias on a standard method of inference.

As a simple example, suppose that an investigator has a group of n persons exposed to some experience or causal force and a second group of n comparable persons who are not exposed to this force. After a suitable period of time, a measurement y of the response that is of interest is made on every person. Let y_{1j} and y_{2j} denote the values of y for the jth members of the first and second groups. As a measure of the effect of the exposure, the investigator takes $\bar{d} = \bar{y}_{1.} - \bar{y}_{2.}$, the difference between the means of the two groups. (The dot in the subscripts of $\bar{y}_{i.}$ indicates that we have averaged over the values of j.)

With human beings it is almost always found that even within a specific group, for example, the "exposed," the values of y_{1j} vary from person to person because of natural human variability. Consequently, $\bar{d}$ does not measure the effect of the exposure exactly, even in the simplest situation. Instead, $\bar{d}$ is more or less in error because of these fluctuations in the y_{1j} and y_{2j}. In handling this problem, the statistical approach starts by setting up a mathematical representation or model of the nature of the data.

First, we postulate that the n exposed people were drawn at random from a large population of exposed persons, and the n unexposed people from a large population of unexposed persons. Sometimes the data were actually obtained in this way. In comparing two types of workers A and B in a large factory, the investigator might have obtained a list of all the workers of types A and B in the factory. Then, with a table of random numbers, the investigator might have selected $n = 30$ persons of each type. Sometimes the data were obtained by random drawings from a small population that the investigator hopes is representative of a larger one. From a school system with 29 junior high schools, 16 might have been selected at random for the study, although the investigator's aim is to draw conclusions applicable to junior high schools across the nation. In many studies, however, there is no deliberate random drawing of the $2n$ people from larger populations. Some volunteering may be involved, or one of the two groups may be more or less a captive one. Some studies on hospital patients are based on those patients who are receiving treatment during the three weeks after the start of the study; the persons studied, for example, may be graduate students in psychology courses given by the investigator, or pupils in a school that agrees to cooperate with the investigator, or an unusual religious sect whose dietary habits interest a nutritionist.

Thus, in many studies the postulate is unrealistic that the data were drawn at random from specific populations. Nevertheless, the standard techniques of statistical inference, the best methods available at present for coping with this problem of person-to-person variability, apply only to the populations of which the data may be regarded as random samples. Consequently, the investigator who has a nonrandom sample must envision the kind of population from which the sample might be regarded as drawn at random. It is helpful to give a name—the *sampled* population—to this kind of population and to describe the ways in which it differs from the *target* population about which we would like to draw conclusions. From this we may be able to form some judgment about the ways in which these differences would alter the conclusions. These judgments are worth including in published reports, though they should be clearly labeled as such. This distinction between sampled and target populations will recur frequently throughout this book and is discussed further in Section 4.7.

To revert to the statistical analysis of the data on exposed and unexposed people, the simplest form of mathematical representation is as follows:

$$y_{1j} = \mu + \delta + e_{1j}; \qquad y_{2j} = \mu + e_{2j} \tag{2.1.1}$$

In this model, μ is a fixed parameter representing the average level of response in the unexposed population. The parameter δ stands for the

average effect of the exposure, therefore the mean of the first population is $\mu + \delta$. The quantities e_{1j} and e_{2j} are called random variables; they vary from one member of each population to another, and allow for the observed fact that the y_{1j} and the y_{2j} vary. It is assumed that over the respective populations the average values of the e_{1j} and the e_{2j} are both zero.

When we write down any mathematical model to be used as a basis for the analysis of data, it is essential to reflect on any assumptions implied by the model about the nature of our data. Analysis is likely to be misleading if derived from a model which makes erroneous assumptions. The simple model (2.1.1) implies an assumption that is at best dubious for most observational studies in practice. If δ is zero, it follows from (2.1.1) that y_{1j} and y_{2j} are drawn from populations having the same mean μ. In the simplest types of controlled experiment, the investigator often takes a step that is designed to ensure that this assumption holds. In order to form two samples of size n, one exposed and one unexposed, he first draws a sample of size $2n$ from a *single* population. He then divides this into two groups of size n by a process of randomization, usually from a table of random numbers. Since y_{1j} and y_{2j} initially come from the same population and differ only as the result of the randomization, this should guarantee the stated assumption.

In observational studies, however, exposed and unexposed groups are rarely found in this way. Nearly always, these groups are formed by forces beyond the investigator's control. People decide themselves whether to wear seat belts or to smoke; in the case of children, their parents decide whether to send them to public or private schools. Thus for observational studies a more realistic model is

$$y_{1j} = \mu_1 + \delta + e_{1j}; \qquad y_{2j} = \mu_2 + e_{2j} \tag{2.1.2}$$

with $E(e_{1j}) = E(e_{2j}) = 0$ as before (where the operator E represents the operation of taking the expected value), but μ_1 is not assumed to be equal to μ_2 since the investigator has been unable to take any step to ensure this equality.

The simplest estimate of δ is the difference between the means of the two samples; that is, $\bar{d} = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$. If (2.1.1) can be assumed, $\bar{d} = \delta + \bar{e}_{1\cdot} - \bar{e}_{2\cdot}$ and the mean value of $\bar{d}$ in repeated sampling is δ. But from (2.1.2),

$$\bar{d} = \delta + (\mu_1 - \mu_2) + \bar{e}_{1\cdot} - \bar{e}_{2\cdot}$$

and the expected value of $\bar{d}$ is $\delta + (\mu_{1\cdot} - \mu_2)$ instead of δ. We say that the estimate $\bar{d}$ is subject to a bias of amount $\mu_1 - \mu_2$. This indicates the reason for an interest in biased estimates in observational studies. Some illustrations of such sources of bias will be given in Section 2.4.

Reverting to the "no bias" situation with $\mu_1 = \mu_2$, even here $\bar{d}$ does not give us the correct answer δ, but an estimate of δ that is subject to an error of amount $\bar{e}_{1\cdot} - \bar{e}_{2\cdot}$. The best-known aids for answering the question "What can we say about δ?" are two techniques called the "test of significance" and the "construction of confidence intervals." The backgrounds of both techniques will be reviewed briefly.

2.2 TESTS OF SIGNIFICANCE

The test of significance relates to the question: Is there convincing evidence that exposure to the possible causal force has any effect at all? The question is not at all specific about the actual value of δ; it merely tries to distinguish between the verdict $\delta = 0$ and the verdict $\delta \neq 0$. The test requires some additional assumptions about the data. Suppose for simplicity that in their populations the e_{1j} and the e_{2j} both have the same standard deviation σ, though a test can be made without this assumption. If the e_{ij} are assumed to be normally and independently distributed, then theory says that $\bar{d}$ is normally distributed with population mean δ and standard deviation $\sigma\sqrt{2/n}$. The value of σ is not known, but an estimate s can be made from the pooled within-group mean square, where

$$s^2 = \frac{\sum(y_{1j} - \bar{y}_{1\cdot})^2 + \sum(y_{2j} - \bar{y}_{2\cdot})^2}{2(n-1)} \tag{2.2.1}$$

Furthermore, the quantity

$$\frac{\bar{d} - \delta}{s\sqrt{2/n}} \tag{2.2.2}$$

follows Student's t distribution with $2(n - 1)$ degrees of freedom. Moderate departures from normality in the data have little effect on this result.

If δ were zero, we would calculate from the data the quantity $t' = \bar{d}/s\sqrt{2/n}$, and find that (2.2.2) and the statement following it show that t' would follow the t distribution. From the tables of the t distribution for $2(n - 1)$ degrees of freedom, we calculate the probability P that a value of t as large as or larger than our calculated value t' would be obtained. If P is small enough, we argue that if δ were zero it is very unlikely that we would get a value of t as large as we did. We conclude that δ is not zero; that is, there was *some* effect of exposure. In practice, the most common dividing line between "small" and "not small" values of P is taken as $P = 0.05$, although no strong logical reason lies behind this choice. There is something to be said for reporting the actual value of P, particularly for the reader who wishes to summarize this investigation along with others, or wishes to form his own judgment as to whether δ differs from zero.

If P is not small, we have learned that a value of t as large as the observed value would quite frequently turn up if δ were zero. Such a result fails to provide convincing evidence in favor of the argument that δ differs from zero. Equally, the result by no means proves that δ *is* zero. Faced with the question "Is δ different from zero?" this result sits on the fence.

In practice, investigators use tests of significance in different ways. In some fields, the finding of a significant $\bar{d}$ has been regarded as necessary evidence which an investigator must produce to verify that an agent has an effect. This use has probably been beneficial in research. Some investigators with a wide fund of ideas have a fondness for making claims based on little solid work. The significance-level "yardstick" encourages them to produce firmer evidence if they want their claims to be recognized.

The finding of a nonsignificant $\bar{d}$, on the other hand, is often regarded as proof that $\delta = 0$. This conclusion has no logical foundation; it may have been suggested by the jargon "we accept the null hypothesis" often used in statistical teaching. If faced with a nonsignificant $\bar{d}$ in a test of an agent A, the investigator may decide to act as if the effect δ of A is zero or small, so that he drops any study of A and proceeds to some other agent that looks more promising. This decision, however, should be based on the investigator's judgment about the alternatives available. In fact, the probability that $\bar{d}$ is nonsignificant depends primarily on the smallness of the quantity $\delta\sqrt{n}/\sigma$, where n is the size of each sample and σ the standard deviation within each population. If we want to form a judgment about δ in the light of a nonsignificant $\bar{d}$, the values of n and σ are both relevant. Fortunately, this judgment is aided by the construction of confidence limits for δ, as discussed in Section 2.3.

There are two forms of the test of significance: the two-tailed and the one-tailed forms. In the *two-tailed* form, used in practice more frequently, we calculate the absolute value of t' (denoted $|t'|$), ignoring its sign. In the t table we look up the probability of getting a value of t greater than the observed t' in either direction; that is, in mathematical terms, the probability that $|t|$ exceeds the observed $|t'|$. (For the vertical bars read: "absolute value of".) The two-tailed test is appropriate when our initial judgment is that the true effect δ could be either positive or negative. In this event a value of δ far removed from zero can reveal itself either by making t' large and negative or by making t' large and positive.

The *one-tailed* test is appropriate* only when we know in advance what sign δ must have if it does not equal zero. In a teaching program designed to increase a child's knowledge of a certain subject, application of a one-tailed test implies that the program either increases the child's knowledge or has no effect; it cannot possibly decrease knowledge. When $\bar{d}$ is in the anticipated direction, the values of P in a one-tailed test are exactly half those in a two-tailed test having the same t'; therefore for given $\bar{d}$ a verdict that δ is unlikely to be zero can be reached for smaller sample sizes. When $\bar{d}$ is in the wrong direction, the conscientious user of a one-tailed test does not compute t'; rather the user concludes that the result does not justify rejection of the idea that δ is zero.

Some investigators misuse one-tailed tests. Before beginning the study, they are convinced that δ must be positive. If $\bar{d}$ is positive, they apply a one-tailed test as planned; if $\bar{d}$ is negative they apply a two-tailed test, having now recognized, perhaps reluctantly, that δ could be negative. The net effect is to make the actual significance probability level 1.5 times the announced significance level; for example, if the test is announced as being at the 5% level, it is actually at the 7.5% level.

2.3 CONFIDENCE INTERVALS

Confidence intervals relate to the question "How large is δ?" We have an estimate $\bar{d}$, but recognize that this will be more or less in error. Once again the t distribution is used. Let $t_{0.025}$ be the two-tailed 5% level of t for $2(n-1)$ degrees of freedom (d.f.). We know that $t = (\bar{d} - \delta)/s\sqrt{2/n}$ follows the t distribution when model (2.1.1) holds. Hence, unless our samples are of an unusual type that turns up only once in 20 times,

$$-t_{0.025} \leqslant \frac{\bar{d}-\delta}{s\sqrt{2/n}} \leqslant +t_{0.025}$$

Rearrangement gives

$$\bar{d} - t_{0.025}s\sqrt{2/n} \leqslant \delta \leqslant \bar{d} + t_{0.025}s\sqrt{2/n} \qquad (2.2.3)$$

*The editors do not regard Cochran's interpretation of the one-tailed test as the only appropriate one. In testing composite hypotheses, the null hypothesis might include both zero and losses, with the alternative of interest including all possible gains. We may be looking for gains from innovations and not be much interested in following up losses. In such circumstances, we think a one-tailed test is appropriate.

Thus, apart from an unlucky 1 in 20 chance that made our samples unusually dissimilar to the sampled populations, δ lies somewhere between the two limits in (2.2.3), called the 95% confidence limits. The width of the interval between the lower and the upper limits, $2\sqrt{2}\,t_{0.025}s/\sqrt{n}$, is a random quantity; its distribution depends on the variability in the populations, on the size of the samples, and on the confidence probability. The interval for 80% confidence probability is about $\frac{2}{3}$ as wide as the 95% interval and that for 50% probability is about $\frac{1}{3}$ as wide.

To summarize, our state of knowledge about the size of δ may be expressed by an estimate $\bar{d}$ and a pair of confidence limits within which δ is likely to lie, with an attached confidence probability indicating how likely. This probability is verifiable experimentally by setting up normal populations whose means differ by a known δ, drawing repeated pairs of samples, and computing the limits for a specified confidence probability, say 80%, at each draw. The values of the limits will vary from draw to draw, but about 80% of them will be found to enclose δ.

In published reports of studies, confidence limits are seldom stated explicitly. A more common practice is to give $\bar{d}$ and its standard error, $s_{\bar{d}} = \sqrt{2}\,s/\sqrt{n}$. The reader may then calculate his own limits by use of a t table. The number of degrees of freedom in $s_{\bar{d}}$ should also be given, but if they exceed 30, the 50%, 80%, and 95% limits are approximately $\bar{d} \pm 0.65s_{\bar{d}}$, $\bar{d} \pm 1.3s_{\bar{d}}$, and $\bar{d} \pm 2s_{\bar{d}}$, respectively.

These confidence limits also supply a two-tailed test of significance. If the 95% limits include 0, $\bar{d}$ is not significantly different from 0 at the 5% significance level. When a value of $\bar{d}$ is not significantly different from 0, it is worth examining the corresponding confidence limits. Sometimes, particularly with large samples, both limits are close to 0. For example, suppose that $\bar{d} = +0.3$, with 95% limits -0.6 and $+1.2$. In the context of the problem, it might be clear that even if δ is as large as 1.2, this is of minor practical importance. In this event the conclusion "exposure had no appreciable effect" would be justified as a practical approximation. On the other hand, with small samples and high variability, we might find that $\bar{d} = +0.3$ as before, but that the limits are -3.0 and $+3.6$. If values of δ as low as -3.0 and as high as $+3.6$ have important practical consequences of different kinds, a realistic conclusion is that the study did not succeed in delimiting the value of δ sufficiently to determine whether exposure has an important effect. In this event it would be risky to conclude that δ can be assumed to be 0.

To summarize, even under ideal conditions the information about the size of the effect supplied by a two-group study on human subjects is not as clear-cut as is desirable. But with the aid of techniques such as tests of significance and confidence intervals, and with careful thought about the implications of the results, we should be able to avoid serious mistakes in the conclusions. The logical ideas behind these techniques may be debated, but the techniques serve well if regarded as a guide to, and not as a substitute for, our thinking.

2.4 SYSTEMATIC DIFFERENCES BETWEEN THE POPULATIONS

When comparing samples from exposed and unexposed populations in an observational study, a frequent source of misleading conclusions, as mentioned in Section 2.1, is that the two populations differ systematically with respect to other characteristics or variables that affect the response variable y. This section illustrates a few of the many situations that occur. The first example comes from the large-sample studies of the relation between smoking and death rates of men (Cochran, 1968). In these studies, information about smoking habits (including type of smoking and amount smoked per day) was first obtained by a mail questionnaire to a large sample of men. The three types considered here are nonsmokers, smokers of cigarettes only, and smokers of cigars and/or pipes (the cigar and pipe smokers were combined because the sample numbers are smaller). After receipt of the questionnaires the investigators were notified of any deaths that occurred among the sample members in subsequent months. From the data supplied to the U.S. Surgeon General's Committee on Smoking and Health, Table 2.4.1 shows the death rates for the three groups of men in a Canadian, a British, and a U.S. study.

To take the figures at face value, it looks as if cigar or pipe smoking results in a substantial increase in death rates, the differences from the nonsmokers being statistically significant in all three studies. For cigarette smokers, the British study shows an elevated death rate, significant at just about the 5% level, but the Canadian and U.S. studies show no elevations of this magnitude.

Table 2.4.1. Death Rates per 1,000 Person-Years

Smoking Group	Canadian (6 years)	British (5 years)	United States (20 months)
Nonsmokers	20.2	11.3	13.5
Cigarettes only	20.5	14.1	13.5
Cigars and/or pipes	35.5	20.7	17.4

We remember, however, that the groups being compared are self-selected. Before rushing out to warn the cigar and pipe smokers we should ask ourselves whether there are other characteristics affecting death rates or variables in which these groups might differ systematically. For men under 40, I think it is correct to say that there is no such variable known to have a *major* effect on death rates. For older men, age is a variable that becomes of predominating importance. The death rate rises a lot as age increases, with a steadily increasing steepness of slope. An obvious precaution is to examine the mean ages of the men in each group, as shown in Table 2.4.2.

In all three studies, the cigar or pipe smokers are older on the average than the men in the other groups. This is not surprising, since cigar and pipe smoking are more frequent among older men. In both the Canadian and U.S. studies, which showed death rates about the same for cigarette smokers and nonsmokers, the cigarette smokers are younger than the nonsmokers. Clearly, no conclusion should be made about the relation between smoking and death rates without taking steps to try to remove the effects of these systematic differences in ages among the groups. The investigators who directed these studies were well aware of this problem and planned from the beginning to handle it. The available procedures for dealing with disturbing variables of this type in the statistical analysis are discussed in Chapters 5 and 6.

The possibility of biases of this type is now widely recognized whenever groups are self-selected. Suppose a television station has a one-hour adult educational program and wishes to measure how much the viewers of this program have learned from it. The station maintains a representative list of viewers of its programs. After the program the station invites a random sample of viewers, some of whom have seen this program and some who have not, to take an examination intended to reveal what has been learned from the program. Even if all those invited conscientiously take the examination, the station recognizes that the computed $\bar{d}$ almost certainly overestimates the effect of the program, because those who chose to view this program were probably more interested in the subject and better informed about it from the beginning, and would do better in the examina-

Table 2.4.2. Mean Ages (in Years) of Men in Each Group

Smoking Group	Canadian	British	United States
Nonsmokers	54.9	49.1	57.0
Cigarettes only	50.5	49.8	53.2
Cigars and/or pipes	65.9	55.7	59.7

tion even if they had not viewed this program. In fact, the station would regard the problem of circumventing this overestimation as the major obstacle in conducting the study.

In a second example the effect of self-selection is less clear-cut. Several studies have compared the health and well-being of families who move from slum housing into new public housing, with those of families who remain in slum housing. The objective is to measure the supposed beneficial effects of public housing. However, in order to become eligible for public housing, the parents of a family may have to possess both initiative and some determination in dealing with a bureaucracy. One might argue that these individuals may already possess a desire to rise in the world that might show up in better health and well-being when the response measurements are made. Insofar as such effects are present, they would increase $\bar{d}$ and the unwary investigator may attribute this to the beneficial effects of public housing.

This example illustrates another aspect of the problem. While admitting the argument in the preceding paragraph, an investigator might retort, "Why should this overestimation be sufficiently large to seriously distort the conclusions? Why should this initiative and determination have much effect on the number of colds Johnny catches next winter?" A critic might reply that such parents have high aspirations for their children and may take better preventive medical care of the children than do the slum parents in this study. The point is that in many studies, sources of potential bias of this type are either unavoidable or overlooked until too late and that we can only guess about the size of bias that is created. In such cases our judgment about the direction and amount of bias, even if slender, is worthwhile when conclusions are being drawn. Incidentally, in a Baltimore public housing study, Wilner et al. (1955) neatly attempted to avoid the bias in question by noticing that the list of processed and eligible applicants for public housing was much greater than the available space. Consequently, both the "public housing" sample and the "slum housing" sample were drawn from this list. The initial "slum" sample was made larger than the "public housing" sample, because it was known that as housing became available some of the "slum" families would move into public housing during the course of the study.

Sometimes systematic differences between the exposed and unexposed groups are introduced in the process of measuring the responses, especially when these measurements involve an element of human judgment. A consulting statistician soon observes that some investigators become emotionally interested in the causal force they are studying and they want the force to show some effects. This is not said in deprecation. It is the business of the research worker to form pictures of what the world is like, and an imaginative interest in one's pictures is conducive to good research. Consequently, the investigator, when measuring the responses, may find that the levels of measurement may change unconsciously when moving from the unexposed to the exposed group.

This danger is widely recognized in medical studies of the progress of patients. I once watched an expert on leprosy for an hour while he examined a patient in meticulous detail in order to measure the patient's progress during the preceding two-months' treatment in an experiment on leprosy. The examination was complex, since bacteriological, neurological, and dermatological symptoms are all involved. At the end of the exam, the doctor dismissed the patient and said, "I rate this patient *Much Improved*"—this being the highest category of improvement that the scale allowed. At this point a blabbermouth at the back of the room, who had the code sheet, said "You'll be interested to know, doctor, that the patient was on placebo"—a placebo being an inert drug with no bactericidal effect on leprosy, intended only as a measure of comparison for the real treatments. The immediate retort of the doctor was "I would never have rated that patient 'Much Improved' if I had known he was on placebo. Call him back." There was silence for about 30 seconds; members of the planning team either stared at the expert or the blabbermouth in mixtures of sorrow and anger, or engaged in silent prayer. Finally, the expert said in a low voice, as if arguing with himself: "No. Let it stand."

For this reason a standard precaution in medical studies, as in this one, is that the doctor who is measuring the response variable should not be informed of the treatment the patient is receiving. In some studies no workable precaution may occur to the investigators. In studies on the possible inheritance of some form of cancer, cancer patients may be better informed about cancer in their relatives than are noncancer controls, and hence report more relatives with cancer. In a study of the inheritance of neuroticism it was proposed to enlist both neurotic and nonneurotic subjects and obtain information about the parental generation by questioning each group of subjects about neuroticism in their parents. Bradford Hill (1953) remarks: "What the adult neurotic thinks of his father may not always be the truth."

2.5 THE MODEL WHEN BIAS IS PRESENT

When systematic differences between the exposed and unexposed groups are present, the simplest model appears to be

$$y_{ij} = \mu_1 + \delta + e_{1j}; \qquad y_{2j} = \mu_2 + e_{2j}$$

giving

$$\bar{d} = \delta + (\mu_1 - \mu_2) + \bar{e}_{1\cdot} - \bar{e}_{2\cdot} = \delta + B + \bar{e}_{1\cdot} - \bar{e}_{2\cdot}$$

where $B = \mu_1 - \mu_2$ represents the amount of bias in the estimate $\bar{d}$. The change from the "no bias" situation is that $\bar{d}$ now estimates $\delta + B$. While this is obvious, it emphasizes a point that investigators or their critics sometimes overlook. Once the presence of bias is admitted, they sometimes take a morbid view of the situation, implying that nothing can be learned about δ from $\bar{d}$. Actually, since we can rarely be certain in observational studies that estimates are completely free from bias, a not unreasonable view is that *all* estimates are biased to some extent in observational studies. The problem is to keep the bias small enough so that we are not seriously led astray in our conclusions.

We shall examine the effect of bias on the probability that $\bar{d}$ lies within the interval $(\delta - L, \delta + L)$, in other words, that $\bar{d}$ is correct to within given limits $\pm L$. With no bias, $\bar{d}$ is assumed approximately normally distributed with mean δ and standard deviation $\sigma_{\bar{d}}$. Hence, the quantity $(\bar{d} - \delta)/\sigma_{\bar{d}}$ is a standard normal deviate, and we calculate the probability α that $\bar{d} - \delta$ lies within $\pm L$ by setting $z = L/\sigma_{\bar{d}}$ and locating in the normal tables the probability α that a normal deviate z lies within the limits $\pm z_{\alpha/2}$.

When the mean of $\bar{d}$ is $\delta + B$, on the other hand, the normal deviate is $(\bar{d} - \delta - B)/\sigma_{\bar{d}}$. This equals $(-L - B)/\sigma_{\bar{d}}$ when $\bar{d} - \delta = -L$, and $(L - B)/\sigma_{\bar{d}}$ when $\bar{d} - \delta = +L$. If $B = fL$, these limits become $-L(1+f)/\sigma_{\bar{d}}$ and $+L(1-f)/\sigma_{\bar{d}}$ or $-z_{\alpha/2}(1+f)$ and $+z_{\alpha/2}(1-f)$. For a given probability in the "no bias" case, $z_{\alpha/2}$ is known; for example, $z_{\alpha/2} = 1.96$ for $1 - \alpha = 0.95$. Thus, given f and the "no bias" probability, we can read from the normal tables the corresponding probability when bias is present.

Table 2.5.1 shows the probabilities that $\bar{d}$ lies within $(\delta - L, \delta + L)$ for $P = 0.99, 0.95, 0.90, 0.80, 0.70$, and 0.60 and $f = B/L$ running from 0.1 to 1.0, by intervals of 0.1, and also $f = 1.5$ and 2.0. If f is less than 0.2, the reduction in the "no bias" probability is trivial. Even for $f = 0.5$, the reduction remains moderate, for example, from 0.95 to 0.84 and from 0.80 to 0.71. When $f = 1.0$, however, all the probabilities are reduced to 0.5 or less, and decrease steadily for higher f.

To put it another way, the effect of a bias of amount B cannot make the probability that $\bar{d}$ is correct to within $\pm B$ more than $\frac{1}{2}$, no matter how large the sample is. However, the probability that $\bar{d}$ is correct to $\pm 2B$ is decreased only moderately by the bias; the probability that $\bar{d}$ is correct to $\pm 5B$ is decreased only trivially. Some investigators prefer to express quantities like B and L as percentages of δ. In these terms, a 15% bias makes the probability that $\bar{d}$ is correct to within 15% of δ at most $\frac{1}{2}$, but reduces only moderately the probability that $\bar{d}$ is correct to within $\pm 30\%$.

For a given value of f, we might expect that a higher "no bias" probability will result in a correspondingly higher probability when bias is present. Table 2.5.1 shows this happens when $f \leqslant 1$, but for $f > 1$ the probabilities go in the opposite direction when bias is present. With $f = 1.5$, for instance, a "no bias" P of 0.99 is reduced to $P = 0.10$, but a "no bias" P of 0.60 is reduced only to 0.32. The explanation is that $\bar{d}$ in the biased case is an unbiased estimate of $\delta + B$, which lies outside the limits $\delta \pm L$. As the random-sampling error $\sigma_{\bar{d}}$ decreases in this case, we are doing a better job of estimating the wrong quantity $\delta + B$, but a poorer job of estimating δ. With $f = 2$, $\bar{d}$ is an unbiased estimate of $\delta + 2L$, and can fall in the desired interval $\delta \pm L$ only by making a negative error of more than L in estimating $\delta + 2L$. Thus with $f = 2$, the probability that $\bar{d}$ lies in the desired interval is $(1 - P)/2$, which for $P = 0.95$ gives 0.025.

The lessons from this example are important. Even without bias, other sources of variability impose a limit on the accuracy that can be attained with high probability in observational studies. These probabilities are not drastically reduced by bias, provided that B is substantially less than the limit of error L that can be tolerated. If an investigator takes pains to remove the effects of suspected sources of bias, the effect of undetected bias may be to reduce a presumed 95% probability of lying within prescribed limits to something like 60 or 70%. In such cases the deleterious effect of

Table 2.5.1. Effect of a Bias of Amount $B = fL$ on the Probability P that $\bar{d}$ Lies within Limits $(\delta - L, \delta + L)$

			Probability (P)			
No Bias	0.99	0.95	0.90	0.80	0.70	0.60
$f = 0.1$	0.99	0.95	0.90	0.80	0.70	0.60
0.2	0.98	0.93	0.88	0.78	0.69	0.59
0.3	0.96	0.91	0.86	0.77	0.68	0.58
0.4	0.94	0.88	0.83	0.74	0.66	0.57
0.5	0.90	0.84	0.79	0.71	0.64	0.56
0.6	0.85	0.78	0.74	0.68	0.61	0.54
0.7	0.78	0.72	0.69	0.64	0.58	0.52
0.8	0.70	0.65	0.63	0.59	0.55	0.50
0.9	0.60	0.58	0.56	0.54	0.52	0.48
1.0	0.50	0.50	0.50	0.49	0.48	0.45
1.5	0.10	0.16	0.21	0.26	0.30	0.32
2.0	0.005	0.025	0.05	0.10	0.15	0.20

bias is not that it makes the results completely wrong, but that we do not know how far the results can be trusted. The most misleading situation is that of a relatively large bias when the samples are large. In this case $\bar{d}$ is likely to have a small standard error, so that the 95% confidence interval is narrow and we congratulate ourselves on our accurate results. The actual probability that δ lies within these limits may, however, be tiny, as the results for $f = 2$ indicate.

This example is also relevant when we come later to study the techniques for removing *suspected* bias in observational studies. There is evidence that in many circumstances the available methods are not fully effective. They remove some, hopefully most, of the bias, but leave a residual part. Our hope is that for this residual part, B/σ is substantially less than the value of L which makes our results useful.

On occasion it helps to think in terms of B/δ or of $B/\bar{d}$. Suppose we have been unable to remove or reduce a specific source of bias, and are making speculative calculations as to how large a bias from this source can be. We can sometimes reach a firm judgment that even under the most unfavorable circumstances, B/δ is bound to be small. This seems to be the situation in prospective studies of the relation between heavy cigarette smoking and the death rate from lung cancer. Cigarette smokers are self-selected, and numerous possible sources of bias in comparing them with nonsmokers have been mentioned in the literature. But the increase in the lung-cancer death rate for heavy cigarette smokers versus nonsmokers is so large that estimates of the bias, often admittedly speculative, all seem to result in relatively small changes in the estimated δ.

The preceding discussion has dealt with the effect of bias on attempts to estimate δ correct to within limits $\pm L$. It is also worth considering the effect of bias on a test of significance of the null hypothesis $\delta = 0$ in relation to sample size. If we can assume, that apart from the bias, we have independent random samples from the two populations, then $\sigma_{\bar{d}} = \sqrt{2}\,\sigma/\sqrt{n}$. Consider first a two-tailed test. The type-I error is the probability that $\bar{d}$ lies outside the limits $(-\sqrt{2}\,\sigma z_{\alpha/2}/\sqrt{n}, +\sqrt{2}\,\sigma z_{\alpha/2}/\sqrt{n})$. With a bias of amount B, this is easily found to be the probability that a normal deviate z lies outside the limits $(-z_{\alpha/2} - B\sqrt{n}/\sqrt{2}\,\sigma, z_{\alpha/2} - B\sqrt{n}/\sqrt{2}\,\sigma)$. Suppose that $B\sqrt{n}/\sqrt{2}\,\sigma = \pm 0.5$ in a 5% test, for which $z_{\alpha/2} = 1.96$. The limits are $(-2.46, 1.46)$, or $(-1.46, 2.46)$, and the probability of type-I error P is 0.079. With $B\sqrt{n}/\sqrt{2}\,\sigma = \pm 1$, $P = 0.170$, and with $B\sqrt{n}/\sqrt{2}\,\sigma = \pm 2$, $P = 0.516$. Regardless of the sign of the bias, the type-I error of a two-tailed test is increased by bias. If n is large, the type-I error is so distorted that the test becomes meaningless.

While I can give only an unsubstantiated opinion, things may not be this bad in practice. Careful precautions against bias might reduce B/σ to say 0.05. For two samples of sizes 50, 100, and 200, the P's for the type-I errors become 0.057, 0.064, and 0.079, respectively; only for samples larger than 200 does the distortion become intolerable.

The effect of bias on a one-sided test depends of course on the direction of the bias in relation to the direction of the one-sided test. If these directions are the same, that is, if the alternative hypothesis specifies $\delta > 0$ and if $B > 0$, bias increases the type-I error still more rapidly than it does for two-tailed tests. Consider a one-tailed 5% test that assumes $\delta \geqslant 0$, with $z_\alpha = 1.645$. For $B\sqrt{n}/\sqrt{2}\,\sigma = 0.5$, 1, and 2 (the figures given in the two-tailed example), the P's for the type-I errors become 0.126, 0.260, and 0.639, respectively. On the other hand, if B is negative in this situation, the type-I error is decreased. In fact, if we are sure that $\delta \geqslant 0$ and if $B\sqrt{n}/\sqrt{2}\,\sigma$ is negative and sufficiently large, we might detect the presence of bias by noting that our $\bar{d}$ would be significant in the wrong direction, a logical contradiction of the notion that $\delta \geqslant 0$, unless negative bias is the explanation.

Admitting the inevitability of bias and the difficulty of securing random samples in observational studies, some writers argue that the test of significance is useless for guidance in such studies. In large-sample studies this can be so, only because any difference large enough to be of interest is certain to be significant by a standard test that ignores bias. One frequently finds no mention of tests in the discussion of the results of such studies. Tests can also be useless if our results are subject to biases that are completely unknown in size and direction. However, as Kish (1959) points out, such biases would render ineffective *any* attempt to draw conclusions from observational studies, not merely tests of significance.

The positive attitude toward this problem is to exercise precautions against bias in the planning and analysis of observational studies, in the hope that the remaining bias will not greatly disturb type-I errors or confidence probabilities. In the interpretation of tests of significance, whether the verdict is significant or nonsignificant, any judgment that we can form about the direction and size of the remaining bias is clearly relevant.

2.6 SUMMARY

This chapter discusses some of the difficulties in trying to draw sound conclusions from the results of a study. Even in the simplest type of study—a comparison of a group of people exposed to some causal force and a second group not so exposed—the study provides only an estimate of the average effect of the causal force that is subject to error. In the interpretation of this estimated difference, two major statistical aids are *tests of significance* and *confidence limits*.

The test of significance relates to the question of whether there is convincing evidence of some real effect. Confidence limits supply estimated upper and lower bounds to the size of the effect. The finding of a nonsignificant difference by no means proves that there was no real effect. It merely reports that a difference as large as that observed could have occurred with nonnegligible probability without the presence of a real effect different from zero. The interpretation of a nonsignificant difference is often helped by calculating confidence limits. With small sample sizes or a highly variable population, the confidence limits may show that the study failed to measure the size of the effect accurately enough for competent decisions. The soundest conclusion is that more work on the problem is needed.

Further difficulties are often present. Tests of significance and confidence limits apply only to the population of which the data are a random sample. This population—the sampled population—is often difficult to describe accurately and may differ in one or more respects from the target population to which the investigator wants his conclusions to apply. A useful part of the investigator's summary of results includes a statement of known differences between sampled and target populations and a judgment about the extent to which these differences may affect the conclusions as they apply to the target population.

In observational studies the exposed and nonexposed populations are usually created by forces beyond the control of the investigator, and may differ systematically in respects other than that of exposure. The consequence is that the sample mean difference $\bar{d}$ does not estimate the real effect of exposure, but $\delta + B$, where B is a bias term caused by these other systematic differences. The presence of bias decreases the probability that $\bar{d}$ as an estimate of δ is correct within given limits $\pm L$. This reduction in probability is moderate if B/L is less than 0.5, as, for example, $B/\delta = 10\%$ and limits of accuracy $L/\delta = \pm 20\%$ are sufficient for our purpose. This result is encouraging. In statistical studies it is hard to ensure that bias has been completely eliminated, particularly in observational studies, but measures taken to control bias in planning and analysis may reduce B/L to a value that is small enough for practical decisions.

Similarly, a bias in either direction increases the probability of a type-I error in a two-tailed test of significance, sometimes to an extent that makes the test meaningless in large samples. This fact emphasizes the importance of exercising precautions against bias in the planning and analysis of observational studies and of using any judgments about the direction and

size of the remaining bias in the interpretation of the results of tests of significance.

REFERENCES

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 295–313 [Collected Works #90].

Hill, A. B. (1953). Observation and experiment. *New Engl. J. Med.*, **248**, 995–1001.

Kish, L. (1959). Some statistical problems of research design. *Am. Sociological Rev.*, **24**, 328–338.

Wilner, D. M., R. P. Walkley, and S. W. Cook (1955). *Human Relations in Interracial Housing.* University of Minnesota Press, Minneapolis.