2

Randomized Experiments

2.1 Introduction and Example: A Randomized Clinical Trial

Observational studies and controlled experiments have the same goal, inference about treatment effects, but random assignment of treatments is present only in experiments. This chapter reviews the role of randomization in experiments, and so prepares for discussion of observational studies in later chapters. A theory of observational studies must have a clear view of the role of randomization, so it can have an equally clear view of the consequences of its absence. Sections 2.1 and 2.2 give two examples: a large controlled clinical trial, and then a small but famous example due to Sir Ronald Fisher, who is usually credited with the invention of randomization, which he called the "reasoned basis for inference" in experiments. Later sections discuss the meaning of this phrase, that is, the link between randomization and statistical methods. Most of the material in this chapter is quite old.

Randomized Trial of Coronary Surgery

The US Veterans Administration (Murphy et al. 1977) conducted a randomized controlled experiment comparing coronary artery bypass surgery with medical therapy as treatments for coronary artery disease. Bypass surgery is an attempt to repair the arteries that supply blood to the heart, arteries that have been narrowed by fatty deposits. In bypass surgery, a

TABLE 2.1. Base-Line Comparison of Coronary Patients in the Veterans Administration Randomized Trial.

Covariate	Medical %	Surgical %
New York Heart Association Class II & III	94.2	95.4
History of myocardial infarction (MI)	59.3	64.0
Definite or possible MI based on electrocardiogram	36.1	40.5
Duration of chest pain > 25 months	50.0	51.8
History of hypertension	30.0	27.6
History of congestive heart failure	8.4	5.2
History of cerebral vascular episode	3.2	2.1
History of diabetes	12.9	12.2
Cardiothoracic ratio > 0.49	10.4	12.2
Serum cholesterol > 249 mg/100 ml	31.6	20.6

bypass or bridge is formed around a blockage in a coronary artery. In contrast, medical therapy uses drugs to enhance the flow of blood through narrowed arteries. The study involved randomly assigning 596 patients at 13 Veterans Administration hospitals, of whom 286 received surgery and 310 received drug treatments. The random assignment of a treatment for each patient was determined by a central office after the patient had been admitted into the trial.

Table 2.1 is taken from their study. It compares the medical and surgical treatment groups in terms of 10 important characteristics of patients measured at "base-line," that is, prior to the start of treatment. A variable measured prior to the start of treatment is called a *covariate*. Similar tables appear in reports of most clinical trials.

Table 2.1 shows the two groups of patients were similar in many important ways prior to the start of treatment, so that comparable groups were being compared. When the percentages for medical and surgical are compared, the difference is not significant at the 0.05 level for nine of the variables in Table 2.1, but is significant for serum cholesterol. This is in line with what one would expect from 10 significance tests if the only dif-

ferences were due to chance, that is, due to the choice of random numbers used in assigning treatments.

For us, Table 2.1 is important for two reasons. First, it is an example showing that randomization tends to produce relatively comparable or balanced treatment groups in large experiments. The second point is separate and more important. The 10 covariates in Table 2.1 were not used in assigning treatments. There was no deliberate balancing of these variables. Rather the balance we see was produced by the random assignment, which made no use of the variables themselves. This gives us some reason to hope and expect that other variables, not measured, are similarly balanced. Indeed, as shown shortly, statistical theory supports this expectation. Had the trial not used random assignment, had it instead assigned patients one at a time to balance these 10 covariates, then the balance might well have been better than in Table 2.1, but there would have been no basis for expecting other unmeasured variables to be similarly balanced.

The VA study compared survival in the two groups three years after treatment. Survival in the medical group was 87% and in the surgical group 88%, both with a standard error of 2%. The 1% difference in mortality was not significant. Evidently, when comparable groups of patients received medical and surgical treatment at the VA hospitals, the outcomes were quite similar.

The statement that randomization tends to balance covariates is at best imprecise; taken too literally, it is misleading. For instance, in Table 2.1, the groups do differ slightly in terms of serum cholesterol. Presumably there are other variables, not measured, exhibiting imbalances similar to if not greater than that for serum cholesterol. What is precisely true is that random assignment of treatments can produce some imbalances by chance, but common statistical methods, properly used, suffice to address the uncertainty introduced by these chance imbalances. To this subject, we now turn.

2.2 The Lady Tasting Tea

"A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup," or so begins the second chapter of Sir Ronald Fisher's (1935, 1949) book *The Design of Experiments*, which introduced the formal properties of randomization. This example is part of the tradition of statistics, and in addition it was well selected by Fisher to illustrate key points. He continues:

> Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order. The subject has been told in advance of what the test will consist, namely that she

> will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such a manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

Fisher then asks what would be expected if the Lady was "without any faculty of discrimination," that is, if she makes no changes at all in her judgments in response to changes in the order in which tea and milk are added to the cups. To change her judgments is to have some faculty of discrimination, however slight. So suppose for the moment that she cannot discriminate at all, that she gives the same judgments no matter which four cups receive milk first. Then it is only by accident or chance that she correctly identifies the four cups in which milk was added first. Since there are $\binom{8}{4} = 70$ possible divisions of the eight cups into two groups of four, and randomization has ensured that these are equally probable, the chance of this accident is $1/70$. In other words, the probability that the random ordering of the cups will yield perfect agreement with the Lady's fixed judgments is $1/70$. If the Lady correctly classified the cups, this probability, $0.014 = 1/70$, is the significance level for testing the null hypothesis that she is without the ability to discriminate.

Fisher goes on to describe randomization as the "reasoned basis" for inference and "the physical basis of the validity of the test"; indeed, these phrases appear in section headings and are clearly important to Fisher. He explains:

> We have now to examine the physical conditions of the experimental technique needed to justify the assumption that, if discrimination of the kind under test is absent, the result of the experiment will be wholly governed by the laws of chance ... .
>
> It is [not sufficient] to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation ... .
>
> The element in the experimental procedure which contains the essential safeguard is that the two modifications of the test beverage are to be prepared "in random order." This, in fact, is the only point in the experimental procedure in which the laws of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced.

Fisher discusses this example for 15 pages, though its formal aspects are elementary and occupy only a part of a paragraph. He is determined to establish that randomization has justified or grounded a particular inference, formed its "reasoned basis," a basis that would be lacking had the same pattern of responses, the same data, been observed in the absence of randomization.

The example serves Fisher's purpose well. The Lady is not a sample from a population of Ladies, and even if one could imagine that she was, there is but one Lady in the experiment and the hypothesis concerns her alone. Her eight judgments are not independent observations, not least because the rules require a split into four and four. Later cups differ from earlier ones, for by cup number five, the Lady has surely tasted one with milk first and one with tea first. There is no way to construe, or perhaps misconstrue, the data from this experiment as a sample from a population, or as a series of independent and identical replicates. And yet, Fisher's inference is justified, because the only probability distribution used in the inference is the one created by the experimenter.

What are the key elements in Fisher's argument? First, experiments do not require, indeed cannot reasonably require, that experimental units be homogeneous, without variability in their responses. Homogeneous experimental units are not a realistic description of factory operations, hospital patients, agricultural fields. Second, experiments do not require, indeed, cannot reasonably require, that experimental units be a random sample from a population of units. Random samples of experimental units are not the reality of the industrial laboratory, the clinical trial, or the agricultural experiment. Third, for valid inferences about the effects of a treatment on the units included in an experiment, it is sufficient to require that treatments be allocated at random to experimental units—these units may be both heterogeneous in their responses and not a sample from a population. Fourth, probability enters the experiment only through the random assignment of treatments, a process controlled by the experimenter. A quantity that is not affected by the random assignment of treatments is a fixed quantity describing the units in the experiment.

The next section repeats Fisher's argument in more general terms.

2.3 Randomized Experiments

2.3.1 Units and Treatment Assignments

There are N units available for experimentation. A unit is an opportunity to apply or withhold the treatment. Often, a unit is a person who will receive either the treatment or the control as determined by the experimenter. However, it may happen that it is not possible to assign a treatment to a single person, so a group of people form a single unit, perhaps all children

in a particular classroom or school. On the other hand, a single person may present several opportunities to apply different treatments, in which case each opportunity is a unit; see Problem 2. For instance, in §2.2, the one Lady yielded eight units.

The N units are divided into S strata or subclasses on the basis of covariates, that is, on the basis of characteristics measured prior to the assignment of treatments. The stratum to which a unit belongs is not affected by the treatment, since the strata are formed prior to treatment. There are n_s units in stratum s for $s = 1, \ldots, S$, so $N = \sum n_s$.

Write $Z_{si} = 1$ if the ith unit in stratum s receives the treatment and write $Z_{si} = 0$ if this unit receives the control. Write m_s for the number of treated units in stratum s, so $m_s = \sum_{i=1}^{n_s} Z_{si}$, and $0 \leq m_s \leq n_s$. Finally, write $\mathbf{Z}$ for the N-dimensional column vector containing the Z_{si} for all units in the lexical order; that is,

$$\mathbf{Z} = \begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1,n_1} \\ Z_{21} \\ \vdots \\ Z_{S,n_S} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_S \end{bmatrix}, \qquad \text{where} \quad \mathbf{Z}_s = \begin{bmatrix} Z_{s1} \\ \vdots \\ Z_{s,n_s} \end{bmatrix}. \tag{2.1}$$

This notation covers several common situations. If no covariates are used to divide the units, then there is a single stratum containing all units, so $S = 1$. If $n_s = 2$ and $m_s = 1$ for $s = 1, \ldots, S$, then there are S pairs of units matched on the basis of covariates, each pair containing one treated unit and one control. The situation in which $n_s \geq 2$ and $m_s = 1$ for $s = 1, \ldots, S$ is called matching with multiple controls. In this case there are S matched sets, each containing one treated unit and one or more controls.

The case of a single stratum, that is $S = 1$, is sufficiently common and important to justify slight modifications in notation. When there is only a single stratum the subscript s is dropped, so Z_i is written in place of Z_{1i}. The same convention applies to other quantities that have subscripts s and i.

2.3.2 Several Methods of Assigning Treatments at Random

In a *randomized experiment*, the experimenter determines the assignment of treatments to units, that is the value of $\mathbf{Z}$, using a known random mechanism such as a table of random numbers. To say that the mechanism is known is to say that the distribution of the random variable $\mathbf{Z}$ is known because it was created by the experimenter. One requirement is placed on this random mechanism, namely, that, before treatments are assigned,

every unit has a nonzero chance of receiving both the treatment and the control, or formally that $0 < \text{prob}(Z_{si} = 1) < 1$ for $s = 1, \ldots, S$ and $i = 1, \ldots, n_s$. Write Ω_0 for the set containing all possible values of $\mathbf{Z}$, that is, all values of $\mathbf{Z}$ which are given nonzero probability by the mechanism.

In practice, many different random mechanisms have been used to determine $\mathbf{Z}$. The simplest assigns treatments independently to different units, taking $\text{prob}(Z_{si} = 1) = 1/2$ for all s, i. This method was used in the Veterans Administration experiment on coronary artery surgery in §2.1. In this case, Ω_0 is the set containing 2^N possible values of $\mathbf{Z}$, namely, all N-tuples of zeros and ones, and every assignment in Ω_0 has the same probability; that is, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/2^N$ for all $\mathbf{z} \in \Omega_0$. The number of elements in a set A is written $|A|$, so in this case $|\Omega_0| = 2^N$. This mechanism has the peculiar property that there is a nonzero probability that all units will be assigned to the same treatment, though this probability is extremely small when N is moderately large. From a practical point of view, a more important problem with this mechanism arises when S is fairly large compared to N. In this case, the mechanism may give a high probability to the set of treatment assignments in which all units in some stratum receive the same treatment. If the strata were types of patients in a clinical trial, this would mean that all patients of some type received the same treatment. If the strata were schools in an educational experiment, it would mean that all children in some school received the same treatment. Other assignment mechanisms avoid this possibility.

The most commonly used assignment mechanism fixes the number m_s of treated subjects in stratum s. In other words, the only assignments $\mathbf{Z}$ with nonzero probability are those with m_s treated subjects in stratum s for $s = 1, \ldots, S$. If m_s is chosen sensibly, this avoids the problem mentioned in the previous paragraph. For instance, if n_s is required to be even and m_s is required to equal $n_s/2$ for each s, then half the units in each stratum receive the treatment and half receive the control, so the final treated and control groups are exactly balanced in the sense that they contain the same number of units from each stratum.

When m_s is fixed in this way, let Ω be the set containing the $K = \prod_{s=1}^{S} \binom{n_s}{m_s}$ possible treatment assignments $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_S \end{bmatrix}$ in which $\mathbf{z}_s$ is an n_s-tuple with m_s ones and $n_s - m_s$ zeros for $s = 1, \ldots, S$. In the most common assignment mechanism, each of these K possible assignments is given the same probability, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$ all $\mathbf{z} \in \Omega$. This type of randomized experiment, with equal probabilities and fixed m_s, will be called a *uniform randomized experiment*. When there is but a single stratum, $S = 1$, it has traditionally been called a *completely randomized experiment*, but when there are two or more strata, $S \geq 2$, it has been called a *randomized block experiment*. If the strata each contain two units, $n_s = 2$, and one

receives the treatment, $m_s = 1$, then it has been called a *paired randomized experiment*.

As a small illustration, consider a uniform randomized experiment with two strata, $S = 2$, four units in the first stratum, $n_1 = 4$, and two in the second, $n_2 = 2$, and $N = n_1 + n_2 = 6$ units in total. Half of the units in each stratum receive the treatment, so $m_1 = 2$ and $m_2 = 1$. There are $K = 12$ possible treatment assignments $\mathbf{z} = (z_{11}, z_{12}, z_{13}, z_{14}, z_{21}, z_{22})^T$ contained in the set Ω, and each has probability $1/12$. So Ω is the following set of $K = 12$ vectors $\mathbf{z}$ of dimension $N = 6$ with binary coordinates such that $2 = z_{11} + z_{12} + z_{13} + z_{14}$ and $1 = z_{21} + z_{22}$.

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right.$$
$$\left. \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

The following proposition is often useful. It says that in a uniform randomized experiment, the assignments in different strata are independent of each other. For the elementary proof, see Problem 3.

Proposition 1 *In a uniform randomized experiment, the* $\mathbf{Z}_1, \ldots, \mathbf{Z}_S$ *are mutually independent, and* $\text{prob}(\mathbf{Z}_s = \mathbf{z}_s) = 1/\binom{n_s}{m_s}$ *for each* n_s*-tuple* $\mathbf{z}_s$ *containing* m_s *ones and* $n_s - m_s$ *zeros.*

The uniform randomized designs are by far the most common randomized experiments involving two treatments, but others are also used, particularly in clinical trials. It is useful to mention one of these methods of randomization to underscore the point that randomized experiments need not give every treatment assignment $\mathbf{z} \in \Omega_0$ the same probability. A distinguishing feature of many clinical trials is that the units are patients who arrive for treatment over a period of months or years. As a result, the number n_s of people who will fall in stratum s will not be known at the start of the experiment, so a randomized block experiment is not possible. Efron (1971) proposed the following method. Fix a probability p with $1/2 < p < 1$. When the ith patient belonging to stratum s first arrives, calculate a current measure of imbalance in stratum s, IMBAL_{si}, defined to be the number of patients so far assigned to treatment in this stratum minus the number so far assigned to control. It is easy to check that

$\text{IMBAL}_{si} = \sum_{j=1}^{i-1}(2Z_{sj} - 1)$. If $\text{IMBAL}_{si} = 0$, assign the new patient to treatment or control each with probability 1/2. If $\text{IMBAL}_{si} < 0$, so there are too few treated patients in this stratum, then assign the new patient to treatment with probability p and to control with probability $1 - p$. If $\text{IMBAL}_{si} > 0$, so there are too many treated patients, then assign the new patient to treatment with probability $1 - p$ and to control with probability p. Efron examines various aspects of this method. In particular, he shows that it is much better than independent assignment in producing balanced treated and control groups, that is, treated and control groups with similar numbers of patients from each stratum. He also examines potential biases due to the experimenter's knowledge of IMBAL_{si}. Zelen (1974) surveys a number of related methods with similar objectives.

2.4 Testing the Hypothesis of No Treatment Effect

2.4.1 The Distribution of a Test Statistic When the Treatment Is Without Effect

In the theory of experimental design, a special place is given to the test of the hypothesis that the treatment is entirely without effect. The reason is that, in a randomized experiment, this test may be performed virtually without assumptions of any kind, that is, relying just on the random assignment of treatments. Fisher discussed the Lady and her tea with such care to demonstrate this. Other activities, such as estimating treatment effects or building confidence intervals, do require some assumptions, often innocuous assumptions, but assumptions nonetheless. The contribution of randomization to formal inference is most clear when expressed in terms of the test of no effect. Does this mean that such tests are of greater practical importance than point or interval estimates? Certainly not. It is simply that the theory of such tests is less cluttered, and so it sets randomized and nonrandomized studies in sharper contrast. The important point is that, in the absence of difficulties such as noncompliance or loss to follow-up, assumptions play a minor role in randomized experiments, and no role at all in randomization tests of the hypothesis of no effect. In contrast, inference in a nonrandomized experiment requires assumptions that are not at all innocuous. So let us follow Fisher and develop this point with care.

Each unit exhibits a response that is observed some time after treatment. To say that the treatment has no effect on this response is to say that each unit would exhibit the same value of the response whether assigned to treatment or control. If the treatment has no effect on a patient's survival, then the patient would live the same number of months under treatment or under control. This is the definition of "no effect." If changing the treatment assigned to a unit changed that unit's response, then certainly the treatment has at least some effect. If a patient would live one

more month under treatment than under control, then the treatment has some effect on that patient.

In the traditional development of randomization inference, chance and probability enter only through the random assignment of treatments, that is, through the known mechanism that selects the treatment assignment $\mathbf{Z}$ from Ω. The only random quantities are $\mathbf{Z}$ and quantities that depend on $\mathbf{Z}$. When the treatment is without effect, the response of a unit is fixed, in the sense that this response would not change if a different treatment assignment $\mathbf{Z}$ were selected from Ω. Again, this is simply what it means for a treatment to be without effect. When testing the null hypothesis of no effect, the response of the ith unit in stratum s is written r_{si} and the N-tuple of responses for all N units is written $\mathbf{r}$. The lowercase notation for r_{si} emphasizes that, under the null hypothesis, r_{si} is a fixed quantity and not a random variable. Later on, when discussing treatments with effects, a different notation is needed.

A test statistic $t(\mathbf{Z}, \mathbf{r})$ is a quantity computed from the treatment assignment $\mathbf{Z}$ and the response $\mathbf{r}$. For instance, the treated-minus-control difference in sample means is the test statistic

$$t(\mathbf{Z}, \mathbf{r}) = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{r}}{\mathbf{Z}^{\mathrm{T}}\mathbf{1}} - \frac{(\mathbf{1}-\mathbf{Z})^{\mathrm{T}}\mathbf{r}}{(\mathbf{1}-\mathbf{Z})^{\mathrm{T}}\mathbf{1}}, \tag{2.2}$$

where $\mathbf{1}$ is an N-tuple of 1s. Other statistics are discussed shortly.

Given any test statistic $t(\mathbf{Z}, \mathbf{r})$, the task is to compute a significance level for a test that rejects the null hypothesis of no treatment effect when $t(\mathbf{Z}, \mathbf{r})$ is large. More precisely:

(i) The null hypotheses of no effect is tentatively assumed to hold, so $\mathbf{r}$ is fixed.

(ii) A treatment assignment $\mathbf{Z}$ has been selected from Ω using a known random mechanism.

(iii) The observed value, say T, of the test statistic $t(\mathbf{Z}, \mathbf{r})$ has been calculated.

(iv) We seek the probability of a value of the test statistic as large or larger than that observed if the null hypothesis were true.

The significance level is simply the sum of the randomization probabilities of assignments $\mathbf{z} \in \Omega$ that lead to values of $t(\mathbf{z}, \mathbf{r})$ greater than or equal to the observed value T, namely,

$$\text{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq T\} = \sum_{\mathbf{z} \in \Omega} [t(\mathbf{z}, \mathbf{r}) \geq T] \cdot \text{prob}(\mathbf{Z} = \mathbf{z}), \tag{2.3}$$

$$\text{where} \quad [\text{event}] = \begin{cases} 1 & \text{if event occurs,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

and $\text{prob}(\mathbf{Z} = \mathbf{z})$ is determined by the known random mechanism that assigned treatments. This is a direct calculation, though not always a straightforward one when Ω is extremely large.

In the case of a uniform randomized experiment, there is a simpler expression for the significance level (2.3) since $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K = 1/|\Omega|$. It is the proportion of treatment assignments $\mathbf{z} \in \Omega$ giving values of the test statistic $t(\mathbf{z}, \mathbf{r})$ greater than or equal to T, namely,

$$\text{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq T\} = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{r}) \geq T\}|}{K}. \tag{2.5}$$

2.4.2 More Tea

To illustrate, consider again Fisher's example of the Lady who tastes $N = 8$ cups of tea, all in a single stratum, so $S = 1$. A treatment assignment is an 8-tuple containing four 1s and four 0s. For instance, the assignment $\mathbf{Z} = (1, 0, 0, 1, 1, 0, 0, 1)^{\mathrm{T}}$ would signify that cups 1, 4, 5, and 8 had milk added first and the other cups had tea added first. The set of treatment assignments Ω contains all possible 8-tuples containing four 1s and four 0s, so Ω contains $|\Omega| = K = \binom{8}{4} = 70$ such 8-tuples. The actual assignment was selected at random in the sense that $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K = 1/70$ for all $\mathbf{z} \in \Omega$. Notice that $\mathbf{z}^{\mathrm{T}}\mathbf{1} = 4$ for all $\mathbf{z} \in \Omega$.

The Lady's response for cup i is either $r_i = 1$ signifying that she classifies this cup as milk first or $r_i = 0$ signifying that she classifies it as tea first. Then $\mathbf{r} = (r_1, \ldots, r_8)^{\mathrm{T}}$. Recall that she must classify exactly four cups as milk first, so $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 4$. The test statistic is the number of cups correctly identified, and this is written formally as $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}r + (\mathbf{1} - \mathbf{Z})^{\mathrm{T}}(\mathbf{1} - \mathbf{r}) = 2\mathbf{Z}^{\mathrm{T}}\mathbf{r}$, where the second equality follows from $\mathbf{1}^{\mathrm{T}}\mathbf{1} = 8$, $\mathbf{Z}^{\mathrm{T}}\mathbf{1} = 4$, and $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 4$. To make this illustration concrete, suppose that $\mathbf{r} = (1, 1, 0, 0, 0, 1, 1, 0)$, so the Lady classifies the first, second, sixth, and seventh cups as milk first. To say that the treatment has no effect is to say that she would give this classification no matter how milk was added to the cups, that is, no matter how treatments were assigned to cups. If changing the cups to which milk is added first changes her responses, then she is discerning something, and the treatment has some effect, however slight or erratic.

There is only one treatment assignment $\mathbf{z} \in \Omega$ leading to perfect agreement with the Lady's responses, namely, $\mathbf{z} = (1, 1, 0, 0, 0, 1, 1, 0)$, so if $t(\mathbf{Z}, \mathbf{r}) = 8$ the significance level (2.5) is $\text{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq 8\} = 1/70$. This says that the chance of perfect agreement by accident is $1/70 = 0.014$, a small chance. In other words, if the treatment is without effect, the chance that a random assignment of treatments will just happen to produce perfect agreement is $1/70$.

It is not possible to have seven agreements since to err once is to err twice. How many assignments $\mathbf{z} \in \Omega$ lead to exactly $t(\mathbf{Z}, \mathbf{r}) =$ six agreements? One such assignment with six agreements is $\mathbf{z} = (1, 0, 1, 0, 0, 1, 1, 0)$. Starting

with perfect agreement, $\mathbf{z} = (1,1,0,0,0,1,1,0)$, any one of the four 1s may be made a 0 and any of the four 0s may be made a 1, so there are $16 = 4 \times 4$ assignments with exactly $t(\mathbf{Z},\mathbf{r}) = 6$ agreements. Hence, there are 17 assignments leading to six or more agreements. With six agreements the significance level (2.5) is $\text{prob}\{t(\mathbf{Z},\mathbf{r}) \geq 6\} = 17/70 = 0.24$, no longer a small probability. It would not be surprising to see six or more agreements if the treatment were without effect—it happens by chance as frequently as seeing two heads when flipping two coins.

The key point deserves repeating. Probability enters the calculation only through the random assignment of treatments. The needed probability distribution is known, not assumed. The resulting significance level does not depend upon assumptions of any kind. If the same calculation were performed in a nonrandomized study, it would require an assumption that the distribution of treatment assignments, $\text{prob}(\mathbf{Z} = \mathbf{z})$, is some particular distribution, perhaps the assumption that all assignments are equally probable, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$. In a nonrandomized study, there may be little basis on which to ground or defend this assumption, it may be wrong, and it will certainly be open to responsible challenge and debate. In other words, the importance of the argument just considered is that it is one way of formally expressing the claim that randomized experiments are not open to certain challenges that can legitimately be made to nonrandomized studies.

2.4.3 Some Common Randomization Tests

Many commonly used tests are randomization tests in that their significance levels can be calculated using (2.5), though the tests are sometimes derived in other ways as well. This section briefly recalls and reviews some of these tests. The purpose of the section is to provide a reference for these methods in a common terminology so they may be discussed and used at later stages. Though invented at different times, it is natural to see the methods as members of a few classes whose properties are similar, and this is done beginning in §2.4.4. In most cases, the methods described have various optimality properties which are not discussed here; see Cox (1970) for the optimality properties of the procedures for binary outcomes and Lehmann (1975) for optimality properties of the nonparametric procedures. In all cases, the experiment is the uniform randomized experiment in §2.3.2 with $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$ for all $\mathbf{z} \in \Omega$.

Fisher's exact test for a 2×2 contingency table is, in fact, the test just used for the example of the Lady and her tea. Here, there is one stratum, $S = 1$; the outcome r_i is *binary*, that is, $r_i = 1$ or $r_i = 0$; and the test statistic is the number of responses equal to 1 in the treated group, that is, $t(\mathbf{Z},\mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{r}$. The 2×2 contingency table records the values of Z_i and r_i, as shown in Table 2.2 for Fisher's example. Notice that the marginal totals in this table are fixed by the structure of the experiment, because

TABLE 2.2. The 2 × 2 Table for Fisher's Exact Test for the Lady Tasting Tea.

		Response, r_i		
		1	0	Total
Treatment or control, Z_i	1	$\mathbf{Z}^T\mathbf{r}$	$4 - \mathbf{Z}^T\mathbf{r}$	4
	0	$4 - \mathbf{Z}^T\mathbf{r}$	$\mathbf{Z}^T\mathbf{r}$	4
	Total	4	4	8

$N = 8$ cups, $\mathbf{1}^T\mathbf{r} = 4$ and $\mathbf{1}^T\mathbf{Z} = 4$ are fixed in this experiment. Under the hypothesis of no effect, the randomization distribution of the test statistic $\mathbf{Z}^T\mathbf{r}$ is the hypergeometric distribution. The usual *chi-square* test for a 2×2 table is an approximation to the randomization significance level when N is large.

The *Mantel–Haenszel* (1959) *statistic* is the analogue of Fisher's exact test when there are two or more strata, $S \geq 2$, and the outcome r_{si} is binary. It is extensively used in epidemiology and certain other fields. The data may be recorded in a $2 \times 2 \times S$ contingency table giving treatment Z by outcome r by stratum s. The test statistic is again the number of 1 responses among treated units, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^T\mathbf{r} = \sum_{s=1}^{S} \sum_{i=1}^{n_s} Z_{si} r_{si}$. Under the null hypothesis, the contribution from stratum s, that is, $\sum_{i=1}^{n_s} Z_{si} r_{si}$, again has a hypergeometric distribution, and (2.5) is the distribution of the sum of S independent hypergeometric variables. The Mantel–Haenszel statistic yields an approximation to the distribution of $\mathbf{Z}^T\mathbf{r}$ based on its expectation and variance, as described in more general terms in the next section. One technical attraction of this statistic is that the large sample approximation tends to work well for a $2 \times 2 \times S$ table with large N even if S is also large, so there may be few subjects in each of the S tables. In particular, the statistic is widely used in matching with multiple controls, in which case $m_s = 1$ for each s.

McNemar's (1947) *test* is for paired binary data, that is, for S pairs with $n_s = 2$, $m_s = 1$, and $r_{si} = 1$ or $r_{si} = 0$. The statistic is, yet again, the number of 1 responses among treated units; that is, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^T\mathbf{r}$. McNemar's statistic is, in fact, a special case of the Mantel–Haenszel statistic, though the $2 \times 2 \times S$ table now describes S pairs and certain simplifications are possible. In particular, the distribution of $\mathbf{Z}^T\mathbf{r}$ in (2.5) is that of a constant plus a certain binomial random variable.

Developing these methods for $2 \times 2 \times S$ tables in a different way, Birch (1964) and Cox (1966, 1970) show that these three tests with binary responses possess an optimality property, so there is a sense in which Fisher's exact test, the Mantel–Haenszel test, and McNemar's test are the best tests for the problems they address. Specifically, they show that the test statistic $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^T\mathbf{r}$ together with the significance level (2.5) is a uniformly most powerful unbiased test against alternatives defined in terms of constant odds ratios.

Mantel's (1963) *extension* of the Mantel–Haenszel test is for responses r_{si} that are confined to a small number of values representing a numerical scoring of several ordered categories. As an example of such an outcome, the New York Heart Association classifies coronary patients into one of four categories based on the degree to which the patient is limited in physical activity by coronary symptoms such as chest pain. The categories are:

(1) no limitation of physical activity;

(2) slight limitation, comfortable at rest, but ordinary physical activity results in pain or other symptoms;

(3) marked limitation, minor activities result in coronary symptoms; and

(4) unable to carry on any physical activity without discomfort, which may be present even at rest.

The outcome r_{si} for a patient is then one of the integers 1, 2, 3, or 4. In this case the data might be recorded as a $2 \times 4 \times S$ contingency table for $Z \times r \times s$. Mantel's test statistic is the sum of the response scores for treated units; that is, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{r}$. Birch (1965) shows that the test is optimal in a certain sense.

In the case of a single stratum, $S = 1$, *Wilcoxon's* (1945) *rank sum test* is commonly used to compare outcomes taking many numerical values. In this test, the responses are ranked from smallest to largest. If all N responses were different numbers, the ranks would be the numbers $1, 2, \ldots, N$. If some of the responses were equal, then the average of their ranks would be used. Write q_i for the rank of r_i, and write $\mathbf{q} = (q_1, \ldots, q_N)^{\mathrm{T}}$. For instance, if $N = 4$, and $r_1 = 2.3$, $r_2 = 1.1$, $r_3 = 2.3$, and $r_4 = 7.9$, then $q_1 = 2.5$, $q_2 = 1$, $q_3 = 2.5$, and $q_4 = 4$, since r_2 is smallest, r_4 is largest, and r_1 and r_3 share the ranks 2 and 3 whose average rank is $2.5 = (2 + 3)/2$. Note that the ranks $\mathbf{q}$ are a function of the responses $\mathbf{r}$ which are fixed if the treatment has no effect, so $\mathbf{q}$ is also fixed. The rank sum statistic is the sum of the ranks of the treated observations, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$, and its significance level is determined from (2.5). The properties of the rank sum test have been extensively studied; for instance, see Lehmann (1975, §1) or Hettmansperger (1984, §3). Wilcoxon's rank sum test is equivalent to the *Mann and Whitney* (1947) *test*.

In the case of S matched pairs with $n_s = 2$ and $m_s = 1$ for $s = 1, \ldots, S$, Wilcoxon's (1945) signed rank test is commonly used for responses taking many values. Here, $(Z_{s1}, Z_{s2}) = (1, 0)$ if the first unit in pair s received the treatment or $(Z_{s1}, Z_{s2}) = (0, 1)$ if the second unit received the treatment. In this test, the absolute differences in responses within pairs $|r_{s1} - r_{s2}|$ are ranked from 1 to S, with average ranks used for ties. Let d_s be the rank of $|r_{s1} - r_{s2}|$ thus obtained. The signed rank statistic is the sum of the ranks for pairs in which the treated unit had a higher response than the control unit.

To write this formally, let $c_{s1} = 1$ if $r_{s1} > r_{s2}$ and $c_{s1} = 0$ otherwise, and similarly, let $c_{s2} = 1$ if $r_{s2} > r_{s1}$ and $c_{s2} = 0$ otherwise, so $c_{s1} = c_{s2} = 0$ if $r_{s1} = r_{s2}$. Then $Z_{s1}c_{s1} + Z_{s2}c_{s2}$ equals 1 if the treated unit in pair s had a higher response than the control unit, and equals zero otherwise. It follows that the signed rank statistic is $\sum_{s=1}^{S} d_s \sum_{i=1}^{2} c_{si} Z_{si}$. Note that d_s and c_{si} are functions of $\mathbf{r}$ and so are fixed under the null hypothesis of no treatment effect. Also, if $r_{s1} = r_{s2}$, then pair s contributes zero to the value of the statistic no matter how treatments are assigned. As with the rank sum test, the signed rank test is widely used and has been extensively studied; for instance, see Lehmann (1975, §3) or Hettmansperger (1984, §2). Section 3.2.4 below contains a numerical example using the sign-rank statistic in an observational study.

For stratified responses, a method that is sometimes used involves calculating the rank sum statistic separately in each of the S strata, and taking the sum of these S rank sums as the test statistic. This is the *stratified rank sum statistic*. It is easily checked that this statistic has the form $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$ resembling the rank sum statistic; however, the ranks in $\mathbf{q}$ are no longer a permutation of the numbers $1, 2, \ldots, N$, but rather of the numbers $1, \ldots, n_1,\ 1, \ldots, n_2, \ldots,\ 1, \ldots, n_S$, with adjustments for ties if needed. Also Ω has changed.

Hodges and Lehmann (1962) find the stratified rank sum statistic to be inefficient when S is large compared to N. In particular, for paired data with $S = N/2$, the stratified rank test is equivalent to the sign test, which in turn is substantially less efficient than the signed rank test for data from short-tailed distributions such as the Normal. They suggest as an alternative the method of aligned ranks: the mean in each stratum is subtracted from the responses in that stratum creating aligned responses that are ranked from 1 to N, momentarily ignoring the strata. Writing $\mathbf{q}$ for these aligned ranks, the *aligned rank statistic* is the sum of the aligned ranks in the treated group, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$. See also Lehmann (1975, §3.3).

Another statistic is the *median test*. Let $c_{si} = 1$ if r_{si} is greater than the median of the responses in stratum s and let $c_{si} = 0$ otherwise, and let $\mathbf{c}$ be the N-tuple containing the c_{si}. Then $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{c}$ is the number of treated responses that exceed their stratum medians. With a single stratum, $S = 1$, the median test is quite good in large samples if the responses have a double exponential distribution, a distribution with a thicker tail than the normal; see, for instance, Hettmansperger (1984, §3.4, p. 146) and the more critical comments by Freidlin and Gastwirth (2000). In this test, the median is sometimes replaced by other quantiles or other measures of location.

Start with any statistic $t(\mathbf{Z}, \mathbf{r})$ and the randomization distribution of $t(\mathbf{Z}, \mathbf{r})$ may be determined from (2.5). This is true even of statistics that are commonly referred to a theoretical distribution instead, for instance, the *two-sample* or *paired t-tests*, among others. Welch (1937) and Wilk (1955) studied the relationship between the randomization distribution and the theoretical distribution of statistics that were initially derived from

assumptions of Normally and independently distributed responses. They suggest that the theoretical distribution may be viewed as a computationally convenient approximation to the desired but computationally difficult randomization distribution. That is, they suggest that t-tests, like rank tests or Mantel–Haenszel tests, may be justified solely on the basis of the use of randomization in the design of an experiment, without reference to Normal independent errors. These findings depend on the good behavior of moments of sums of squares of responses over the randomization distribution; therefore, they depend on the absence of extreme responses. Still, the results are important as a conceptual link between Normal theory and randomization inference.

2.4.4 Classes of Test Statistics

The similarity among the commonly used test statistics in §2.4.3 is striking but not accidental. In this book, these statistics are not discussed individually, except when used in examples. The important properties of the methods are shared by large classes of statistics, so it is both simpler and less repetitive to discuss the classes.

Though invented by different people at different times for different purposes, the commonly used statistics in §2.4.3 are similar for good reason. As the sample size N increases, the number K of treatment assignments in Ω grows very rapidly, and the direct calculation in (2.5) becomes very difficult to perform, even with the fastest computers. To see why this is true, take the simplest case consisting of one stratum, $S = 1$, and an equal division of the n subjects into $m = n/2$ treated subjects and $m = n/2$ controls. Then there are $K = \binom{n}{n/2}$ treatment assignments in Ω. If one more unit is added to the experiment, increasing the sample size to $n+1$, then K is increased by a factor of $(n+1) / \{(n/2) + 1\}$, that is, K nearly doubles. Roughly speaking, if the fastest computer can calculate (2.5) directly for at most a sample of size n, and if computing power doubles every year for 10 years, then 10 years hence computing power will be $2^{10} = 1024$ times greater than today and it will be possible to handle a sample of size $n+10$. Direct calculation of (2.5) is not practical for large n.

The usual solution to this problem is to approximate (2.5) using a large sample or asymptotic approximation. The most common approximations use the moments of the test statistic, its expectation and variance, and sometimes higher moments. The needed moments are easily derived for certain classes of statistics, including all those in §2.4.3.

As an alternative to asymptotic approximation, there are several proposals for computing (2.5) exactly, but they are not, as yet, commonly used. One is to compute (2.5) exactly but indirectly using clever computations that avoid working with the set Ω. For some statistics this can be done by calculating the characteristic function of the test statistic and inverting it

using the fast Fourier transform; see Pagano and Tritchler (1983). A second approach is to design experiments differently so that Ω is a much smaller set, perhaps containing 10,000 or 100,000 treatment assignments. In this case, direct calculation is possible and any test statistic may be used; see Tukey (1985) for discussion.

The first class of statistics will be called *sum statistics* and they are of the form $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$, where $\mathbf{q}$ is some function of $\mathbf{r}$. A sum statistic sums the scores q_{si} for treated units. All of the statistics in §2.4.4 are sum statistics for suitable choices of $\mathbf{q}$. In Fisher's exact test, the Mantel–Haenszel test, and McNemar's test, $\mathbf{q}$ is simply equal to $\mathbf{r}$. In the rank sum test, $\mathbf{q}$ contains the ranks of $\mathbf{r}$. In the median test, $\mathbf{q}$ is the vector of ones and zeros identifying responses r_{si} that exceed stratum medians. In the signed rank statistic, $q_{si} = d_s c_{si}$.

Simple formulas exist for the moments of sum statistics under the null hypothesis that the treatment is without effect. In this case, $\mathbf{r}$ is fixed, so $\mathbf{q}$ is also fixed. The moment formulas use the properties of simple random sampling without replacement. Recall that a simple random sample without replacement of size m from a population of size n is a random subset of m elements from a set with n elements where each of the $\binom{n}{m}$ subsets of size m has the same probability $1/\binom{n}{m}$. Cochran (1963) discusses simple random sampling. In a uniform randomized experiment, the m_s treated units in stratum s are a simple random sample without replacement from the n_s units in stratum s. The following proposition is proved in Problem 4.

Proposition 2 *In a uniform randomized experiment, if the treatment has no effect, the expectation and variance of a sum statistic* $\mathbf{Z}^{\mathrm{T}}\mathbf{q}$ *are*

$$E(\mathbf{Z}^{\mathrm{T}}\mathbf{q}) = \sum_{s=1}^{S} m_s \bar{q}_s,$$

and

$$\mathrm{var}(\mathbf{Z}^{\mathrm{T}}\mathbf{q}) = \sum_{s=1}^{S} \frac{m_s(n_s - m_s)}{n_s(n_s - 1)} \sum_{i=1}^{n_s} (q_{si} - \bar{q}_s)^2,$$

where

$$\bar{q}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} q_{si}.$$

Moments are easily determined for sum statistics, but other classes of statistics have other useful properties. The first such class, the *sign-score statistics*, is a subset of the sum statistics. A statistic is a sign-score statistic if it is of the form $t(\mathbf{Z}, \mathbf{r}) = \sum_{s=1}^{S} d_s \sum_{i=1}^{n_s} c_{si} Z_{si}$, where c_{si} is binary, $c_{si} = 1$ or $c_{si} = 0$, and both d_s and c_{si} are functions of $\mathbf{r}$. Fisher's exact test, the

Mantel–Haenszel test, and McNemar's test are sign-score statistics with $d_s = 1$ and $c_{si} = r_{si}$. The signed rank and median test statistics are also sign-score statistics, but with c_{si} and d_s defined differently. A sign-score statistic is a sum statistic with $q_{si} = d_s c_{si}$, but many sum statistics, including the rank sum statistic, are not sign-score statistics. In Chapter 4, certain calculations are simpler for sign-score statistics than for certain other sum statistics, and this motivates the distinction.

Another important class of statistics is the class of *arrangement increasing functions* of $\mathbf{Z}$ and $\mathbf{r}$, which are defined in a moment. Informally, a statistic $t(\mathbf{Z}, \mathbf{r})$ is arrangement-increasing if it increases in value as the coordinates of $\mathbf{Z}$ and $\mathbf{r}$ are rearranged into an increasingly similar order within each stratum. In fact, all of the statistics in §2.4.3 are arrangement-increasing, so anything that is true of arrangement-increasing statistics is true of all the commonly used statistics in §2.3.2. Hollander, Proschan, and Sethuraman (1977) discuss many properties of arrangement-increasing functions.

A few preliminary terms are useful. The numbers S and n_s, $s = 1, \ldots, S$ with $N = \sum n_s$, are taken as given. A *stratified N-tuple* $\mathbf{a}$ is an N-tuple in which the N coordinates are divided into S strata with n_s coordinates in stratum s, where a_{si} is the ith of the n_s coordinates in stratum s. For instance, $\mathbf{Z}$ and $\mathbf{r}$ are each stratified N-tuples. If $\mathbf{a}$ is a stratified N-tuple, and if i and j are different positive integers less than or equal to n_s, then let $\mathbf{a}_{sij}$ be the stratified N-tuple formed from $\mathbf{a}$ by interchanging a_{si} and a_{sj}, that is, by placing the value a_{sj} in the ith position in stratum s and placing the value a_{si} in the jth position in stratum s. To avoid repetition, whenever the symbol $\mathbf{a}_{sij}$ appears, it is assumed without explicit mention that the subscripts are appropriate, so s is a positive integer between 1 and S and i and j are different positive integers less than or equal to n_s. A function $f(\mathbf{a}, \mathbf{b})$ of two stratified N-tuples is *invariant* if $f(\mathbf{a}, \mathbf{b}) = f(\mathbf{a}_{sij}, \mathbf{b}_{sij})$ for all s, i, j, so renumbering units in the same stratum does not change the value of $f(\mathbf{a}, \mathbf{b})$. For instance, the function $\mathbf{z}^T\mathbf{q}$ is an invariant function of $\mathbf{z}$ and $\mathbf{q}$.

Definition 3 *An invariant function $f(\mathbf{a},\mathbf{b})$ of two stratified N-tuples is arrangement-increasing (or AI) if $f(\mathbf{a}, \mathbf{b}_{sij}) \geq f(\mathbf{a},\mathbf{b})$ whenever*

$$(a_{si} - a_{sj}) \cdot (b_{si} - b_{sj}) \leq 0.$$

Notice what this definition says. Consider the ith and jth unit in stratum s. If $(a_{si} - a_{sj})(b_{si} - b_{sj}) < 0$, then of these two units, the one with the higher value of a has the lower value of b, so these two coordinates are out of order. However, in $\mathbf{a}$ and $\mathbf{b}_{sij}$, these two coordinates are in the same order, for b_{si} and b_{sj} have been interchanged. The definition says that an arrangement increasing function will be larger, or at least no smaller, when these two coordinates are switched into the same order.

TABLE 2.3. A Hypothetical Example Showing an Arrangement-Increasing Statistic.

i		$\mathbf{z}$	$\mathbf{q}$	$\mathbf{q}_{23}$
1	Treated	1	4	4
2	Treated	1	2	3
3	Control	0	3	2
4	Control	0	1	1
Rank sum			6	7

Notice also what the definition says when $(a_{si}-a_{sj})(b_{si}-b_{sj}) = 0$. In this case, either $a_{si} = a_{sj}$ or $b_{si} = b_{sj}$ or both. In this case, $f(\mathbf{a}, \mathbf{b}_{sij}) = f(\mathbf{a}, \mathbf{b})$.

Consider some examples. The function $\mathbf{z}^T\mathbf{q}$ is arrangement-increasing as a function of $\mathbf{z}$ and $\mathbf{q}$. To see this, note that $\mathbf{z}^T\mathbf{q}_{sij}-\mathbf{z}^T\mathbf{q} = (z_{si}q_{sj}+z_{sj}q_{si})-(z_{si}q_{si}+z_{sj}q_{sj}) = -(z_{si}-z_{sj})(q_{si}-q_{sj})$, so if $(z_{si}-z_{sj})(q_{si}-q_{sj}) \leq 0$ then $\mathbf{z}^T\mathbf{q}_{sij} - \mathbf{z}^T\mathbf{q} \geq 0$. This shows $\mathbf{z}^T\mathbf{q}$ is arrangement-increasing.

Table 2.3 is a small illustration for the rank sum statistic with a single stratum, $S = 1$, $n = 4$ units, of whom $m = 2$ received the treatment. Here, $(z_2 - z_3)(q_2 - q_3) = (1 - 0)(2 - 3) = -1 \leq 0$, and the rank sum $\mathbf{z}^T\mathbf{q} = 6$ is increased to $\mathbf{z}^T\mathbf{q}_{23} = 7$ by interchanging q_2 and q_3.

As a second example, consider the function $t(\mathbf{z}, \mathbf{r}) = \mathbf{z}^T\mathbf{q}$, where $\mathbf{q}$ is a function of $\mathbf{r}$, which may be written explicitly as $\mathbf{q}(\mathbf{r})$. Then $t(\mathbf{z}, \mathbf{r})$ may or may not be arrangement-increasing in $\mathbf{z}$ and $\mathbf{r}$ depending upon how $\mathbf{q}(\mathbf{r})$ varies with $\mathbf{r}$. The common statistics in §2.4.3 all have the following two properties:

(i) permute $\mathbf{r}$ within strata and $\mathbf{q}$ is permuted in the same way; and

(ii) within each stratum, larger r_{si} receive larger q_{si}.

One readily checks that $t(\mathbf{z}, \mathbf{r}) = \mathbf{z}^T\mathbf{q}$ is arrangement-increasing if $\mathbf{q}(\mathbf{r})$ has these two properties, because the first property ensures that $t(\mathbf{z}, \mathbf{r})$ is invariant, and the second ensures that $r_{si} - r_{sj} \geq 0$ implies $q_{si} - q_{sj} \geq 0$, so $(z_{si} - z_{sj})(r_{si} - r_{sj}) \leq 0$ implies $(z_{si} - z_{sj})(q_{si} - q_{sj}) \leq 0$, and the argument of the previous paragraph applies. The important conclusion is that all of the statistics in §2.4.3 are arrangement-increasing.

In describing the behavior of a statistic when the null hypothesis does not hold and instead the treatment has an effect, a final class of statistics is useful. Many statistics that measure the size of the difference between treated and control groups would tend to increase in value if responses in the treated group were increased and those in the control group were decreased. Statistics with this property will be called effect increasing, and the idea will now be expressed this formally. A treated unit has $2Z_{si}-1 = 1$, since $Z_{si} = 1$, and a control unit has $2Z_{si}-1 = -1$ since $Z_{si} = 0$. Let $\mathbf{z} \in \Omega$

TABLE 2.4. Hypothetical Example of an Effect Increasing Statistic.

i		z_i	$2z_i - 1$	r_i	r_i^*
1	Treated	1	1	5	6
2	Treated	1	1	2	4
3	Control	0	−1	3	2
4	Control	0	−1	1	1
Rank sum				6	7

be a possible treatment assignment and let $\mathbf{r}$ and $\mathbf{r}^*$ be two possible values of the N-tuple of responses such that $(r_{si}^* - r_{si})(2z_{si} - 1) \geq 0$ for all s, i. With treatments given by $\mathbf{z}$, this says that $r_{si}^* \geq r_{si}$ for every treated unit and $r_{si}^* \leq r_{si}$ for every control unit. In words, if higher responses indicated favorable outcomes, then every treated unit does better with $\mathbf{r}^*$ than with $\mathbf{r}$, and every control does worse with $\mathbf{r}^*$ than with $\mathbf{r}$. That is, the difference between treated and control groups looks larger with $\mathbf{r}^*$ than with $\mathbf{r}$. The test statistic is *effect increasing* if $t(\mathbf{z}, \mathbf{r}) \leq t(\mathbf{z}, \mathbf{r}^*)$ whenever $\mathbf{r}$ and $\mathbf{r}^*$ are two possible values of the response such that $(r_{si}^* - r_{si})(2z_{si} - 1) \geq 0$ for all s, i. All of the commonly used statistics in §2.4.3 are effect increasing.

Table 2.4 contains a small hypothetical example to illustrate the idea of an effect increasing statistic. Here there is a single stratum, $S = 1$, and four subjects, $n = 4$, of whom $m = 2$ received the treatment. Notice that when r_i and r_i^* are compared, treated subjects have $r_i^* \geq r_i$ while controls have $r_i^* \leq r_i$. If the responses are ranked 1, 2, 3, 4, and the ranks in the treated group are summed to give Wilcoxon's rank sum statistic, then the rank sum is larger for r_i^* than for r_i.

In summary, this section has considered four classes of statistics:

(i) the sum statistics;

(ii) the arrangement-increasing statistics;

(iii) the effect increasing statistics; and

(iv) the sign-score statistics.

All of the commonly used statistics in §2.4.3 are members of the first three classes, and most are sign-score statistics; however, the rank sum statistic, the stratified rank sum statistic, and Mantel's extension are not sign-score statistics.

*2.4.5 *No Effect Means No Effect*

No effect means no effect. A nonzero effect that varies from one unit to the next and that is hard to fathom or predict is, nonetheless, a nonzero effect. It may not be an immediately useful effect, but it is an effect, perhaps an effect that can someday be understood, tamed, and made useful.

Empirically, it may be difficult to discern erratic unsystematic effects, but logically they are distinct from no effect.

To emphasize this point, consider the extreme case. Suppose that we somehow discerned that the treatment erratically benefits some patients and harms others, but that we have no way of predicting who will benefit or who will be harmed, so the average effect of the treatment is essentially zero in every large group of patients defined by pretreatment variables. In point of fact, it is very difficult to discern something like this, unless we covertly introduce more information that does distinguish these supposedly indistinguishable patients. Suppose, however, we can discern this, perhaps because the treatment produces one of two easily distinguished biochemical reactions, one beneficial, the other harmful, and neither reaction is ever seen among controls; however, we are completely at a loss to identify in advance those patients who will have beneficial reactions. This is a nonzero treatment effect, perhaps not a very useful one given current knowledge, but a nonzero effect nonetheless. What would a scientist do with such an effect? Might the scientist sometimes return with the treatment to the laboratory in an effort to understand why only some patients exhibit the beneficial biochemical reaction? In contrast, no treatment effect—really no treatment effect—would send the scientist in search of another treatment.

No effect is one hypothesis among many. It is rarely, perhaps never, sufficient to know whether the null hypothesis of no treatment effect is compatible with observed data. And yet, it is typically of interest to know this along with much more. Section 2.5 and Chapter 5 discuss models for treatment effects and associated methods of inference, including confidence intervals.

Fisher (1935) and Neyman (1935), two brilliant founders of statistics, did not agree about the meaning of the null hypothesis of no treatment effect. The hypothesis of no effect as I have described it is Fisher's version. Fisher's conception is particular: randomization justifies causal inferences about particular treatment effects, on particular units, at a particular time, under particular circumstances. Change the units or the times or the circumstances and the findings may change to an extent not adequately addressed by statistical standard errors. These standard errors measure one very important source of uncertainty, namely, uncertainty about how units would have responded to a treatment they did not receive, that is, uncertainty about the effects caused by the treatment. Campbell and Stanley (1963) say that randomization ensures *internal validity* but not *external validity*; see §2.7.1 and the discussion of efficacy and effectiveness in §5.4. Neyman's (1935, p. 110) conception is general: we can "repeat the experiment indefinitely without any change of vegetative conditions or of arrangement so that ... the yields from this plot will form a population ...." For Neyman, the variations we do not understand become, by assumption, variations from sampling a population. In point of fact, we cannot repeat the experiment indefinitely, and we cannot ensure the same experimental

conditions, but this conception concerns a hypothetical world in which we can. This was not a disagreement about matters of fact, but about matters of art, the art of developing statistical concepts for scientific applications.

In most cases, their disagreement is entirely without technical consequence: the same procedures are used, and the same conclusions are reached. Perhaps this is expressed most beautifully by Lehmann (1959, §5). First, Lehmann (1959, §5.7, Theorem 3) shows that inferences under a population model can be distribution-free only if they are made particular by conditioning on observed responses, yielding Fisher's randomization test. Lehmann (1959, §5.8) then uses a population model and the Neyman—Pearson lemma to obtain most powerful permutation tests; that is, he uses Neyman's conception to obtain the best tests of the type Fisher was proposing. Whatever Fisher and Neyman may have thought, in Lehmann's text they work together. The importance to mathematical statistics and to science of infinite population models and Neyman's contributions are, today, surely unquestioned.

And yet, when one is thinking about the science of an experiment, it is surely true that random assignment of treatments justifies inferences that are particular, that is, particular to certain units at certain times under certain circumstances. If the inference reaches beyond that to infinite populations extending into the indefinite future, then this has been accomplished by assuming those populations into existence, and assuming away much that is true of the world we actually inhabit. In those instances where their conceptions point in scientifically different directions—for instance, the unpredictable but distinguishable biochemical reactions above—it seems to me that Fisher's conception more closely describes how scientists think and work. Much that we cannot currently predict and do not currently fathom is not random error. The variation we do not fathom today we intend to decipher tomorrow.

2.5 Simple Models for Treatment Effects

2.5.1 Responses When the Treatment Has an Effect

If the treatment has an effect, then the observed N-tuple of responses for the N units will be different for different treatment assignments $\mathbf{z} \in \Omega$—this is what it means to say the treatment has an effect. In earlier sections, the null hypothesis of no treatment effect was assumed to hold, so the observed responses were fixed, not varying with $\mathbf{z}$, and the response was written $\mathbf{r}$. When the null hypothesis of no effect is not assumed to hold, the response changes with $\mathbf{z}$, and the response observed when the treatment assignment is $\mathbf{z} \in \Omega$ will be written $\mathbf{r}_\mathbf{z}$. The null hypothesis of no treatment effect says that $\mathbf{r}_\mathbf{z}$ does not vary with $\mathbf{z}$, and instead $\mathbf{r}_\mathbf{z}$ is a constant the same for all $\mathbf{z}$; in this case, $\mathbf{r}$ was written for this constant. Notice that, for each

$\mathbf{z} \in \Omega$, the response $\mathbf{r_z}$ is some nonrandom N-tuple—probability has not yet entered the discussion. Write $r_{si\mathbf{z}}$ for the (s, i) coordinate of $\mathbf{r_z}$, that is, for the response of the ith unit in stratum s when the N units receive the treatment assignment $\mathbf{z}$.

To make this definite, return for a moment to Fisher's Lady tasting tea. If the Lady could not discriminate at all, then no matter how milk is added to the cup—that is, no matter what $\mathbf{z}$ is—she will classify the cups in the same way; that is, she will give the same binary 8-tuple of responses $\mathbf{r}$. On the other hand, if she discriminates perfectly, always classifying cups correctly, then her 8-tuple of responses will vary with $\mathbf{z}$; indeed, the responses will match the treatment assignments so that $\mathbf{r_z} = \mathbf{z}$.

If treatments are randomly assigned, then the treatment assignment $\mathbf{Z}$ is a random variable, so the observed responses are also random variables as they depend on $\mathbf{Z}$. Specifically, the observed response is the random variable $\mathbf{r_Z}$, that is, one of the many possible $\mathbf{r_z}$, $\mathbf{z} \in \Omega$, selected by picking a treatment assignment $\mathbf{Z}$ by the random mechanism that governs the experiment. Write $\mathbf{R} = \mathbf{r_Z}$ for the observed response, where $\mathbf{R}$ like $\mathbf{Z}$ is a random variable.

In principle, each possible treatment assignment $\mathbf{z} \in \Omega$ might yield a pattern of responses $\mathbf{r_z}$ that is unrelated to the pattern observed with another $\mathbf{z}$. For instance, in a completely randomized experiment with 50 subjects divided into two groups of 25, there might be $|\Omega| = \binom{50}{25} \doteq 1.3 \times 10^{14}$ different and unrelated 50-tuples $\mathbf{r_z}$. Since it is difficult to comprehend a treatment effect in such terms, we look for regularities, patterns, or models of the behavior of $\mathbf{r_z}$ as $\mathbf{z}$ varies over Ω. The remainder of §2.5 discusses the most basic models for $\mathbf{r_z}$ as $\mathbf{z}$ varies over Ω. Chapter 5 discusses additional models for treatment effects.

2.5.2 No Interference Between Units

A first model is that of "no interference between units" which means that "the observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox, 1958a, §2.4). Rubin (1986) calls this SUTVA for the "stable unit treatment value assumption." Formally, no interference means that $r_{si\mathbf{z}}$ varies with z_{si} but not with the other coordinates of $\mathbf{z}$. In other words, the response of the ith unit in stratum s depends on the treatment assigned to this unit, but not on the treatments assigned to other units, so this unit has only two possible values of the response rather than $|\Omega|$ possible values. When this model is assumed, write r_{Tsi} and r_{Csi} for the responses of the ith unit in stratum s when assigned, respectively, to treatment or control; that is, r_{Tsi} is the common value of $\mathbf{r}_{si\mathbf{z}}$ for all $\mathbf{z} \in \Omega$ with $z_{si} = 1$, and r_{Csi} is the common value of $r_{si\mathbf{z}}$ for all $z \in \Omega$ with $z_{si} = 0$. Then the observed response from the ith unit in stratum s is $R_{si} = r_{Tsi}$ if $Z_{si} = 1$ or $R_{si} = r_{Csi}$ if $Z_{si} = 0$, which may also be written $R_{si} = Z_{si}r_{Tsi} + (1 - Z_{si})r_{Csi}$. This model, with

one potential response for each unit under each treatment, has been important both to experimental design—see Neyman (1923), Welch (1937), Wilk (1955), Cox (1958b, §5), and Robinson (1973)—and to causal inference more generally—see Rubin (1974, 1977) and Holland (1986). When there is no interference between units, write $\mathbf{r}_T$ for $(r_{T11}, \ldots, r_{TS,n_S})^T$ and $\mathbf{r}_C$ for $(r_{C11}, \ldots, r_{CS,n_S})^T$.

"No interference between units" is a model and it can be false. No interference is often plausible when the units are different people and the treatment is a medical intervention with a biological response. In this case, no interference means that a medical treatment given to one patient affects only that patient, not other patients. That is often true. However, a vaccine given to many people may protect unvaccinated individuals by reducing the spread of a virus (so called herd immunity) and this is a form of interference. No interference is less plausible in some social settings, such a workplace or a classroom, where a reward given to one person may be visible to others, and may affect their behavior. No interference is often implausible when the strata are people and the units are repeated measures on a person; then a treatment given at one time may affect responses at later times; see Problem 2. In randomized single subject experiments, such as the Lady tasting tea, no interference is typically implausible.

2.5.3 The Model of an Additive Effect, and Related Models

The model of an additive treatment effect assumes units do not interfere with each other, and the administration of the treatment raises the response of a unit by a constant amount τ, so that $r_{Tsi} = r_{Csi} + \tau$ for each s, i. The principal attraction of the model is that there is a definite parameter to estimate, namely, the additive treatment effect τ. As seen in §2.7, in a uniform randomized experiment, many estimators do indeed estimate τ when this model holds.

In understanding the model of an additive treatment effect, it is important to keep in mind that the pair of responses, (r_{Tsi}, r_{Csi}), is never jointly observed for one unit (s, i). Therefore the model of an additive effect, $r_{Tsi} = r_{Csi} + \tau$, cannot be checked directly by comparing r_{Tsi} and r_{Csi} for particular units. The treatment Z_{si} and the observed response $R_{si} = Z_{si} r_{Tsi} + (1 - Z_{si}) r_{Csi}$ are observed, and one can check what the model, $r_{Tsi} = r_{Csi} + \tau$, implies about these observable quantities. In a completely randomized experiment with a single stratum, $S = 1$, dropping the s, the model of an additive treatment effect, $r_{Ti} = r_{Ci} + \tau$, implies that, as sample sizes m and $n - m$ increase, the distribution of observed responses R_i for treated units $Z_i = 1$ will be shifted by τ when compared to the distribution of observed responses R_i for controls $Z_i = 0$, so the distributions will have the same shape and dispersion. That is, the histograms or boxplots would look the same, but one would be moved left or right relative to the other. This is a shift model, commonly used in nonparametrics; see

Lehmann (1975). In a uniform randomized experiment with several strata $S > 1$ and $r_{Tsi} = r_{Csi} + \tau$, the distribution of responses may have different shapes and dispersions in different strata, but within each stratum, the treated and control distributions are shifted by τ. This is a fairly weak form of no interaction between treatment group and stratum in the $2 \times S$ table of observable distributions, and it implies much less about observable distributions than the analogous nonparametric analysis of variance model, which typically assumes a common shape and dispersion in all $2S$ cells. If the only data are (Z_{si}, R_{si}), does the additive model have content beyond its implications for observable distributions? See Problem 7.

Under the additive model, the observed response from the ith unit in stratum s is $R_{si} = r_{Csi} + \tau Z_{si}$, or $\mathbf{R} = \mathbf{r}_C + \tau \mathbf{Z}$. It follows that the *adjusted responses*, $\mathbf{R} - \tau \mathbf{Z} = \mathbf{r}_C$, are fixed, not varying with the treatment assignment, $\mathbf{Z}$, so the adjusted responses satisfy the null hypothesis of no effect. This fact will be useful in drawing inferences about τ.

There are many similar models, including the model of a multiplicative effect, $r_{Tsi} = \sigma r_{Csi}$. Chapter 5 discusses quite different models for treatment effects.

2.5.4 *Positive Effects and Larger Effects

The model of an additive effect assumes a great deal about the relationship between r_{Tsi} and r_{Csi}. At times, it is desirable to describe the behavior of statistical procedures while assuming much less. When there is no interference between units, an effect is a pair $(\mathbf{r}_T, \mathbf{r}_C)$ giving the responses of each unit under each treatment. Two useful concepts are positive effects and larger effects. Unlike the model of an additive treatment effect, positive effects and larger effects are meaningful not just for continuous responses, but also for binary responses, for ordinal responses, and as seen later in §2.8, for censored responses and multivariate responses.

A treatment has a *positive effect* if $r_{Tsi} \geq r_{Csi}$ for all units (s, i) with strict inequality for at least one unit. A more compact way of writing this is that $(\mathbf{r}_T, \mathbf{r}_C)$ is a positive effect if $\mathbf{r}_T \geq \mathbf{r}_C$ with $\mathbf{r}_T \neq \mathbf{r}_C$. This says that application of the treatment never decreases a unit's response and sometimes increases it. For instance, there is a positive effect if the effect is additive and $\tau > 0$. Hamilton (1979) discusses this model in detail when the outcome is binary.

Consider two possible effects, say $(\mathbf{r}_T, \mathbf{r}_C)$ and $(\mathbf{r}^*_T, \mathbf{r}^*_C)$. Then $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ is a *larger effect* than $(\mathbf{r}_T, \mathbf{r}_C)$ if $r^*_{Tsi} \geq r_{Tsi}$ and $r^*_{Csi} \leq r_{Csi}$ for all s, i. For instance, the simplest example occurs when the treatment effect is additive with the same responses under control, namely, $\mathbf{r}^*_C = \mathbf{r}_C$, $\mathbf{r}_T = \mathbf{r}_C + \tau \mathbf{1}$, and $\mathbf{r}^*_T = \mathbf{r}_C + \tau^* \mathbf{1}$, for in this case $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ exhibits a *larger effect* than $(\mathbf{r}_T, \mathbf{r}_C)$ if $\tau^* \geq \tau$. In general, write $\mathbf{R}$ and $\mathbf{R}^*$ for the observed responses from, respectively, the effects $(\mathbf{r}_T, \mathbf{r}_C)$ and $(\mathbf{r}^*_T, \mathbf{r}^*_C)$, so $R^*_{si} = r^*_{Tsi}$ if $Z_{si} = 1$ and $R^*_{si} = r^*_{Csi}$ if $Z_{si} = 0$.

If a statistical test rejects the null hypothesis 5% of the time when it is true, one would hope that it would reject at least 5% of the time when it is false in the anticipated direction. Recall that a statistical test is *unbiased* against a collection of alternative hypotheses if the test is at least as likely to reject the null hypothesis when one of the alternatives is true as when the null hypothesis is true. The next proposition says that all of the common tests in §2.4.3 are unbiased tests against positive treatment effects, and the test statistic is larger when the effect is larger. The proposition is proved in somewhat more general terms in the appendix, §2.9.

Proposition 4 *In a randomized experiment, a test statistic that is effect increasing yields an unbiased test of no effect against the alternative of a positive effect, and if* $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ *is a larger effect than* $(\mathbf{r}_T, \mathbf{r}_C)$*, then* $t(\mathbf{Z}, \mathbf{R}^*) \geq t(\mathbf{Z}, \mathbf{R})$.

2.6 Confidence Intervals

2.6.1 Testing General Hypotheses

So far, the test statistic $t(\mathbf{Z}, \mathbf{R})$ has been used to test the null hypothesis of no treatment effect. There is an extension to test hypotheses that specify a particular treatment effect. In §2.6.2, this extension is used to construct confidence intervals. As always, the confidence interval is the set of hypotheses not rejected by a test.

Consider testing the hypothesis $H_0 : \tau = \tau_0$ in the model of an additive effect, $\mathbf{R} = \mathbf{r}_C + \tau\mathbf{Z}$. The idea is as follows. If the null hypothesis $H_0 : \tau = \tau_0$ were true, then $\mathbf{r}_C = \mathbf{R} - \tau_0\mathbf{Z}$, so testing $H_0 : \tau = \tau_0$ is the same as testing that $\mathbf{R} - \tau_0\mathbf{Z}$ satisfies the null hypothesis of no treatment effect.

More precisely, if $\mathbf{r}_C$ were known, the probability, say α, that $t(\mathbf{Z}, \mathbf{r}_C)$ is greater than or equal to some fixed number T could be determined from (2.3). If the null hypothesis were true, then $\mathbf{r}_C$ would equal the *adjusted responses*, $\mathbf{R} - \tau_0\mathbf{Z}$, so under the null hypothesis, $\mathbf{r}_C$ can be calculated from τ_0 and the observed data. If the hypothesis $H_0 : \tau = \tau_0$ is true, then the chance that $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z}) \geq T$ is α, where α is calculated as described above with $\mathbf{r}_C = \mathbf{R} - \tau_0\mathbf{Z}$.

Now, suppose the null hypothesis is not true, say instead $\tau > \tau_0$, and consider the behavior of the above test. In this case, $\mathbf{R} = \mathbf{r}_C + \tau\mathbf{Z}$ and the adjusted responses $\mathbf{R} - \tau_0\mathbf{Z}$ equal $\mathbf{r}_C + (\tau - \tau_0)\mathbf{Z}$, so the adjusted responses will vary with the assigned treatment $\mathbf{Z}$. If a unit receives the treatment, it will have an adjusted response that is $\tau - \tau_0$ higher than if this unit receives the control. If the test statistic is effect increasing, as is true of all the statistics in §2.4.3, then $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z}) = t\{\mathbf{Z}, \mathbf{r}_C + (\tau - \tau_0)\mathbf{Z}\} \geq t(\mathbf{Z}, \mathbf{r}_C) = t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$, where the inequality follows from the definition of an effect increasing statistic. In words, if the null hypothesis is false and

TABLE 2.5. Example of Confidence Interval Computations.

Unit	Control Response	Group	Observed Response	Adjusted Response	Ranks of Adjusted Responses
i	r_{Ci}	Z_i	$R_i = r_{Ci} + \tau Z_i$	$R_i - \tau_0 Z_i$	q_i
1	2	1	9	8	7
2	1	0	1	1	1
3	3	0	3	3	2
4	4	0	4	4	3
5	0	1	7	6	5
6	4	1	11	10	8
7	1	1	8	7	6
8	5	0	5	5	4
			$\tau = 7, \tau_0 = 1$		

instead $\tau > \tau_0$, then an effect increasing test statistic $t(\mathbf{Z}, \mathbf{R} - \tau_0 \mathbf{Z})$ will be larger with the incorrect τ_0 than it would have been had we tested the correct value τ.

Table 2.5 illustrates these computations with a rank sum test. It is a hypothetical uniform randomized experiment with $N = 8$ units, all in a single stratum $S = 1$, with $m = 4$ units assigned to treatment, and an additive treatment effect $\tau = 7$, though the null hypothesis incorrectly says $H_0 : \tau = \tau_0 = 1$. The rank sum computed from the adjusted responses $\mathbf{R} - 1\mathbf{Z}$ is $7 + 5 + 8 + 6 = 26$, which is the largest possible rank sum for $N = 8$, $m = 4$, and the one-sided significance level is $\binom{8}{4}^{-1} = 1/70 = 0.014$. The two-sided significance level is $2 \times 0.014 = 0.028$. After removing the hypothesized $\tau_0 = 1$ from treated units, the treated units continue to have higher responses than the controls.

2.6.2 Confidence Intervals by Inverting a Test

Under the model of an additive treatment effect, $\mathbf{R} = \mathbf{r}_C + \tau \mathbf{Z}$, a $1 - \alpha$ confidence set for τ is obtaining by testing each value of τ as in §2.6.1 and collecting all values not rejected at level α into a set A. More precisely, A is the set of values of τ that, when tested, yield significance levels or P-values greater than or equal to α. For instance, in the example in Table 2.5, the value $\tau = 1$ would not be contained in a 95% confidence set. When the true value τ is tested, it is rejected with probability no greater than α, so the random set A contains the true τ with probability at least $1 - \alpha$. This is called "inverting" a test, and it is the standard way of obtaining a confidence set from a test; see, for instance, Cox and Hinkley (1974, §7.2) or Lehmann (1959, §3.5). For many test statistics, a two-sided test yields

a confidence set that is an interval, whose endpoints may be determined by a line search, as illustrated in §4.3.5. Section 3.2.4 uses this confidence interval in an observational study of lead in the blood of children.

2.7 Point Estimates

2.7.1 Unbiased Estimates of the Average Effect

The most quoted fact about randomized experiments is that they lead to unbiased estimates of the average treatment effect. Take the simplest case, a uniform randomized experiment with a single stratum, with no interference between units. In this case, there are m treated units, $N-m$ control units, $E(Z_i) = m/N$, $R_i = r_{Ti}$ if $Z_i = 1$, and $R_i = r_{Ci}$ if $Z_i = 0$. The difference between the mean response in the treated group, namely, $(1/m)\sum Z_i R_i$, and the mean response in the control group, namely, $\{1/(N-m)\}\sum(1-Z_i)R_i$, has expectation

$$E\left\{\sum \frac{Z_i R_i}{m} - \frac{(1-Z_i)R_i}{N-m}\right\} = E\left\{\sum \frac{Z_i r_{Ti}}{m} - \frac{(1-Z_i)r_{Ci}}{N-m}\right\}$$

$$= \sum \frac{(m/N)r_{Ti}}{m} - \frac{(1-m/N)r_{Ci}}{N-m} = \frac{1}{N}\sum r_{Ti} - r_{Ci},$$

and the last term is the average of the N treatment effects $r_{Ti} - r_{Ci}$ for the N experimental units. In words, the difference in sample means is unbiased for the average effect of the treatment. Notice carefully that this is true assuming only that there is no interference between units. There is no assumption that the treatment effect $r_{Ti} - r_{Ci}$ is constant from unit to unit, no assumption about interactions.

The estimate is unbiased for the average effect on the N units in this study, namely, $(1/N)\sum r_{Ti} - r_{Ci}$, but this says nothing about the effect on other units not in the study. Campbell and Stanley (1963) say that a randomized experiment has *internal validity* in permitting inferences about effects for the N units in the study, but it need not have *external validity* in that there is no guarantee that the treatment will be equally effective for other units outside the study; see also §2.4.5. The related issue of efficacy and effectiveness is discussed in §5.4.

The difference in sample means may be biased when there are two or more strata and the experimenter assigns disproportionately more subjects to the treatment in some strata than in others. However, there is an unbiased estimate that corrects the imbalance. It consists of calculating, within stratum s, the difference between the average response in the treated group, namely, $(1/m_s)\sum_i Z_{si}R_{si}$, and the average response in the control group,

namely, $\{1/(n_s - m_s)\} \sum_i (1 - Z_{si}) R_{si}$, and weighting this difference by the proportion of units in stratum s, namely, n_s/N. The estimate, called *direct adjustment*, is then:

$$\sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{Z_{si} R_{si}}{m_s} - \frac{(1 - Z_{si}) R_{si}}{n_s - m_s} \right\}. \tag{2.6}$$

To check that (2.6) is unbiased, recall that, in a uniform randomized experiment, Z_{si} has expectation m_s/n_s. It follows that (2.6) has expectation

$$\begin{aligned}
&E\left[\sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{Z_{si} R_{si}}{m_s} - \frac{(1 - Z_{si}) R_{si}}{n_s - m_s} \right\} \right] \\
&= E\left[\sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{Z_{si} r_{Tsi}}{m_s} - \frac{(1 - Z_{si}) r_{Csi}}{n_s - m_s} \right\} \right] \\
&= \sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{(m_s/n_s) r_{Tsi}}{m_s} - \frac{(1 - m_s/n_s) r_{Csi}}{n_s - m_s} \right\} \\
&= \frac{1}{N} \sum r_{Ti} - r_{Ci},
\end{aligned}$$

so direct adjustment does indeed give an unbiased estimate of the average effect. In a very clear discussion, Rubin (1977) does calculations of this kind.

In effect, direct adjustment views the treated units and the control units as two stratified random samples from the N units in the experiment. Then (2.6) is the usual stratified estimate of mean response to treatment in the population of N units minus the usual estimate of the mean response to control in the population of N units. Notice again that direct adjustment is unbiased for the average treatment effect even if that effect varies from unit to unit or from stratum to stratum. On the other hand, the average effect is but a summary of the effects, and not a complete description, when the effect varies from one stratum to another.

2.7.2 Hodges–Lehmann Estimates of an Additive Effect

Under the model of an additive effect, $\mathbf{R} = \mathbf{r}_C + \tau \mathbf{Z}$, there are many estimates of τ. One due to Hodges and Lehmann (1963) is closely tied to the test in §2.4 and the confidence interval in §2.6. Recall that $H_0 : \tau = \tau_0$ is tested using $t(\mathbf{Z}, \mathbf{R} - \tau_0 \mathbf{Z})$, that is, by subtracting the hypothesized treatment effect $\tau_0 \mathbf{Z}$ from the observed responses $\mathbf{R}$, and asking whether the adjusted responses $\mathbf{R} - \tau_0 \mathbf{Z}$ appear to be free of a treatment effect. The Hodges–Lehmann estimate of τ is that value $\hat{\tau}$ such that the adjusted responses $\mathbf{R} - \hat{\tau} \mathbf{Z}$ appear to be exactly free of a treatment effect. Consider this

in detail. Throughout this section, the experiment is a uniform randomized experiment.

Suppose that we can determine the expectation, say $\bar{\bar{t}}$, of the statistic $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ when calculated using the correct τ, that is, when calculated from responses $\mathbf{R} - \tau\mathbf{Z}$ that have been adjusted so they are free of a treatment effect. For instance, in an experiment with a single stratum, the rank sum statistic has expectation $\bar{\bar{t}} = m(N+1)/2$ if the treatment has no effect. This is true because, in the absence of a treatment effect, the rank sum statistic is the sum of m scores randomly selected from N scores whose mean is $(N+1)/2$. In the same way, in a stratified experiment, the stratified rank sum statistic has expectation $\bar{\bar{t}} = \frac{1}{2}\sum m_s(n_s+1)$ in the absence of a treatment effect. In an experiment comprised of S pairs, in the absence of a treatment effect, the expectation of the signed rank statistic is $\bar{\bar{t}} = (S+1)/4$, since we expect to sum half of S scores which average $(S+1)/2$. In the absence of an effect, in an experiment with a single stratum, the difference in sample means (2.2) has expectation $\bar{\bar{t}} = 0$. In each of these cases, $\bar{\bar{t}}$ may be determined without knowing τ, so there is a Hodges–Lehmann estimate.

Roughly speaking, the Hodges–Lehmann estimate is the solution $\hat{\tau}$ of the equation $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$. In other words, $\hat{\tau}$ is the value such that the adjusted responses $\mathbf{R} - \hat{\tau}\mathbf{Z}$ appear to be entirely free of a treatment effect, in the sense that the test statistic $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ exactly equals its expectation in the absence of an effect.

If $t(\cdot, \cdot)$ is an effect increasing statistic, as is true of all of the statistics in §2.3, then $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ is monotone decreasing as a function of $\hat{\tau}$ with $\mathbf{Z}$ and $\mathbf{R}$ fixed. This says: The larger the treatment effect $\hat{\tau}\mathbf{Z}$ removed from the observed responses $\mathbf{R}$, the smaller the statistic becomes. This is useful in solving $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$. If a $\hat{\tau}$ has been tried such that $\bar{\bar{t}} < t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$, then a larger $\hat{\tau}$ will tend to make $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ smaller, moving it toward $\bar{\bar{t}}$. Similarly, if $\bar{\bar{t}} > t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$, then a smaller $\hat{\tau}$ is needed.

Problems arise immediately. For rank statistics, such as the rank sum and the signed rank, $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ varies in discrete jumps as $\hat{\tau}$ is varied, so there may be no value $\hat{\tau}$ such that $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$. To see this, take a trivial case, a uniform experiment in one stratum, sample size $N = 2$, one treated unit $m = 1$. Then the rank sum statistic is either 1 or 2 depending upon which of the two units receive the treatment, but $\bar{\bar{t}} = 1.5$, so it is not possible to find a $\hat{\tau}$ such that $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$.

Not only may $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ have no solution $\hat{\tau}$, but it may have infinitely many solutions. If $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ varies in discrete jumps, it will be constant for intervals of values of $\hat{\tau}$.

Hodges and Lehmann resolve these problems in the following way. They define the solution of an equation $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ as SOLVE$\{\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} -$

TABLE 2.6. Computing a Hodges–Lehmann Estimate.

τ	4.9999	5	5.0001	5.9999	6	6.0001
$t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$	20	19	18	18	17	15

$\hat{\tau}\mathbf{Z})\}$ defined by

$$\begin{aligned} \hat{\tau} &= \text{SOLVE}\{\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})\} \\ &= \frac{\inf\{\tau : \bar{\bar{t}} > t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})\} + \sup\{\tau : \bar{\bar{t}} < t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})\}}{2}. \end{aligned}$$

This defines the Hodges–Lehmann estimate. Roughly speaking, if there is no exact solution, then average the smallest τ that is too large and the largest τ that is too small.

Consider the small example in Table 2.5. Under the null hypothesis of no effect, the rank sum statistic has expectation $\bar{\bar{t}} = m(N+1)/2 = 4(8+1)/2 = 18$, that is, half of the sum of all eight ranks, $36 = 1+2+\cdots+8$. Table 2.6 gives values of $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ for several values of τ. As noted, since $t(\cdot, \cdot)$ is effect increasing, in Table 2.6, $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ decreases in τ. We want as our estimate a value $\hat{\tau}$ such that $18 = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$, but the table indicates that any value between 5 and 6 will do. As the table suggests, $\inf\{\tau : \bar{\bar{t}} > t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})\} = 6$ and $\sup\{\tau : \bar{\bar{t}} < t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})\} = 5$, so the Hodges–Lehmann estimate is $\hat{\tau} = (6+5)/2 = 5.5$.

For particular test statistics, there are other ways of computing $\hat{\tau}$. This is true, for instance, for a single stratum using the rank sum test. In this case, it may be shown that $\hat{\tau}$ is the median of the $m(N-m)$ pairwise differences formed by taking each of the m treated responses and subtracting each of the $N - m$ control responses.

The Hodges–Lehmann estimate $\hat{\tau}$ inherits properties from the test statistic $t(\cdot, \cdot)$. Consistency is one such property. Recall that a test is *consistent* if the probability of rejecting each false hypothesis tends to one as the sample size increases. Recall that an estimate is consistent if the probability that it is close to the true value tends to one as the sample size increases. As one would expect, these ideas are interconnected. A test that rejects incorrect values of τ leads to an estimate that moves away from these incorrect values. In other words, under mild conditions, consistent tests lead to consistent Hodges–Lehmann estimates; see Maritz (1981, §1.4) for some details.

2.8 *More Complex Outcomes

*2.8.1 *Partially Ordered Outcomes*

So far, the outcome R_{si} has been a number, possibly a continuous measurement, possibly a binary event, possibly a discrete score, but always a single number. However, for more complex responses, much of the earlier discussion continues to apply with little or no change. The purpose of §2.8 is to discuss issues that arise with certain complex outcomes, including multivariate responses and censored observations.

When the outcome R_{si} is a single number, it is clear what it means to speak of a high or low response, and it is clear what it means to ask whether responses are typically higher among treated units than among controls. For more complex responses, it may happen that some responses are higher than some others; and yet not every pair of possible responses can be ordered. For example, unit 1 may have a more favorable outcome than units 2 and 3, but units 2 and 3 may have different outcomes neither of which can be described as entirely more favorable than the other. For instance, patient 1 may live longer and have a better quality of life than patients 2 and 3, but patient 2 may outlive patient 3 though patient 3 had a better quality of life than patient 2. In this case, outcomes may be partially ordered rather than totally ordered, an idea that is formalized in a moment. Common examples are given in §2.8.2 and 2.8.3.

A *partially ordered set* or *poset* is a set A together with a relation $\lesssim$ on A such that three conditions hold:

(i) $a \lesssim a$ for all $a \in A$;

(ii) $a \lesssim b$ and $b \lesssim a$ implies $a = b$ for all $a, b \in A$; and

(iii) if $a \lesssim b$ and $b \lesssim c$ then $a \lesssim c$ for all $a, b, c \in A$.

There is *strict inequality* between a and b if $a \lesssim b$ and $a \neq b$. A poset A is *totally ordered* if $a \lesssim b$ or $b \lesssim a$ for every $a, b \in A$. The real numbers with conventional inequality $\leq$ are totally ordered. If A is partially ordered but not totally ordered, then for some $a, b \in A$, $a \neq b$, neither a nor b is higher than the other; that is, neither $a \lesssim b$ nor $b \lesssim a$. Sections 2.8.2 and 2.8.3 discuss two common examples of partially ordered outcomes, namely, censored and multivariate outcomes. Following this, in §2.8.4, general methods for partially ordered outcomes are discussed.

*2.8.2 *Censored Outcomes*

In some experiments, an outcome records the time to some event. In a clinical trial, the outcome may be the time between a patient's entry into the trial and the patient's death. In a psychological experiment, the outcome

may be the time lapse between administration of a stimulus by the experimenter and the production of a response by an experimental subject. In a study of remedial education, the outcome may be the time until a certain level of proficiency in reading is reached.

Times may be censored in the sense that, when data analysis begins, the event may not yet have occurred. The patient may be alive at the close of the study. The stimulus may never elicit a response. The student may not develop proficiency in reading during the period under study.

If the event occurs for a unit after, say, 3 months, the unit's response is written 3. If the unit entered the study 3 months ago, if the event has not yet occurred, and if the analysis is done today, then the unit's response is written 3+ signifying that the event has not occurred in the initial 3 months.

Censored times are partially ordered. To see this, consider a simple illustration. In a clinical trial, patient 1 died at 3 months, patient 2 died at 12 months, and patient 3 entered the study 6 months ago and is alive today yielding a survival of 6+ months. Then patient 1 had a shorter survival than patients 2 and 3, but it is not possible to say whether patient 2 survived longer than patient 3 because we do not know whether patient 3 will survive for a full year.

The set A of censored survival times contains the nonnegative real numbers together with the nonnegative real numbers with a plus appended. Define the partial order $\lesssim$ on A as follows: if a and b are nonnegative real numbers, then:

(i) $a \lesssim b$ if and only if $a \leq b$;

(ii) $a \lesssim b+$ if and only if $a \leq b$; and

(iii) $a \lesssim a$ and $a+ \lesssim a+$.

Here, (i) indicates that "a" and "b" are both deaths and "a" died first. In (ii), "a" died before "b" was censored, so "b" certainly outlived "a." Of course, (iii) is just the case of equality—every censored time is equal to itself, and so is less than or equal to itself. It is easy to check that this is indeed a partial order, and that strict inequality indicates certainty about who died first.

*2.8.3 *Multivariate Outcomes and Other Partially Ordered Outcomes*

Quite often, a single number is not enough to describe the outcome for a unit. In an educational intervention, there may be test scores in several areas, such as reading and mathematics. In a clinical trial, the outcome may involve both survival and quality of life. A multivariate response is a p-tuple of outcomes describing an individual. If the p components are

numbers, then the multivariate response inherits a partial order as follows: $(a_1, \ldots, a_p) \lesssim (b_1, \ldots, b_p)$ if and only if $a_1 \leq b_1, a_2 \leq b_2, \ldots$, and $a_p \leq b_p$. It is easy to check that this defines a partial order. As an example, if the outcome is the 2-tuple consisting of a reading score and a mathematics score, then one student has a higher multivariate response than another only if the first student did at least as well as the second student on both tests.

In fact, the components of the p-tuple need not be numbers—rather they may be any partially ordered outcomes. In the same way, the p-tuple inherits a partial order from the partial orders of individual outcomes. For instance, the outcome might be a 2-tuple consisting of a censored survival time and a number measuring quality of life. The censored survival times are partially but not totally ordered. In this case, a patient who died early with a poor quality of life would have a lower outcome than a patient who was censored late with a good quality of life.

Multivariate responses may be given other partial orders appropriate to particular contexts. Here is one that gives greatest emphasis to the first coordinate and about equal emphasis to the other two: $(a_1, a_2, a_3) \lesssim (b_1, b_2, b_3)$ if $a_1 \leq b_1$ or if $\{a_1 = b_1$ and $a_2 \leq b_2$ and $a_3 \leq b_3\}$. In an educational setting, this might say that a student who graduates had a better outcome than one who did not regardless of test scores, but among those who graduate, one student is better than another only if both reading and math scores are as good or better.

2.8.4 *A Test Statistic for Partially Ordered Outcomes

The task is to test the null hypothesis of no treatment effect against the alternative that treated units tend to have higher responses than controls in the sense of a partial order $\lesssim$ on the outcomes. For this purpose, define indicators L_{sij} for $s = 1, \ldots, S$, $i = 1, \ldots, n_s$, $j = 1, \ldots, n_s$, as follows:

$$L_{sij} = \begin{cases} 1 & \text{if } R_{sj} \lesssim R_{si} \text{ with } R_{si} \neq R_{sj}, \\ -1 & \text{if } R_{si} \lesssim R_{sj} \text{ with } R_{si} \neq R_{sj}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

In words, L_{sij} compares the ith and jth units in stratum s, and L_{sij} is 1 if the ith is strictly greater than the jth, is -1 if the ith is strictly smaller than the jth, and is zero in all other cases. The statistic is

$$t(\mathbf{Z}, \mathbf{R}) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} Z_{si}(1 - Z_{sj}) L_{sij}. \tag{2.8}$$

Consider the statistic in detail. The term $Z_{si}(1 - Z_{sj})L_{sij}$ equals 1 if, in stratum s, the ith unit received the treatment, the jth unit received the

control, and these two units had unequal responses with the treated unit having a higher response, $R_{sj} \lesssim R_{si}$. Similarly, $Z_{si}(1 - Z_{sj})L_{sij}$ equals -1 if, in stratum s, the ith unit is treated, the jth is a control, and the control had the higher response, $R_{si} \lesssim R_{sj}$. In all other cases, $Z_{si}(1 - Z_{sj})L_{sij}$ equals zero. So the test statistic is the number of comparisons of treated and control units in the same stratum in which the treated unit had the higher response minus the number in which the control unit had the higher response.

This statistic generalizes several familiar statistics. If the outcome is a single number and the partial order $\lesssim$ is ordinary inequality $\leq$, then (2.8) is equivalent to the Mann–Whitney (1947) statistic and the Wilcoxon (1945) rank sum statistic. If the outcome is censored and $\lesssim$ is the partial order in §2.8.2, then the statistic is Gehan's (1965) statistic.

A device due to Mantel (1967) shows that (2.8) is a sum statistic. The steps are as follows. First note that, for any subset B of $\{1, 2, \ldots, n_s\}$,

$$\sum_{i \in B} \sum_{j \in B} L_{sij} = 0 \tag{2.9}$$

since L_{sij} and L_{sji} both appear in the sum, with $L_{sij} = -L_{sji}$, and they cancel. Using this fact with $B = \{i : 1 \leq i \leq n_s \text{ with } Z_{si} = 1\}$ yields

$$0 = \sum_{i \in B} \sum_{j \in B} L_{sij} = \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} Z_{si} Z_{sj} L_{sij},$$

which permits the test statistic (2.8) to be rewritten as the sum statistic

$$t(\mathbf{Z}, \mathbf{R}) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} Z_{si} \sum_{j=1}^{n_s} L_{sij} = \sum_{s=1}^{S} \sum_{i=1}^{n_s} Z_{si} q_{si} \qquad \text{with} \quad q_{si} = \sum_{j=1}^{n_s} L_{sij}.$$

As a result, the expectation and variance of the test statistic under the null hypothesis are given by Proposition 2. In fact, in that Proposition, $\bar{q}_s = 0$ for each s using (2.9).

The score q_{si} has an interpretation. It is the number of units in stratum s with outcomes less than unit i minus the number with outcomes greater than i. The score q_{si} is large if unit i has a response larger than that of most units in stratum s. For instance, in Gehan's statistic for censored outcomes, the score q_{si} is the number of patients in stratum s who definitely died before patient i minus the number who definitely died after patient i.

*2.8.5 *Effect Increasing Statistics, Positive Effects, Larger Effects*

In §2.4 and 2.5, three terms were discussed, namely, effect increasing statistics, positive effects, and larger effects. These terms apply to partially ordered outcomes with virtually no change, as shown in a moment. In each

case, the definitions in §2.4 and 2.5 are the special case of the definitions in this section with the partial order $\lesssim$ given by ordinary inequality $\leq$ of real numbers.

Let $\mathbf{r}$ and $\mathbf{r}^*$ be two possible values of the N-tuple of partially ordered outcomes. If $r_{si} \lesssim r^*_{si}$ for every treated unit and $r^*_{si} \lesssim r_{si}$ for every control unit, then the treated and control groups appear farther apart for outcome $\mathbf{r}^*$ than for outcome $\mathbf{r}$. A test statistic $t(\cdot, \cdot)$ is *effect increasing* if $t(\mathbf{z}, \mathbf{r}) \leq t(\mathbf{z}, \mathbf{r}^*)$ whenever $\mathbf{r}$ and $\mathbf{r}^*$ are two possible values of the response such that $r_{si} \lesssim r^*_{si}$ if $z_{si} = 1$ and $r^*_{si} \lesssim r_{si}$ if $z_{si} = 0$ for all s, i. In words, the statistic is larger when the outcomes in treated and control groups are farther apart. The statistic in §2.8.4 is effect increasing; see Problem 6.

If there is no interference between units, then $(\mathbf{r}_T, \mathbf{r}_C)$ is a *positive effect* if $\mathbf{r}_T \neq \mathbf{r}_C$ and $r_{Csi} \lesssim r_{Tsi}$ for every s, i. In the case of censored survival times, this would mean that each patient would definitely survive at least as long under the treatment as under the control, or else would continue to be censored at the same time due to the end of the study. An effect $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ is a *larger effect* than $(\mathbf{r}_T, \mathbf{r}_C)$ if $r_{Tsi} \lesssim r^*_{Tsi}$ and $r^*_{Csi} \lesssim r_{Csi}$, for all s, i, that is, if the treated responses are higher and the control responses are lower.

The following proposition is the extension of Proposition 4 to partially ordered responses. Again, the proof is given in the appendix, §2.9.

Proposition 5 *In a randomized experiment, a test statistic that is effect increasing yields an unbiased test of no effect against the alternative of a positive effect, and if* $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ *is a larger effect than* $(\mathbf{r}_T, \mathbf{r}_C)$ *then* $t(\mathbf{Z}, \mathbf{R}^*) \geq t(\mathbf{Z}, \mathbf{R})$.

2.9 *Appendix: Effect Increasing Tests Under Alternatives

This appendix proves Propositions 4 and 5 which describe the behavior of effect increasing test statistics under the alternative hypotheses of positive effects or larger effects. It may be of interest to contrast these propositions with a result in Lehmann (1959, §5.8, Lemma 2) which is similar in spirit though quite different in detail. It suffices to prove Proposition 5 since Proposition 4 is the special case of the former in which the partial order is ordinary inequality. The proof depends on the following lemma.

Lemma 6 *Let* $t(\cdot, \cdot)$ *be effect increasing. If* $(\mathbf{r}_T, \mathbf{r}_C)$ *is a positive effect, then* $t(\mathbf{z}, \mathbf{r}_\mathbf{z}) \geq t(\mathbf{z}, \mathbf{r}_\mathbf{a})$ *for all* $\mathbf{z}$, $\mathbf{a} \in \Omega$. *If* $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ *is a larger effect than* $(\mathbf{r}_T, \mathbf{r}_C)$, *then* $t(\mathbf{z}, \mathbf{r}^*_\mathbf{z}) \geq t(\mathbf{z}, \mathbf{r}_\mathbf{z})$ *for all* $\mathbf{z} \in \Omega$.

Proof of Lemma. Let $(\mathbf{r}_T, \mathbf{r}_C)$ be a positive effect, let $\mathbf{z}$, $\mathbf{a} \in \Omega$, and consider $\mathbf{r}_\mathbf{z}$ and $\mathbf{r}_\mathbf{a}$. If $z_{si} = 1$, then $r_{si\mathbf{z}} = \mathbf{r}_{Tsi}$ while $\mathbf{r}_{si\mathbf{a}}$ may equal either

$\mathbf{r}_{Tsi}$ or $\mathbf{r}_{Csi}$ depending on a_{si}, but in either case $r_{si\mathbf{a}} \lesssim r_{si\mathbf{z}}$ since $(\mathbf{r}_T, \mathbf{r}_C)$ is a positive effect. Similarly, if $z_{si} = 0$, then $r_{si\mathbf{z}} = r_{Csi} \lesssim r_{si\mathbf{a}}$. Since $t(\cdot, \cdot)$ is effect increasing, this implies $t(\mathbf{z}, \mathbf{r}_\mathbf{z}) \geq t(\mathbf{z}, \mathbf{r}_\mathbf{a})$, proving the first part of the lemma.

Now let $\mathbf{z} \in \Omega$, let $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ be a larger effect than $(\mathbf{r}_T, \mathbf{r}_C)$, and consider $\mathbf{r}_\mathbf{z}^*$ and $\mathbf{r}_\mathbf{z}$. If $z_{si} = 1$, then $r_{si\mathbf{z}} = r_{Tsi} \lesssim r_{Tsi}^* = r_{si\mathbf{z}}^*$. If $z_{si} = 0$, then $r_{si\mathbf{z}}^* = r_{Csi}^* \lesssim r_{Csi} = r_{si\mathbf{z}}$. Hence $t(\mathbf{z}, \mathbf{r}_\mathbf{z}^*) \geq t(\mathbf{z}, \mathbf{r}_\mathbf{z})$ since $t(\cdot, \cdot)$ is effect increasing, completing the proof. ■

Proof of Proposition 5. The lemma directly shows that if $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ is a larger effect than $(\mathbf{r}_T, \mathbf{r}_C)$, then $t(\mathbf{Z}, \mathbf{R}^*) \geq t(\mathbf{Z}, \mathbf{R})$. To prove unbiasedness, let $\mathbf{Z}$ be randomly selected from Ω where $\text{prob}(\mathbf{Z} = \mathbf{z})$ is known but need not be uniform. If the random treatment assignment turns out to be $\mathbf{Z} = \mathbf{a}$, then the observed outcome is $\mathbf{R} = \mathbf{r}_\mathbf{a}$. If the null hypothesis were true, if the treatment had no effect, the observed response would be the same $\mathbf{r}_\mathbf{a}$ no matter how treatments were assigned, that is, the observed response would be $\mathbf{R} = \mathbf{r}_\mathbf{a}$ no matter what value $\mathbf{Z}$ assumed. If the null hypothesis were false and the treatment had a positive effect, the observed response would vary depending upon the treatment assignment, $\mathbf{R} = \mathbf{r}_\mathbf{z}$ if $\mathbf{Z} = \mathbf{z}$. For any fixed number T

$$
\begin{aligned}
&\text{prob}\{t(\mathbf{Z}, \mathbf{R}) \geq T\} \\
&\quad = \sum_{\mathbf{z} \in \Omega} [t(\mathbf{z}, \mathbf{r}_\mathbf{z}) \geq T] \ \text{prob}(\mathbf{Z} = \mathbf{z}) \\
&\quad \geq \sum_{\mathbf{z} \in \Omega} [t(\mathbf{z}, \mathbf{r}_\mathbf{a}) \geq T] \ \text{prob}(\mathbf{Z} = \mathbf{z}) \qquad \text{for } \mathbf{a} \in \Omega \text{ by the lemma.}
\end{aligned}
$$

In other words, the chance that the test statistic $t(\mathbf{Z}, \mathbf{R})$ exceeds any number T is at least as great under the alternative hypothesis of a positive effect as under the null hypothesis of no effect, proving unbiasedness. ■

2.10 *Appendix: The Set of Treatment Assignments

*2.10.1 *Outline and Motivation: The Special Structure of Ω*

The set Ω of treatment assignments plays an important role both in randomized experiments and in the discussion of observational studies in later chapters. This set Ω possess a special structure, first noted by Savage (1964). Using this structure, a single theorem may refer to large classes of test statistics and to all of the simple designs, including matched pairs, matching with multiple controls, two-group comparisons, and stratified comparisons. The purpose of this section is to describe the special structure of Ω. Appendices in later chapters refer back to this appendix.

Savage (1964) observed that the set Ω is a finite distributive lattice. This is useful because there are tidy theorems about probability distributions on a finite distributive lattice, including the FKG inequality and Holley's inequality. This section:

(i) offers a little motivation;

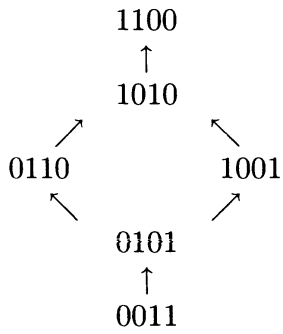
(ii) reviews the definition of a distributive lattice;

(iii) shows that Ω is indeed such a lattice; and

(iv) discusses the relevant probability inequalities.

The material in this appendix may be read without previous experience with lattices.

For motivation, consider a simple case. There is a single stratum, $S = 1$, so the s subscript is dropped in this example, and there are $n = 4$ units of which $m = 2$ receive the treatment. Then Ω contains $\binom{4}{2} = 6$ possible treatment assignments. Assume for this motivating example that the null hypothesis of no treatment effect holds, and renumber the four subjects so their observed responses are in decreasing order, $r_1 \geq r_2 \geq r_3 \geq r_4$. Since no quantity we calculate ever depends on the numbering of subjects, this renumbering changes nothing, but it is notationally convenient. The six possible treatment assignments appear in (2.10).



(2.10)

The treatment assignment $\mathbf{z} = (1, 1, 0, 0)$ at the top in (2.10) is the one that would suggest the largest positive treatment effect, since this assignment places the two largest responses, r_1 and r_2, in the treated group. The assignment below this, namely, $\mathbf{z} = (1, 0, 1, 0)$ would suggest a smaller treatment effect than (1, 1, 0, 0), since r_3 has replaced r_2, but it would suggest a larger treatment effect than any other assignment. The assignments (0, 1, 1, 0) and (1, 0, 0, 1) are not directly comparable to each other, since the latter places the largest and smallest responses in the treated group while the former places the two middle responses in the treated group; however, both are lower than (1, 0, 1, 0) and both are higher than (0, 1, 0, 1).

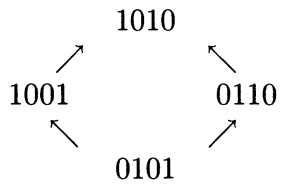
Consider the behavior of a test statistic $t(\mathbf{z}, \mathbf{r})$ as we move through (2.10). Suppose, for instance, there are no ties among the responses, $r_1 > r_2 >$

$r_3 > r_4$, and $t(\mathbf{z}, \mathbf{r})$ is the rank sum statistic. Then $t(\mathbf{z}, \mathbf{r}) = 7$ for $\mathbf{z} = 1100$, $t(\mathbf{z}, \mathbf{r}) = 6$ for 1010, $t(\mathbf{z}, \mathbf{r}) = 5$ for both 1001 and 0110, $t(\mathbf{z}, \mathbf{r}) = 4$ for 0101, and $t(\mathbf{z}, \mathbf{r}) = 3$ for 0011, so $t(\mathbf{z}, \mathbf{r})$ increases steadily along upward paths in (2.10). If, instead, $t(\mathbf{z}, \mathbf{r})$ were the difference between the mean response in treated and control groups, it would again be increasing along upward paths.

Suppose, instead, that r_2 and r_3 were tied, so $r_1 > r_2 = r_3 > r_4$. In this case, the rank sum statistic would give average rank 2.5 to both r_2 and r_3, so moving from 1100 to 1010 would not change $t(\mathbf{z}, \mathbf{r})$. Notice, however, that even with ties, $t(\mathbf{z}, \mathbf{r})$ is monotone increasing (i.e., nondecreasing) along upward paths.

Actually, the order in (2.10) applies to many statistics whether ties are present or not. If $t(\mathbf{z}, \mathbf{r})$ is any arrangement-increasing statistic, then $t(\mathbf{z}, \mathbf{r})$ is monotone-increasing on upward paths in (2.10). Most reasonable statistics will assign a higher value to 1100 than to 1010, but reasonable statistics can differ in how they order assignments that are not comparable like 1001 and 0110.

Take a look at a second example, the case of $S = 2$ matched pairs, so $n_s = 2$ and $m_s = 1$ for $s = 1, 2$. Then Ω contains $2^2 = 4$ treatment assignments $\mathbf{z} = (z_{11}, z_{12}, z_{21}, z_{22})$. Again, assume the null hypothesis of no treatment effect and renumber the units in each pair so that in the first pair $r_{11} \geq r_{12}$, and in the second pair $r_{21} \geq r_{22}$. The set Ω appears in (2.11).



(2.11)

The assignment $\mathbf{z}$ in (2.11) suggesting the largest positive treatment effect is $\mathbf{z} = (1, 0, 1, 0)$ since in both pairs the treated unit had a higher response than the control. For $\mathbf{z} = 1001$ and $\mathbf{z} = 0110$, the treated unit had the higher response in one pair and the lower response in the other. In the assignment $\mathbf{z} = 0101$ the treated unit had a lower response than the control in both pairs.

Once again, common statistics are monotone-increasing along upward paths in (2.11). For instance, this is true of the signed rank statistic, which equals zero at the bottom of (2.11), equals one or two in the middle, and equals three at the top. Indeed, all arrangement-increasing functions are monotone-increasing along upward paths in (2.11).

What does all this suggest? There are certain treatment assignments $\mathbf{z} \in \Omega$ that are higher than others, and this is true without reference to the nature of the response $\mathbf{r}$ or the specific test statistic $t(\mathbf{z}, \mathbf{r})$. The responses might be continuous or they might be discrete scores or they might be binary. The test statistic might be the signed rank statistic or the McNemar

statistic. In all these cases, $\mathbf{z} = 1010$ is higher than $\mathbf{z} = 1001$ in (2.11). Certain statements about treatment assignments $\mathbf{z} \in \Omega$ should be true generally, without reference to the specific nature of the outcome or the test statistic.

*2.10.2 *A Brief Review of Lattices*

Briefly, a lattice is a partially ordered set in which each pair of elements has a greatest lower bound and a least upper bound. This terminology is discussed formally in a moment, but first consider what this means in (2.10). A point $\mathbf{z}$ in (2.10) is below another $\mathbf{z}^*$ if there is a path up from $\mathbf{z}$ to $\mathbf{z}^*$; for instance, 0110 is below 1100. The points 1001 and 0110 are not comparable—there is not a path up from one to the other—so Ω is partially but not totally ordered. The least upper bound of 0110 and 1001 is 1010, for it is the smallest element above both of them. The least upper bound of 1010 and 1100 is 1100. A nice introduction to lattices is given by MacLane and Birkoff (1988).

A set Ω is *partially ordered* by a relation $\lesssim$ if for all $\mathbf{z}$, $\mathbf{z}^*$, $\mathbf{z}^{**} \in \Omega$:

(i) $\mathbf{z} \lesssim \mathbf{z}$;

(ii) $\mathbf{z} \lesssim \mathbf{z}^*$ and $\mathbf{z}^* \lesssim \mathbf{z}$ implies $\mathbf{z} = \mathbf{z}^*$; and

(iii) $\mathbf{z} \lesssim \mathbf{z}^*$ and $\mathbf{z}^* \lesssim \mathbf{z}^{**}$ implies $\mathbf{z} \lesssim \mathbf{z}^{**}$.

An *upper bound* for $\mathbf{z}$, $\mathbf{z}^* \in \Omega$ is an element $\mathbf{z}^{**}$ such that $\mathbf{z} \lesssim \mathbf{z}^{**}$ and $\mathbf{z} \lesssim \mathbf{z}^{**}$. A *least upper bound* $\mathbf{z}^{**}$ for $\mathbf{z}, \mathbf{z}^*$ is an upper bound that is below all other upper bounds for $\mathbf{z}, \mathbf{z}^*$; that is, if $\mathbf{z}^{***}$ is any upper bound for $\mathbf{z}, \mathbf{z}^*$, then $\mathbf{z}^{**} \lesssim \mathbf{z}^{***}$. If a least upper bound for $\mathbf{z}, \mathbf{z}^*$ exists, then it is unique by (ii). Lower bound and *greatest lower bound* are defined similarly. A *lattice* is a partially ordered set Ω in which every pair $\mathbf{z}$, $\mathbf{z}^*$ of elements has a least upper bound, written $\mathbf{z} \vee \mathbf{z}^*$, and a greatest lower bound, written $\mathbf{z} \wedge \mathbf{z}^*$. A lattice Ω is *finite* if the set Ω contains only finitely many elements. In (2.10), both 1010 and 1100 are upper bounds for the pair 1001 and 0110, but the least upper bound is $1001 \vee 0110 = 1010$.

The partial order $\lesssim$ and the operations $\vee$ and $\wedge$ are tied together by the following relationship: $\mathbf{z} \lesssim \mathbf{z}^*$ if and only if $\mathbf{z} \vee \mathbf{z}^* = \mathbf{z}^*$ and $\mathbf{z} \wedge \mathbf{z}^* = \mathbf{z}$. In fact, using this relationship, a lattice may be defined beginning with the operations $\vee$ and $\wedge$ rather than beginning with the partial order $\lesssim$, that is, defining the partial order in terms of the operations. The following theorem is well known; see MacLane and Birkoff (1988, §XIV, 2) for proof.

Theorem 7 *A set Ω with operations $\vee$ and $\wedge$ is a lattice if and only if for all $\mathbf{z}$, $\mathbf{z}^*$, $\mathbf{z}^{**} \in \Omega$:*

L1. $\mathbf{z} \vee \mathbf{z} = \mathbf{z}$ *and* $\mathbf{z} \wedge \mathbf{z} = \mathbf{z}$;

L2. $\mathbf{z} \vee \mathbf{z}^* = \mathbf{z}^* \vee \mathbf{z}$ *and* $\mathbf{z} \wedge \mathbf{z}^* = \mathbf{z}^* \wedge \mathbf{z}$*;*

L3. $\mathbf{z} \vee (\mathbf{z}^* \vee \mathbf{z}^{**}) = (\mathbf{z} \vee \mathbf{z}^*) \vee \mathbf{z}^{**}$ *and* $\mathbf{z} \wedge (\mathbf{z}^* \wedge \mathbf{z}^{**}) = (\mathbf{z} \wedge \mathbf{z}^*) \wedge \mathbf{z}^{**}$*; and*

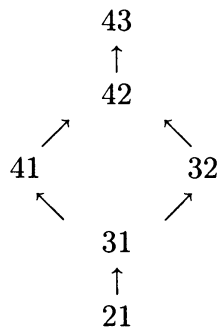
L4. $\mathbf{z} \wedge (\mathbf{z} \vee \mathbf{z}^*) = \mathbf{z} \vee (\mathbf{z} \wedge \mathbf{z}^*) = \mathbf{z}$.

Here, L2 and L3 are the commutative and associate laws, L1 is called idempotence, and L4 is called absorption. A lattice is *distributive* if the distributive law also holds,

$$\mathbf{z} \vee (\mathbf{z}^* \wedge \mathbf{z}^{**}) = (\mathbf{z} \vee \mathbf{z}^*) \wedge (\mathbf{z} \vee \mathbf{z}^{**}) \qquad \text{for all} \quad \mathbf{z}, \mathbf{z}^*, \mathbf{z}^{**} \in \Omega.$$

*2.10.3 *The Set of Treatment Assignments Is a Distributive Lattice*

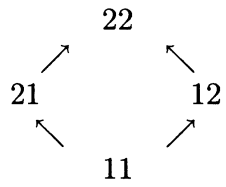
This section gives Savage's (1964) demonstration that Ω is a distributive lattice. With each N-dimensional $\mathbf{z} \in \Omega$, associate a vector $\mathbf{c}$ of dimension $\sum m_s$, as follows. The vector $\mathbf{c}$ is made up of S pieces, where piece s has m_s coordinates. It is suggestive and almost accurate to say that $\mathbf{c}$ contains the ranks of the responses of treated units, each stratum being ranked separately, the ranks being arranged in decreasing order in each stratum. This would be exactly true if there were no ties, but it is not exactly true in the case of ties. Here is the exact definition, with or without ties. If $z_{s1} = 0$, $z_{s2} = 0, \ldots, z_{s,i-1} = 0$, $z_{si} = 1$, then $c_{s1} = n_s - i + 1$. Continuing, if $z_{s,i+1} = 0, \ldots, z_{s,j-1} = 0$, $z_{sj} = 1$, then $c_{s2} = n_s - j + 1$, and so on. In terms of the $\mathbf{c}$, (2.10) becomes (2.12), and (2.11) becomes (2.13). For instance, in (2.10), $\mathbf{z} = 1100$ becomes $\mathbf{c} = 43$, since the first 1 in $\mathbf{z}$ appears in position $i = 1$, so $n - i + 1 = 4 - 1 + 1 = 4$ and the second 1 in $\mathbf{z}$ appears in position $j = 2$, so $n - j + 1 = 4 - 2 + 1 = 3$.



(2.12)

If there are ties among the responses in a stratum, then $\mathbf{c}$ is no longer a collection of ranks, because $\mathbf{c}$ distinguishes units with the same tied response. In the end, this is not a problem. The lattice order makes a few distinctions among treatment assignments that statistical procedures will

ignore.



(2.13)

It is readily checked that each $\mathbf{z}$ has one and only one corresponding $\mathbf{c}$. Given $\mathbf{z}$, $\mathbf{z}^* \in \Omega$, with corresponding $\mathbf{c}$ and $\mathbf{c}^*$, the operations $\vee$ and $\wedge$ are defined as follows. Define $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$ as the vectors containing, respectively, $\max(c_{si}, c^*_{si})$ and $\min(c_{si}, c^*_{si})$. Define $\mathbf{z} \vee \mathbf{z}^*$ and $\mathbf{z} \wedge \mathbf{z}^*$ as the elements of Ω corresponding to $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$. It is readily checked that this definition makes sense, that is, that $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$ always correspond to elements of Ω. For instance, in (2.10), $\mathbf{z} = 0110$ and $\mathbf{z}^* = 1001$ correspond to $\mathbf{c} = 32$ and $\mathbf{c}^* = 41$, so $\mathbf{c} \vee \mathbf{c}^* = 42$ and $\mathbf{c} \wedge \mathbf{c}^* = 31$, so $\mathbf{z} \vee \mathbf{z}^* = 1010$ and $\mathbf{z} \wedge \mathbf{z}^* = 0101$, as is consistent with (2.10). Notice carefully that the coordinate (s, i) of $\mathbf{z} \vee \mathbf{z}^*$ is not generally equal to $\max(z_{si}, z^*_{si})$.

To show that Ω is a lattice with these operations, one needs to check L1 to L4 in Theorem 7, but L1 to L3 hold trivially for $\max(c_{si}, c^*_{si})$ and $\min(c_{si}, c^*_{si})$. To show $\mathbf{z} \wedge (\mathbf{z} \vee \mathbf{z}^*) = \mathbf{z}$ in L4, it suffices to show $\mathbf{c} \wedge (\mathbf{c} \vee \mathbf{c}^*) = \mathbf{c}$. If $c_{si} \geq c^*_{si}$, then $\min\{c_{si}, \max(c_{si}, c^*_{si})\} = \min(c_{si}, c_{si}) = c_{si}$, while if $c_{si} < c^*_{si}$, then $\min\{c_{si}, \max(c_{si}, c^*_{si})\} = \min(c_{si}, c^*_{si}) = c_{si}$, so $\mathbf{c} \wedge (\mathbf{c} \vee \mathbf{c}^*) = \mathbf{c}$ as required. The second part of L4 is proved in the same way. So Ω is a lattice.

More than this, Ω is a distributive lattice. As proof, it suffices to show $\mathbf{c} \vee (\mathbf{c}^* \wedge \mathbf{c}^{**}) = (\mathbf{c} \vee \mathbf{c}^*) \wedge (\mathbf{c} \vee \mathbf{c}^{**})$, that is, to show

$$\max\{c_{si}, \min(c^*_{si}, c^{**}_{si})\} = \min\{\max(c_{si}, c^*_{si}), \max(c_{si}, c^{**}_{si})\}.$$

There are two cases. If $c_{si} \geq \min(c^*_{si}, c^{**}_{si})$, then $\max\{c_{si}, \min(c^*_{si}, c^{**}_{si})\} = c_{si}$, but also c_{si} is less than or equal to both $\max(c_{si}, c^*_{si})$ and $\max(c_{si}, c^{**}_{si})$ yet it equals one of them, so

$$\min\{\max(c_{si}, c^*_{si}), \max(c_{si}, c^{**}_{si})\} = c_{si}.$$

On the other hand, if $c_{si} < \min(c^*_{si}, c^{**}_{si})$, then

$$\max\{c_{si}, \min(c^*_{si}, c^{**}_{si})\} = \min(c^*_{si}, c^{**}_{si}),$$

but $\max(c_{si}, c^*_{si}) = c^*_{si}$, and $\max(c_{si}, c^{**}_{si}) = c^{**}_{si}$, so

$$\min\{\max(c_{si}, c^*_{si}), \max(c_{si}, c^{**}_{si})\} = \min(c^*_{si}, c^{**}_{si}),$$

as required to complete the proof.

2.10.4 *Inequalities for Probability Distributions on a Lattice

This section discusses two inequalities for probability distributions on a finite distributive lattice, namely, the FKG inequality and Holley's inequality. These inequalities are the principal tool that makes use of the lattice properties of Ω. The original proofs of these inequalities are somewhat involved, but Ahlswede and Daykin (1978) developed a simpler proof involving nothing more than elementary probability. Their proof is nicely presented in several recent texts (Anderson 1987, §6, Bollobas, 1986, §19), to which the reader may refer.

A real-valued function on Ω, $f : \Omega \to \mathbb{R}$ is isotonic if $\mathbf{z} \lesssim \mathbf{z}^*$ implies $f(\mathbf{z}) \leq f(\mathbf{z}^*)$. Throughout this appendix, $\mathbf{r}$ has been sorted into order within each stratum, $r_{si} \geq r_{s,i+1}$ for each s, i. With this order, the arrangement-increasing statistics $t(\mathbf{z}, \mathbf{r})$ are some of the isotonic functions on Ω. Actually, the arrangement-increasing statistics are the interesting isotonic functions, for they are the isotonic functions that are unchanged by interchanging tied responses in the same stratum. If there are ties, that is, if $r_{si} = r_{s,i+1}$ for some s and i, then there are isotonic functions that are not arrangement-increasing, specifically functions that increase when $z_{si} = 0, z_{s,i+1} = 1$ is replaced by $z_{si} = 1, z_{s,i+1} = 0$; however, these functions are not interesting as test statistics $t(\mathbf{z}, \mathbf{r})$ because they distinguish between people who gave identical responses. From a practical point of view, the important point is that a property of all isotonic functions on Ω is automatically a property of all arrangement-increasing functions, and all of the statistics in §2.4.3 are arrangement-increasing.

The first inequality is due to Fortuin, Kasteleyn, and Ginibre (1971).

Theorem 8 (The FKG Inequality) *Let $f(\cdot)$ and $g(\cdot)$ be isotonic functions on a finite distributive lattice Ω. If a random element $\mathbf{Z}$ of Ω is selected by a probability distribution satisfying*

$$\text{prob}(\mathbf{Z} = \mathbf{z} \vee \mathbf{z}^*) \cdot \text{prob}(\mathbf{Z} = \mathbf{z} \wedge \mathbf{z}^*) \geq \text{prob}(\mathbf{Z} = \mathbf{z}) \cdot \text{prob}(\mathbf{Z} = \mathbf{z}^*)$$
$$\textit{for all} \quad \mathbf{z}, \mathbf{z}^* \in \Omega,$$

then

$$\text{cov}\{f(\mathbf{Z}), g(\mathbf{Z})\} \geq 0.$$

For example, randomization gives equal probabilities to all elements of Ω, so the randomization distribution satisfies the condition for the FKG inequality. Hence, under the null hypothesis of no effect in a randomized experiment, any two arrangement-increasing statistics have a nonnegative correlation.

The next theorem is due to Holley (1974).

Theorem 9 (Holley's Inequality) *Let $f(\cdot)$ be an isotonic function on a finite distributive lattice Ω. If $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are random elements of Ω selected*

by two probability distributions satisfying

$$\text{prob}(\mathbf{Z} = \mathbf{z} \vee \mathbf{z}^*) \cdot \text{prob}(\tilde{\mathbf{Z}} = \mathbf{z} \wedge \mathbf{z}^*) \geq \text{prob}(\mathbf{Z} = \mathbf{z}) \cdot \text{prob}(\tilde{\mathbf{Z}} = \mathbf{z}^*)$$
$$\textit{for all} \quad \mathbf{z}, \mathbf{z}^* \in \Omega,$$

then

$$E\{f(\mathbf{Z})\} \geq E\{f(\tilde{\mathbf{Z}})\}.$$

In other words, the premise of Holley's inequality is a sufficient condition for $\mathbf{Z}$ to be stochastically larger than $\tilde{\mathbf{Z}}$, in the sense that for every arrangement-increasing function $f(\cdot)$, the random variable $f(\mathbf{Z})$ has higher expectation than $f(\tilde{\mathbf{Z}})$. Holley's inequality helps later in comparing a nonrandom assignment of treatments to a random assignment. A related result is given by Krieger and Rosenbaum (1994). Literature related to Holley's inequality is reviewed in Rosenbaum (1999).

*2.10.5 *An Identity in Ω*

There is a useful identity in the set Ω of treatment assignments. The identity links $\vee$ and $\wedge$ to the addition of vectors, and therefore it is useful in verifying the conditions of the FKG inequality and Holley's inequality. It is true for this lattice, but not true generally for all lattices.

Lemma 10 *For all* $\mathbf{z}, \mathbf{z}^* \in \Omega$,

$$\mathbf{z} \vee \mathbf{z}^* + \mathbf{z} \wedge \mathbf{z}^* = \mathbf{z} + \mathbf{z}^*.$$

Proof. Fix a coordinate (s, i), so the task is to show $z_{si} + z^*_{si} = z_{\wedge si} + z_{\vee si}$, where $z_{\wedge si}$ and $z_{\vee si}$ are the (s, i) coordinates of $\mathbf{z} \wedge \mathbf{z}^*$ and $\mathbf{z} \vee \mathbf{z}^*$, respectively. Let $\mathbf{c}$ and $\mathbf{c}^*$ correspond with $\mathbf{z}$ and $\mathbf{z}^*$, respectively. There are three cases, depending upon the value of $z_{si} + z^*_{si}$.

1. If $z_{si} + z^*_{si} = 0$, then $c_{sj} \neq n_s - i + 1$ and $c^*_{sj} \neq n_s - i + 1$ for $j = 1, \ldots, m_s$, so $\max(c_{sj}, c^*_{sj}) \neq n_s - i + 1$ and $\min(c_{sj}, c^*_{sj}) \neq n_s - i + 1$ for $j = 1, \ldots, m_s$, so $z_{\wedge si} + z_{\vee si} = 0$, as required.

2. If $z_{si} + z^*_{si} = 2$, then there is a j and a k such that $c_{sj} = n_s - i + 1$ and $c^*_{sk} = n_s - i + 1$. If $j = k$, then $\max(c_{sj}, c^*_{sj}) = n_s - i + 1$ and $\min(c_{sj}, c^*_{sj}) = n_s - i + 1$, so $z_{\wedge si} = 1$ and $z_{\vee si} = 1$, so that $z_{\wedge si} + z_{\vee si} = 2$, as required. If $j < k$, then $n_s - i + 1 = c_{sj} > c_{sk}$ and $c^*_{sj} > c^*_{sk} = n_s - i + 1$, so $\min(c_{sj}, c^*_{sj}) = c_{sj} = n_s - i + 1$ and $\max(c_{sk}, c^*_{sk}) = c^*_{sk} = n_s - i + 1$, so $z_{\wedge si} = 1$ and $z_{\vee si} = 1$, so that $z_{\wedge si} + z_{\vee si} = 2$, as required. The case $j > k$ is similar.

3. If $z_{si} = 1$ and $z^*_{si} = 0$, so $z_{si} + z^*_{si} = 1$, then there is a j such that $c_{sj} = n_s - i + 1$ but $c^*_{sk} \neq n_s - i + 1$ for $k = 1, \ldots, m_s$. In this case, either $n_s - i + 1 = \max(c_{sj}, c^*_{sj})$ or $n_s - i + 1 = \min(c_{sj}, c^*_{sj})$ but not both, and moreover, $n_s - i + 1 \neq \max(c_{sk}, c^*_{sk})$ and $n_s - i + 1 \neq \min(c_{sk}, c^*_{sk})$ for all

$k \neq j$, so $z_{\wedge si} + z_{\vee si} = 1$, as required. The case $z_{si} = 0$ and $z^*_{si} = 1$ is similar. ■

If there were no ties, so $\mathbf{c}$ and $\mathbf{c}^*$ are ranks, then Lemma 10 has the following interpretation. Within each stratum, the operations $\vee$ and $\wedge$ take the ranks in $\mathbf{c}$ and $\mathbf{c}^*$ and apportion them in forming $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$, but in this process they do not create or delete ranks that appear in $\mathbf{c}$ and $\mathbf{c}^*$.

2.11 Bibliographic Notes

Fisher is usually credited with the invention of randomized experiments. See, in particular, his important and influential book, *The Design of Experiments*, first published in 1935. Randomization is discussed in many articles and textbooks. In particular, see Kempthorne (1952), Cox (1958a, §5) and Cox and Reid (2000) for discussions of randomization in experimental design, and see Lehmann (1975) and Maritz (1981) for discussions of its role in nonparametrics. Mantel's (1963) paper was significant not just for the method he proposed, but also for its strengthening of the link between nonparametric methods and contingency table methods. The model for a treatment effect in §2.5.2 in which each unit has two potential responses, one under treatment and the other under control, has a long history. In an article first published in Polish and recently translated into English, Neyman (1923) used it to study the behavior of statistical tests under random assignment of treatments. Related work was done by Welch (1937), Wilk (1955), Cox (1958b, §5), and Robinson (1973), among others. Rubin (1974, 1977) first used the model in observational studies. In particular, he discussed the conditions under which matching, stratification, and covariance adjustment all estimate the same treatment effect. See also Hamilton (1979) and Holland (1986). Arrangement-increasing functions have been studied under various names by Eaton (1967), Hollander, Proshan, and Sethuraman (1977), and Marshall and Olkin (1979, §6F); see also Savage (1957). Although the Hodges-Lehmann (1963) estimates are often derived from rank tests, these *R-estimates* are very closely related to other families of estimates based on order statistics, *L-estimates*, or based on solving equations, *M-estimates*; see Gastwirth (1966) and Jureckova (1984). An attraction of R-estimates over L-estimates or M-estimates is that R-estimates have associated tests and confidence intervals that are exact, nonparametric, and explicitly linked to randomization in experiments. Sign-score statistics are discussed in Rosenbaum (1988) in connection with sensitivity analysis where these statistics permit certain simplifications. The discussion of complex outcomes in §2.8 draws from Mann and Whitney (1947), Gehan (1965), Mantel (1967), and Rosenbaum (1991, 1994). The material in §2.10 uses ideas from Savage (1964) and Rosenbaum (1989, 1995). The results in §2.10 concern permutations of vectors with binary coordinates, but some

of these results extend to permutations of vectors with real coordinates; see Krieger and Rosenbaum (1994).

2.12 Problems

1. The surprising power of the Lady tasting tea. In §2.2, what is the power of the test? Specifically, suppose the Lady can distinguish milk first from tea first, and is always accurate. What is the power of a one-sided, 0.05 level test? Which 2×2 tables of the form Table 2.2 lead to rejection at the 0.05 level? If the Lady can distinguish, what is the chance of a table that leads to rejection?

2. Interference between units with longitudinal data. Suppose that there are S people, $s = 1, \ldots, S$, and person s is measured once a week for n_s consecutive weeks, $i = 1, \ldots, n_s$. Here, one unit (s, i) is one person in one week. For person s, a fixed number, m_s, of weeks are picked at random, independently for different people, and person s is treated in those weeks. Write $Z_{si} = 1$ if person s is treated in week i, $Z_{si} = 0$ otherwise, so $m_s = \sum_{i=1}^{n_s} Z_{si}$. The observed response of person s in week i is R_{si}, which may be affected by the current treatment Z_{si} and previous treatments, Z_{sj}, $j = 1, \ldots, i$. In addition, person s has a pretreatment baseline response, R_{s0}, which is unaffected by treatment, and so is fixed. Consider the model $R_{si} - R_{s,i-1} = \eta_{si} + \Delta Z_{si}$ for $i = 1, \ldots, n_s$, so the treatment produces additive gains, where Δ and the η_{si} are unknown fixed parameters. Show that this model violates the condition of "no interference between units" in §2.5.2. Let $T = t(\mathbf{Z}, \mathbf{R})$ be the stratified rank sum statistic, applied to the changes, $R_{si} - R_{s,i-1}$, so the n_s changes for person s are ranked from 1 to n_s and T is the sum of the ranks for the $\sum m_s$ treated weeks. Under the null hypothesis, $H_0 : \Delta = 0$, what is the randomization distribution of T? How does it compare to the usual randomization distribution of T of the stratified rank test? How could you use the randomization distribution of T when $\Delta = 0$ to test the general hypothesis $H_0 : \Delta = \Delta_0$? (Hint: Think about adjusted responses, $R_{si} - R_{s,i-1} - \Delta_0 Z_{si}$.) How could you use the randomization distribution of T when $\Delta = 0$ to build a confidence interval for Δ? Does interference between units preclude randomization inference?

3. Proof of Proposition 1. Let A and B be two finite, nonempty, disjoint sets, and let $A \times B$ be the set of all ordered pairs (a, b) with $a \in A$ and $b \in B$. If (a, b) is picked at random from $A \times B$, with each element of $A \times B$ having the same probability, show that a and b are independent. Use this to prove Proposition 1 for $S = 2$. Then use it

again to show that if Proposition 1 is true for S, then it is also true for $S+1$.

4. Proof of Proposition 2. Prove Proposition 2. (Hint: Why does

$$var\left(\sum_{s=1}^{S}\sum_{i=1}^{n_s} Z_{si}q_{si}\right) = \sum_{s=1}^{S} var\left(\sum_{i=1}^{n_s} Z_{si}q_{si}\right)?$$

Why does

$$var\left(\sum_{i=1}^{n_s} Z_{si}q_{si}\right) = var\left\{\sum_{i=1}^{n_s} Z_{si}\left(q_{si}-\bar{q}_s\right)\right\}?$$

Remember $q_{si}-\bar{q}_s$ is fixed. What is $E(Z_{si})$? What is

$$E\left\{\sum_{i=1}^{n_s} Z_{si}(q_{si}-\bar{q}_s)\right\}?$$

What is $E(Z_{si}Z_{sj})$? Be careful about $i=j$ and $i\neq j$.)

5. Different statistics that yield the same randomization test. Let $f(\cdot)$ be a strictly increasing function, so $x<y$ implies $f(x)<f(y)$. Show that a test that rejects at level α when $t(\mathbf{Z},\mathbf{R})\geq k$ is exactly the same test as the test that rejects when $f\{t(\mathbf{Z},\mathbf{R})\}\geq f(k)$. In a uniform randomized experiment with a single stratum, $S=1$, dropping the s subscript, show that a randomization test of no treatment effect based on the total in the treated group, $\sum Z_iR_i$, is exactly the same test as a randomization test based on the difference between the treated and control group means,

$$t(\mathbf{Z},\mathbf{R}) = \frac{\sum Z_iR_i}{m} - \frac{\sum(1-Z_i)R_i}{n-m}.$$

In a uniform randomized experiment with a single stratum, $S=1$, what is the Hodges—Lehmann estimate of an additive treatment effect, $r_{Ti}=r_{Ci}+\tau$ obtained from taking $t(\mathbf{Z},\mathbf{R})$ to be the difference between the treated and control group means?

6. An effect increasing statistic with partially ordered responses. Show that the statistic (2.8) is effect increasing. (Hint: Consider two response vectors, $\mathbf{r}$ and $\mathbf{r}^*$, and the corresponding indicators, L_{sij} and L^*_{sij}.)

7. Metaphysics. Section 2.5.3 discussed the distribution of observable quantities (Z_{si},R_{si}) in a uniform randomized experiment under the model of an additive treatment effect, $r_{Tsi}=r_{Csi}+\tau$. Because (r_{Tsi},r_{Csi}) is not jointly observed, one sees only $R_{si}=r_{Tsi}$ if $Z_{si}=1$ for a treated subject, or else one sees $R_{si}=r_{Csi}$ if $Z_{si}=0$ for a

control subject. Consider the case of a single stratum, $S = 1$, dropping the subscript s, and recall that, in a completely randomized experiment, the observable consequence of the additive effect model, $r_{Ti} = r_{Ci} + \tau$, is that the distribution of treated and control responses have the same shape and dispersion, but different locations, so the treated distribution is shifted by τ. Does the additive model $r_{Tsi} = r_{Csi} + \tau$ have content beyond its implications for observable distributions? Keep in mind that this is a problem in metaphysics, not statistics, so perhaps there is an answer, perhaps not. Hint: It is reasonable to ask of a question whether it is a reasonable question to ask. What does the phrase "content beyond" mean in this question? If "content beyond" were replaced by "observable consequences," what becomes of the question? If "content beyond" were replaced by "a mathematical form different from," what becomes of the question? In parallel, Wittgenstein (1958, #47, p22-23) writes:

> To the *philosophical* question: "Is the visual image of this tree composite, and what are its component parts?" the correct answer is "That depends upon what you understand by 'composite'." (And that is of course not an answer but a rejection of the question.)

2.13 References

Ahlswede, R. and Daykin, D. (1978) An inequality for the weights of two families of sets, their unions, and intersections. *Z. Wahrsch. Verus Gebiete*, **43**, 183–185.

Anderson, I. (1987) *Combinatorics of Finite Sets.* New York: Oxford University Press.

Birch, M. W. (1964) The detection of partial association, I: The 2×2 case. *Journal of the Royal Statistical Society*, Series **B**, **26**, 313–324.

Birch, M. W. (1965) The detection of partial association, II: The general case. *Journal of the Royal Statistical Society*, Series **B**, **27**, 111–124.

Bollobas, B. (1986) *Combinatorics.* New York: Cambridge University Press.

Campbell, D. and Stanley, J. (1963) *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Cochran, W. G. (1963) *Sampling Techniques.* New York: Wiley.

Cox, D. R. (1958a) *Planning of Experiments.* New York: Wiley.

Cox, D. R. (1958b) The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, **45**, 69–73.

Cox, D. R. (1966) A simple example of a comparison involving quantal data. *Biometrika*, **53**, 215–220.

Cox, D. R. (1970) *The Analysis of Binary Data*. London: Methuen.

Cox, D. R. and Hinkley, D.V. (1974) *Theoretical Statistics*. London: Chapman & Hall.

Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments*. New York: CRC Press.

Eaton, M. (1967) Some optimum properties of ranking procedures. *Annals of Mathematical Statistics*, **38**, 124–137.

Eaton, M. (1982) A review of selected topics in probability inequalities. *Annals of Statistics*, **10**, 11–43.

Eaton, M. (1987) *Lectures on Topics in Probability Inequalities*. Amsterdam: Centrum. voor Wiskunde en Informatica.

Efron, B. (1971) Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.

Fisher, R. A. (1935, 1949) *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fortuin, C., Kasteleyn, P., and Ginibre, J. (1971) Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, **22**, 89–103.

Freidlin, B. and Gastwirth, J. L. (2000) Should the median test be retired from general use? *American Statistician*, **54**, 161–164.

Friedman, L. M., DeMets, D. L., and Furberg, C. D. (1998) *Fundamentals of Clinical Trials*. New York: Springer-Verlag.

Gastwirth, J. L. (1966) On robust procedures. *Journal of the American Statistical Association*, **61**, 929-948.

Gehan, E. (1965) A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203–223.

Gibbons, J. D. (1982) Brown-Mood median test. In: *Encyclopedia of Statistical Sciences*, Volume 1, S. Kotz and N. Johnson, eds., New York: Wiley, pp. 322–324.

Hamilton, M. (1979) Choosing a parameter for 2×2 table or $2 \times 2 \times 2$ table analysis. *American Journal of Epidemiology*, **109**, 362–375.

Hettmansperger, T. (1984) *Statistical Inference Based on Ranks.* New York: Wiley.

Hodges, J. and Lehmann, E. (1962) Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics*, **33**, 482–497.

Hodges, J. and Lehmann, E. (1963) Estimates of location based on rank tests. *Annals of Mathematical Statistics*, **34**, 598–611.

Holland, P. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–970.

Hollander, M., Proschan, F., and Sethuraman, J. (1977) Functions decreasing in transposition and their applications in ranking problems. *Annals of Statistics*, **5**, 722–733.

Hollander, M. and Wolfe, D. (1973) *Nonparametric Statistical Methods.* New York: Wiley.

Holley, R. (1974) Remarks on the FKG inequalities. *Communications in Mathematical Physics*, **36**, 227–231.

Jureckova, J. (1984) M-, L- and R-estimators. In: *Handbook of Statistics*, Volume IV, P. R. Krishnaiah and P. K. Sen, eds., New York: Elsevier, pp. 463–485.

Kempthorne, O. (1952) *The Design and Analysis of Experiments.* New York: Wiley.

Krieger, A. M. and Rosenbaum, P. R. (1994) A stochastic comparison for arrangement increasing functions. *Combinatorics, Probability and Computing*, **3**, 345–348.

Lehmann, E. L. (1959) *Testing Statistical Hypotheses.* New York: Wiley.

Lehmann, E. L. (1975) *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

MacLane, S. and Birkoff, G. (1988) *Algebra.* New York: Chelsea.

Mann, H. and Whitney, D. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, **18**, 50–60.

Mantel, N. (1963) Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association*, **58**, 690–700.

Mantel, N. (1967) Ranking procedures for arbitrarily restricted observations. *Biometrics*, **23**, 65–78.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

Maritz, J. (1981) *Distribution-Free Statistical Methods*. London: Chapman & Hall.

Marshall, A. and Olkin, I. (1979) *Inequalities: Theory of Majorization and Its Applications*. New York: Academic.

McNemar, Q. (1947) Note on the sampling error of the differences between correlated proportions or percentage. *Psychometrika*, **12**, 153–157.

Murphy, M., Hultgren, H., Detre, K., Thomsen, J., and Takaro, T. (1977) Treatment of chronic stable angina: A preliminary report of survival data of the randomized Veterans Administration Cooperative study. *New England Journal of Medicine*, **297**, 621–627.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (In Polish) *Roczniki Nauk Roiniczych, Tom X*, pp. 1–51. Reprinted in *Statistical Science 1990*, **5**, 463–480, with discussion by T. Speed and D. Rubin.

Neyman, J. (1935) Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, **2**, 107–180.

Pagano, M. and Tritchler, D. (1983) Obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association*, **78**, 435–440.

Robinson, J. (1973) The large sample power of permutation tests for randomization models. *Annals of Statistics*, **1**, 291–296.

Rosenbaum, P. R. (1988) Sensitivity analysis for matching with multiple controls. *Biometrika*, **75**, 577–581.

Rosenbaum, P. R. (1989) On permutation tests for hidden biases in observational studies: An application of Holley's inequality to the Savage lattice. *Annals of Statistics*, **17**, 643–653.

Rosenbaum, P. R. (1991) Some poset statistics. *Annals of Statistics*, **19**, 1091–1097.

Rosenbaum, P. R. (1994) Coherence in observational studies. *Biometrics*, **50**, 368–374.

Rosenbaum, P. R. (1995) Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association*, **90**, 1424–1431.

Rosenbaum, P. R. (1999) Holley's inequality. *Encyclopedia of Statistical Sciences*, Update Volume **3**, S. Kotz, C. B. Read, D. L. Banks, eds., New York: Wiley, pp. 328–331.

Rubin, D. B. (1974) Estimating the causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Rubin, D. B. (1977) Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1–26.

Rubin, D. B. (1986) Which ifs have causal answers? *Journal of the American Statistical Association*, **81**, 961–962.

Savage, I. R. (1957) Contributions to the theory of rank order statistics: The trend case. *Annals of Mathematical Statistics*, **28**, 968–977.

Savage, I. R. (1964) Contributions to the theory of rank order statistics: Applications of lattice theory. *Review of the International Statistical Institute*, **32**, 52–63.

Tukey, J. W. (1985) Improving crucial randomized experiments—especially in weather modification—by double randomization and rank combination. In: *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, L. Le Cam and R. Olshen, eds., Volume 1, Belmont, CA: Wadsworth, pp. 79–108.

Welch, B. L. (1937) On the z-test in randomized blocks and Latin squares. *Biometrika*, **29**, 21–52.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics*, **1**, 8083.

Wilk, M. B. (1955) The randomization analysis of a generalized randomized block design. *Biometrika*, **42**, 70–79.

Wittgenstein, L. (1958) *Philosophical Investigations* (Third Edition). Englewood Cliffs, NJ: Prentice-Hall.

Zelen, M. (1974) The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, **27**, 365–375.

3

Overt Bias in Observational Studies

3.1 Introduction: An Example and Planning Adjustments

3.1.1 Outline: When Can Methods for Randomized Experiments Be Used?

An observational study is biased if the treated and control groups differ prior to treatment in ways that matter for the outcomes under study. An overt bias is one that can be seen in the data at hand—for instance, prior to treatment, treated subjects are observed to have lower incomes than controls. A hidden bias is similar but cannot be seen because the required information was not observed or recorded. Overt biases are controlled using adjustments, such as matching or stratification. In other words, treated and control subjects may be seen to differ in terms of certain observed covariates, but these visible differences may be removed by comparing treated and control subjects with the same values of the observed covariates, that is, subjects in the same matched set or stratum defined by the observed covariates. It is natural to ask when the standard methods for randomized experiments may be applied to matched or stratified data from an observational study. This chapter discusses a model for an observational study in which there is overt bias but no hidden bias. The model is, at best, one of many plausible models, but it does clarify when methods for randomized experiments may be used in observational studies, and so it becomes the starting point for thinking about hidden biases. Dealing with hidden bias is

the focus of most of the later chapters. To permit discussion of conceptual issues in this chapter, Chapter 10 discusses the algorithmic issues that arise in constructing matched sets or strata with many covariates. The remainder of §3.1 considers an example and then discusses some of the planning steps that precede adjustments for covariates.

3.1.2 An Example with a Single Covariate

Cochran (1968) presents three stark examples of overt biases and their removal through adjustments. We will look at one of these. The data are from a study by Best and Walker of mortality in three groups of men: nonsmokers, cigarette smokers, and cigar and pipe smokers. Nonsmokers had a mortality rate of 20.2 deaths per 1000 people per year, cigarette smokers had 20.5 deaths, and cigar and pipe smokers had 35.5 deaths. The naive interpretation would be that cigarettes are harmless, but either cigars or pipes or both are dangerous. Cochran then gives the mean age in each group: 54.9 years for nonsmokers, 50.5 years for smokers, and 65.9 for cigar and pipe smokers. Clearly, the cigar and pipe smokers are older, so their higher death rate is not surprising, and may not reflect an effect of cigars or pipes. On the other hand, the cigarette smokers are the youngest group, and yet their mortality rate is slightly higher than the somewhat older nonsmokers. Perhaps cigarettes are not harmless.

Cochran then adjusts mortality for age, that is, removes an overt bias in the outcome by adjusting for an imbalance in a covariate. He uses age to divide the men into three strata or subclasses so that men in the same stratum have similar ages. Nonsmokers, cigarette smokers, and cigar and pipe smokers of roughly the same age are then compared to each other within each stratum, and the results are combined into a single rate using direct adjustment, essentially as described in §2.7.1. The adjusted mortality rate is 20.3 deaths per 1000 per year for nonsmokers, 28.3 for cigarette smokers, and 21.2 for cigar and pipe smokers. Now it is cigarettes that appear dangerous.

Which rates should be trusted, unadjusted or adjusted? Neither. The unadjusted rates are clearly wrong as a basis for estimating the effects of smoking, for they compare men who are not comparable in terms of one of the most important features of human mortality, namely, age. The adjusted rates are not clearly wrong. They might estimate the effects of smoking. However, it is possible that there is another covariate that was not recorded that has an impact similar to age; in this case, there would be a hidden bias. The current chapter discusses the conditions under which the methods in Chapter 2 for randomized experiments successfully estimate treatment effects in observational studies. These conditions become the basis in later chapters for thinking about hidden biases.

Cochran used three age strata. One might reasonably ask whether three strata are sufficient, whether such broad age groups suffice to remove the

overt bias due to age, and indeed this is the main question in Cochran's paper. If instead of three strata, twelve strata are used, then the adjusted rates are 20.2 for nonsmokers, 29.5 for cigarette smokers, and 19.8 for cigar and pipe smokers. Three-age strata and twelve-age strata produce similar adjusted rates, both of which are very different from the rates prior to adjustment. Cochran presents a theoretical argument concluding that five strata, each containing 20% of the subjects, will remove about 90% of the bias in a single continuous covariate such as age.

3.1.3 Planning Adjustments for Overt Biases

Options narrow as an investigation proceeds. What is easy early on may become difficult or impossible later. This section discusses the earliest stages of planning and data collection, as they relate to adjustments for bias. The points raised are elementary, but at times ignored. When ignored, the problems created can be far from elementary, at times insurmountable.

The control of overt biases begins before the study is designed. A first step in planning an observational study is to determine what treatments will be studied, and in the process to distinguish outcomes from covariates. Outcomes measure quantities that may be affected by the treatment, while covariates are not affected; see §2.5. Cox (1958, §4.2) uses the term concomitant observations in place of covariate and writes:

> The essential point in our assumptions about these observations is that the value for any unit must be unaffected by the particular assignment of treatments to units actually used. In practice this means that either: (a) the concomitant observations are taken before the assignment of treatments to units is made; or (b) the concomitant observations are made after the assignment of treatments, but before the effect of treatments has had time to develop ... ; or (c) we can assume from our knowledge of the nature of the concomitant observations concerned, that they are unaffected by the treatment.

As an example of type (c), Cox mentions the covariate that records the relative humidity in a textile factory, where it is known that the treatments under study could not possibly affect the relative humidity.

If adjustments are not confined to covariates, then adjustments may remove part or all of the effect of the treatment. To illustrate, consider an extreme, hypothetical example. Imagine a study comparing a placebo and a drug intended to reduce blood pressure, the outcome being the incidence of stroke. If the groups were compared after adjustment for blood pressure levels six months after the start of treatment, then the adjusted incidence of stroke might be similar in drug and placebo groups, not because the drug has failed to work, but rather because the drug reduces the risk of stroke

by reducing blood pressure. If the effect of the drug on blood pressure is removed, the effect on stroke is removed with it.

While adjustments for an outcome can remove part of the treatment effect, adjustments of this sort are occasionally performed. It may be suspected that the treatment has only slight effects on a particular outcome, but this outcome may be strongly related to an important covariate that was not measured. An example occurred in the studies by Coleman, Hoffer, and Kilgore (1982) and Goldberger and Cain (1982) of the effects of Catholic versus public high schools. These studies compared cognitive test scores in the senior year of high school adjusting for various covariates, but the studies also adjusted for an outcome, namely, cognitive test scores in the sophomore year. The sophomore year test scores may already be affected by the difference between Catholic and public high schools, so they are, in principle, outcomes, not covariates. Still, it is natural to suspect that any effect of Catholic versus public high schools is produced gradually and cumulatively, and that only a part of the effect is present in the sophomore year. These studies used this outcome as a surrogate for an important covariate that was unavailable, namely, cognitive test scores prior to the start of high school. There are, then, two hazards: adjusting for sophomore test scores can remove part of the difference between Catholic and public schools; and failing to adjust for an early test score may yield a comparison of students who were not comparable in terms of their cognitive abilities prior to the start of high school. Notice that the second hazard is not present in a randomized experiment, so in an experiment, it is possible to give unequivocal advice that adjustments for outcomes should be avoided when estimating treatment effects. In an observational study, both hazards are present, and must be weighed; see Rosenbaum (1984a, §4) for discussion of alternative methods of analysis. The important point for the initial planning of observational studies is the distinction between outcomes and covariates, and their different status in adjustments.

The next step in planning is to list the covariates that will be measured. It is at this stage that biases become either overt or hidden. Since there is no way to completely address a hidden bias, a small change in this list may determine whether the study is convincing. A small oversight, easily corrected in the planning stage, may be an insurmountable problem at a later stage. In the design of randomized clinical trials, the standard practice is to begin with a written protocol that describes the data that will be collected and the main analyses that will be performed. Before the trial starts, the protocol is circulated for critical comment. Observational studies would, I believe, benefit from a written protocol and critical commentary.

Adjustments for overt biases may begin with data collection rather than with data analysis. Often treated subjects are matched to controls to form pairs or matched sets of subjects who are comparable in terms of observed covariates, and matching may take place before outcomes are measured. Chapter 10 discusses matching methods. Here, three points should be men-

tioned. First, unlike analytical adjustments, adjustments that are built into the study design are irrevocable. In the hypothetical example above concerning drug versus placebo to prevent stroke, it would be a mistake to adjust for blood pressure after treatment. If this mistake were made using an analytical method such as in §3.1.2, then it could be corrected by performing a different analysis, but if the mistake were made by matched sampling then it would be difficult to correct.

Second, certain covariates are more easily controlled through matching in the design than through analytical adjustments. Typically, these are covariates that classify subjects into many small categories. Matching can ensure that treated and control subjects belong to the same categories, but if matching is not used in the design of the study, some categories may have treated subjects and no controls or controls and no treated subjects. For instance, consider a study (Rosenbaum, 1986) that compared cognitive skills in what would be the senior year of high school for sophomores who dropped out of school and similar sophomores who remained in school. This was done with a national sample of students, and the high school was an important covariate with many values. The study used matched pairs of students *from the same school* having similar test scores, academic performance, and disciplinary records in the sophomore year, before the dropout left school.

Cost is an important consideration in deciding whether to match. If some covariate information is readily available, but other data are difficult or expensive to obtain, then matching becomes more attractive, but if data come with negligible costs, then matching during the design becomes somewhat less attractive. The reason is that, in many studies, some controls will be so different from treated subjects that they are of little use for comparisons (Dehejia and Wahba 1999). In the example above, many high-school students look very different from most dropouts in terms of test scores, academic performance, and disciplinary problems, so these students are of limited use in trying to determine how students who drop out would have performed had they remained in school. Matching may avoid collecting data on controls who will later be of little use.

A compromise between selecting matched pairs and using all potential controls is to match each treated subject to several controls. Ury (1975) examines the efficiency of studies that match several controls to each treated subject, finding that there is little to be gained from having more than four controls per treated subject with continuous responses. Smith (1997) presents an interesting case study of multivariate matching with multiple controls. In a single application, he compared pair matching with 1 control, matching with 8 controls and matching with 15 controls, concluding that 8 controls was best in this particular study. See also Ming and Rosenbaum (2000). Chapter 10 discusses the construction of matched sets with equal and variable numbers of controls per treated subject.

Matched studies can often be improved by a pilot study that forms a small number of matched pairs and scrutinizes those pairs using ethnographic or qualitative techniques. For instance, one might interview a few paired subjects or read the text of their hospital charts. This process may begin to reveal the hidden biases not visible in data on observed covariates, or it may suggest more accurate ways of using the data. An example is discussed in detail by Rosenbaum and Silber (2001). Emerson (1981) and Katz (2001) survey ethnographic techniques with reference to a large literature; see also Blumer (1969) and Becker (1996).

Having collected the data on covariates, the question arises: Should adjustments be made for all observed covariates? If not, how should covariates be selected for adjustment? These questions are somewhat controversial, not so much because the issues involved are unclear, but rather because there is no fully satisfactory answer. In principle, there is little or no reason to avoid adjustment for a true covariate, a variable describing subjects before treatment. There is little harm in comparing subjects who were comparable before treatment in ways that are not relevant for the outcomes of interest. In experiments, randomization tends to make treated and control groups comparable in terms of all covariates, relevant and irrelevant. In practice, the situation is often more involved, and increasing the number of covariates used in adjustments increases costs and complexities, and may make it more difficult to adjust for the most important covariates. In part, there are issues of data quality and completeness. As more covariates are collected and analyzed, it becomes increasingly difficult to ensure that all covariates meet high standards of accuracy and completeness, and increasingly difficult to ensure that each covariate receives the needed attention when used in modeling or matching. If there are many covariates, each with some missing data, there may be few subjects with complete data on all covariates, and this may make the analysis more difficult than it would otherwise be. These considerations weigh most heavily on covariates having doubtful relevance to outcomes of interest.

Perhaps the most common method for selecting covariates is also the most widely criticized. It entails comparing treated and control groups with respect to a long list of covariates, say using a t-test, and adjusting only for those covariates for which significant differences are found. There are three problems with this. First, the process does not consider the relationship between covariate and outcome. Second, there is no reason to believe that the absence of statistical significance implies the imbalance in the covariate is small enough to be ignored. Third, the process considers covariates one at a time, while the adjustments will control the covariates simultaneously. Addressing the first two problems, Cochran (1965, §3.1) studied this technique under a simple linear regression model in which all quantities are Normally distributed and a single covariate is the only source of bias. He looked at the coverage probability of the 95% confidence interval for the effect of the treatment on an outcome when no adjustment

had been made for the covariate. This coverage probability was 90% or more providing the t-statistic for the covariate was less than 1.5 in absolute value and providing the squared correlation between the covariate and the outcome was 0.5 or less. This limitation on the square correlation is often reasonable for a covariate whose relevance is in doubt. He concluded: "If a single [covariate] shows a value of t above 1.5, these results suggest that we have another look at this [covariate] when the values of the [outcome] become known." Canner (1984, 1991) discusses closely related issues.

The following approach is often reasonable and practical. Begin by selecting a tentative list of covariates for adjustments using scientific knowledge of the relevant covariates together with exploratory comparisons of covariates in the treated and control groups, perhaps including some version of the technique evaluated by Cochran (1965, §3.1). With this tentative list, determine the tentative method of adjustment; that is, select the matched pairs or sets, define the strata, and determine whatever modeling technique will be used. Apply this method of adjustment to the covariates excluded from the tentative list, identifying any covariates exhibiting a large imbalance after adjustment. Reconsider the tentative list of covariates in light of this analysis. This approach addresses, at least in part, each of the three problems in the previous paragraph. The focus is on covariates known to be relevant since they are included in the initial list. At the same time, the data are given several opportunities to call attention to imbalances that might not be anticipated. Examples along these lines are discussed by Rosenbaum and Rubin (1984) and Silber et al. (2001).

3.2 Adjustments by Exact Stratification and Matching

3.2.1 Treatment Assignment with Unknown Probabilities

When is an observational study free of hidden bias? When do adjustments such as matching and stratification remove all of the bias? This section describes a model for an observational study with overt but no hidden bias. In most observational studies, this model is, at best, one of many plausible models—hidden biases are possible. The model is a start, indicating the inferences that would be appropriate were hidden biases absent. Later chapters try to determine whether hidden biases are present and ask how inferences might change if they are.

Initially, there are M units available for study, and each has a value of an observed covariate $\mathbf{x}$, which may contain several variables. Often the covariates $\mathbf{x}$ are used to reorganize the data prior to analysis, for instance,

by matching or stratifying on $\mathbf{x}$. Number the M units $j = 1, \ldots, M$, so $\mathbf{x}_{[j]}$ is the covariate for the jth unit and the treatment assignment for this unit is $Z_{[j]}$. The bracketed subscript $[j]$ signifies the numbering of units before they are reorganized. After reorganization, a unit will have a different subscript without a bracket.

As a model for an observational study, imagine that unit j is assigned to treatment with probability $\pi_{[j]} = \text{prob}(Z_{[j]} = 1)$ and to control with probability $1 - \pi_{[j]} = \text{prob}(Z_{[j]} = 0)$, with assignments for distinct units being independent, and with $0 < \pi_{[j]} < 1$. The model says that treatments were assigned by flipping biased coins, possibly a different coin with a different bias for each unit, where the biases of the coins or the π's are unknown. The model says:

$$\text{prob}(Z_{[1]} = z_1, \ldots, Z_{[M]} = z_M) = \prod_{j=1}^{M} \pi_{[j]}^{z_j} \{1 - \pi_{[j]}\}^{1-z_j}. \tag{3.1}$$

In an observational study, $\pi_{[j]}$ is unknown, so the distribution of treatment assignments $Z_{[1]}, \ldots, Z_{[M]}$ is unknown, and it is not possible to draw inferences as in Chapter 2 where randomization created a known distribution of treatment assignments.

Consider now the model for an observational study with overt biases but no hidden biases. An observational study is *free of hidden bias* if the π's, though unknown, are known to depend only on the observed covariates $\mathbf{x}_{[j]}$, so two units with the same value of $\mathbf{x}$ have the same chance π of receiving the treatment. Formally, the study is free of hidden bias if there exists a function $\lambda(\cdot)$, whose form will typically be unknown, such that $\pi_{[j]} = \lambda(\mathbf{x}_{[j]})$ for $j = 1, \ldots, M$. If the study is free of hidden bias, then (3.1) becomes

$$\text{prob}(Z_{[1]} = z_1, \ldots, Z_{[M]} = z_M) = \prod_{j=1}^{M} \lambda(\mathbf{x}_{[j]})^{z_j} \{1 - \lambda(\mathbf{x}_{[j]})\}^{1-z_j}. \tag{3.2}$$

In short, an observational study is *free of hidden bias* when (3.2) holds. Rubin (1977) calls (3.2) "randomization on the basis of a covariate."

When the study is free of hidden bias, the function $\lambda(\mathbf{x})$ is called the propensity score. In §10.2, the propensity score $\lambda(\mathbf{x})$ is redefined so that it is still meaningful when hidden biases are present; however, in that case, $\pi_{[j]} \neq \lambda(\mathbf{x}_{[j]})$, and (3.2) does not follow from (3.1). A study is free of hidden bias when the treatment assignment probabilities $\pi_{[j]}$ are given by the propensity score $\lambda(\mathbf{x}_{[j]})$ which is always a function of the observed covariates $\mathbf{x}_{[j]}$. Chapter 4 discusses a model in which there is hidden bias and $\pi_{[j]}$ is not a function of $\mathbf{x}_{[j]}$.

A significance level, such as (2.3), cannot be calculated using (3.2) because $\lambda(\mathbf{x})$ is unknown. To adjust for overt bias in a study that is free of hidden bias is to address the fact that $\lambda(\mathbf{x})$ is unknown. The simplest approach is to stratify on $\mathbf{x}$.

3.2.2 Stratifying on x

Often, units are grouped into strata on the basis of the covariate **x**. From the M units, select $N \leq M$ units and group them into S nonoverlapping strata with n_s units in stratum s. In selecting the N units and assigning them to strata, use only the **x**'s and possibly a table of random numbers. A stratification formed in this way is called a *stratification on* **x**. Renumber the units so the ith unit in stratum s has treatment assignment Z_{si} and covariate $\mathbf{x}_{si}$. Using the same notation as Chapter 2, write **Z** for the N-tuple $(Z_{11}, \ldots, Z_{S,n_S})^T$. Write m_s for the number of treated units in stratum s; that is, $m_s = \sum_i Z_{si}$, and $\mathbf{m} = (m_1 \ldots, m_S)^T$.

An *exact stratification* on **x** has strata that are homogeneous in **x**, so two units are in the same stratum only if they have the same value of **x**, that is $\mathbf{x}_{si} = \mathbf{x}_{sj}$ for all s, i, and j. Exact stratification on **x** is practical only when **x** is of low dimension and its coordinates are discrete; otherwise, it will be difficult to locate many units with the same **x**.

In an exact stratification on **x**, if the study is free of hidden bias, that is, if (3.2) holds, then all units in the same stratum have the same chance of receiving the treatment. In this case, write λ_s in place of $\lambda(\mathbf{x}_{si})$, so (3.2) implies:

$$\text{prob}(\mathbf{Z} = \mathbf{z}) = \prod_{s=1}^{S} \prod_{i=1}^{n_s} \lambda_s^{z_{si}} (1 - \lambda_s)^{1 - z_{si}}. \tag{3.3}$$

In (3.3), the distribution of treatment assignments, $\text{prob}(\mathbf{Z} = \mathbf{z})$, is unknown because λ_s is unknown, and $m_s = \sum_i Z_{si}$ is a random variable. Consider the conditional distribution of **Z** given **m**. It is a distribution on a set Ω whose elements are N-tuples of 0s and 1s such that $\mathbf{z} \in \Omega$ if and only if $m_s = \sum_i z_{si}$ for $s = 1, \ldots, S$, so Ω has $K = \prod_{s=1}^{S} \binom{n_s}{m_s}$ elements. Every treatment assignment $\mathbf{z} \in \Omega$ has the same unconditional probability in (3.3), namely,

$$\text{prob}(\mathbf{Z} = \mathbf{z}) = \prod_{s=1}^{S} \lambda_s^{m_s} (1 - \lambda_s)^{n_s - m_s}. \tag{3.4}$$

It follows that the conditional probability given **m** is constant, $\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{m}) = 1/K$. Of course, this is the distribution of **Z** in a uniform randomized experiment; see §2.3.2.

In short, if an observational study is free of hidden bias, and if one stratifies exactly on **x**, then the conditional distribution of the treatment assignment **Z** given the numbers **m** of treated units in each stratum, namely $\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{m})$, is the same as the distribution of treatment assignments in a uniform randomized experiment. This is true even though the treatment assignment probabilities $\lambda(\mathbf{x})$ are unknown. In this case, given **m**, the statistical procedures discussed in Chapter 2 have the properties described

there. In other words, if the study is free of hidden bias and one stratifies exactly on $\mathbf{x}$, then the study may be analyzed using methods for a uniform randomized experiment.

Be clear on a key point. This result does not say that there is no difference between an experiment and an observational study. The difference is that in a uniform randomized experiment, the assignment probabilities $\text{prob}(\mathbf{Z} = \mathbf{z})$ are known to equal $1/K$ because we forced this to be true by randomizing. In an observational study, the conclusion $\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{m}) = 1/K$ is deduced from the premise that the study is free of hidden bias, a premise we have little reason to believe. In an observational study, this premise is subjected to strict scrutiny, asking whether evidence can support or refute it, asking how findings would change if the premise were in error. This scrutiny is the focus of most later chapters.

3.2.3 Matching on $\mathbf{x}$

In §3.2.2, strata were formed using $\mathbf{x}$ alone. One way that matching differs from stratification is that there are constraints on the number m_s of treated units and the number $n_s - m_s$ of control units in a stratum. For instance, pair matching requires $n_s = 2$ and $m_s = 1$ for each s, while matching with multiple controls requires $n_s \geq 2$ and $m_s = 1$. A *matching on* $\mathbf{x}$ is a matched sample formed by:

(i) placing some restriction on S, $\mathbf{m}$ and $\mathbf{n} = (n_1, \ldots, n_S)^{\mathrm{T}}$, and

(ii) picking a stratification that meets these restrictions based exclusively on the pattern of $\mathbf{x}$s in the strata and possibly a table of random numbers.

For instance, a pair matched sample with $S = 100$ pairs would be formed by considering all possible stratifications with $n_s = 2$ and $m_s = 1$ for $m_s = 1$, $s = 1, \ldots, 100$, and selecting one of these possible stratifications based on the $\mathbf{x}$'s in the strata and possibly random numbers. An *exact matching on* $\mathbf{x}$ is a matching on $\mathbf{x}$ in which $\mathbf{x}$ is the same for all n_s units in each matched set, that is, $\mathbf{x}_{si} = \mathbf{x}_{sj}$ for $i, j = 1, \ldots, n_s$ for each s. As with exact stratification, exact matching is possible only when $\mathbf{x}$ is of low dimension and discrete.

The same argument as in §3.2.2 shows that, in an observational study that is free of hidden bias, if one matches exactly on $\mathbf{x}$, then the conditional distribution of the treatment assignment $\mathbf{Z}$ given $\mathbf{m}$ is the same as in a uniform randomized experiment, $\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{m}) = 1/K$. If the study were free of hidden bias, it could be analyzed as if it were a matched randomized experiment, but of course, the comment at the end of §3.2.2 applies here as well.

3.2.4 An Example: Lead in the Blood of Children

Morton et al. (1982) studied lead in the blood of children whose parents worked in a factory where lead was used in making batteries. They were concerned that children were exposed to lead inadvertently brought home by their parents. Their study included 33 such children from different families—they are the exposed or treated children. The outcome R was the level of lead found in a child's blood in $\mu g/dl$ of whole blood. The covariate $\mathbf{x}$ was two-dimensional, recording age and neighborhood of residence. They matched each exposed child to one control child of the same age and neighborhood whose parents were employed in other industries not using lead. Table 3.1 shows the levels of lead found in the children's blood in $\mu g/dl$ of whole blood.

If this study were free of hidden bias, which may or may not be the case, we would be justified in analyzing Table 3.1 using methods for a uniform randomized experiment with 33 matched pairs. If the null hypothesis of no treatment effect is tested using Wilcoxon's signed rank test, the one-sided significance level is less than 0.0001. The Hodges–Lehmann estimate of the size of an additive effect is 15 $\mu g/dl$ with 95% confidence interval (9.5, 20.5). If the study were free of hidden bias, this would strongly suggest that the parents who worked with lead did raise the level of lead in their children's blood by about 15 $\mu g/dl$, a large increase compared to the level of lead found in controls. In later chapters, these data are examined again without the premise that the study is free of hidden bias.

3.2.5 Stratifying and Matching on the Propensity Score

Often, exact stratification or matching on $\mathbf{x}$ is difficult or impossible. If $\mathbf{x}$ is of high dimension or contains continuous measurements, each of the N units may have a different value of $\mathbf{x}$, so no stratum can contain a treated and control unit with the same $\mathbf{x}$. There are several questions. Do a large number of covariates—that is, a high-dimensional $\mathbf{x}$—make stratification and matching infeasible? Does close but inexact matching on $\mathbf{x}$ remove most of the bias due to $\mathbf{x}$? What algorithms produce good stratifications or matchings? The second and third questions are discussed in Chapter 10. The current section begins to answer the first question. As it turns out, there is a sense in which all matching problems are one-dimensional, so the dimensionality of $\mathbf{x}$ is not critical by itself.

Suppose an observational study is free of hidden bias, so (3.2) holds. Instead of stratifying or matching exactly on $\mathbf{x}$, imagine forming strata or matched sets in which units in the same stratum have the same chance of receiving the treatment $\lambda(\mathbf{x})$. Then within a stratum or matched set, units may have different values of $\mathbf{x}$, but they have the same propensity score $\lambda(\mathbf{x})$. Formally, it may happen that $\mathbf{x}_{si} \neq \mathbf{x}_{sj}$ but always $\lambda(\mathbf{x}_{si}) = \lambda(\mathbf{x}_{sj})$. Call this *exact matching or stratification on the propensity score*. In this

TABLE 3.1. Lead in Children's Blood ($\mu g/dl$).

Pair	Exposed	Control	Difference	Rank
1	38	16	22	22
2	23	18	5	8
3	41	18	23	23.5
4	18	24	−6	9.5
5	37	19	18	21
6	36	11	25	26
7	23	10	13	14
8	62	15	47	32
9	31	16	15	17
10	34	18	16	18.5
11	24	18	6	9.5
12	14	13	1	2.5
13	21	19	2	4
14	17	10	7	11
15	16	16	0	1
16	20	16	4	7
17	15	24	−9	12.5
18	10	13	−3	5.5
19	45	9	36	30
20	39	14	25	26
21	22	21	1	2.5
22	35	19	16	18.5
23	49	7	42	31
24	48	18	30	28
25	44	19	25	26
26	35	12	23	23.5
27	43	11	32	29
28	39	22	17	20
29	34	25	9	12.5
30	13	16	−3	5.5
31	73	13	60	33
32	25	11	14	15.5
33	27	13	14	15.5

case, the arguments in §3.2.2 and §3.2.3 go through without changes. In those arguments, equal $\mathbf{x}$'s within strata were used only to ensure equal $\lambda(\mathbf{x})$'s. In short, in an observational study free of hidden bias, exact matching or stratification on the propensity score yields a conditional distribution of treatment assignments $\mathbf{Z}$ given $\mathbf{m}$ that is the same as a uniform randomized experiment, namely $\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{m}) = 1/K$. In this case, the methods in Chapter 2 for uniform randomized experiments may be applied. The same conclusion is reached if strata or matched sets are formed based on $\lambda(\mathbf{x})$ and parts of $\mathbf{x}$, providing the strata or matched sets are homogeneous in $\lambda(\mathbf{x})$.

In practice, $\lambda(\mathbf{x})$ is unknown, so matching or stratification on $\lambda(\mathbf{x})$ is not possible. The use of estimated propensity scores in matching and stratification is discussed in Chapter 10. Also, §10.2 discusses balancing properties of the propensity score that are true whether or not the study is free of hidden bias.

3.3 Case-Referent Studies

3.3.1 Selecting Subjects Based on Their Outcomes

In Chapter 1, the study of DES and vaginal cancer is a case-referent study, and such a study has two features that distinguish it from the other observational studies in Chapter 1. First, the binary outcome, namely, vaginal cancer, is extremely rare, so a study that simply followed women until they developed the disease would need an enormous number of women to produce even a handful of cases of this rare cancer. This first feature is the motivation for conducting a case-referent study, but it is the second feature that characterizes such a study. In a case-referent study, cases are deliberately over-represented and referents are under-represented. The DES study identified cases of vaginal cancer and compared them to a small number of matched referents in terms of the frequency of maternal exposure to DES. In other words, subjects are included or excluded from the study, in part, on the basis of their outcomes. Instead of comparing the outcomes of treated and untreated groups, the case-referent study compares the frequency of exposure to the treatment among cases and referents.

At first, it is not clear that this makes sense. If the outcome $\mathbf{R} = \mathbf{r}_\mathbf{Z}$ is affected by the treatment $\mathbf{Z}$, then selecting subjects using their outcomes may distort the frequency of exposure to the treatment. Indeed, this seems to have happened in the DES study. Of the 40 women in the study, 7 had mothers who had used DES, which exceeds the frequency of exposure to DES in the general population. When the treatment has an effect, it is related to the outcome, so selecting subjects using their outcomes changes the frequency of exposure to the treatment. How can a case-referent study be interpreted?

TABLE 3.2. Data Before Selecting Cases and Referents.

		Z	
		1	0
R	1	$\sum_{\mathbf{x}} Z_{[j]} r_{T[j]}$	$\sum_{\mathbf{x}} (1 - Z_{[j]}) r_{C[j]}$
	0	$\sum_{\mathbf{x}} Z_{[j]} (1 - r_{T[j]})$	$\sum_{\mathbf{x}} (1 - Z_{[j]})(1 - r_{C[j]})$

TABLE 3.3. Expectations in the Absence of Hidden Bias.

		Z	
		1	0
R	1	$\lambda(\mathbf{x}) \sum_{\mathbf{x}} r_{T[j]}$	$\{1 - \lambda(\mathbf{x})\} \sum_{\mathbf{x}} r_{C[j]}$
	0	$\lambda(\mathbf{x}) \sum_{\mathbf{x}} (1 - r_{T[j]})$	$\{1 - \lambda(\mathbf{x})\} \sum_{\mathbf{x}} (1 - r_{C[j]})$

3.3.2 Synthetic Case-Referent Studies

A synthetic case-referent study starts with the population of M subjects in §3.2.1, and draws a random sample of cases and a separate random sample of referents, possibly after stratification using the observed covariates $\mathbf{x}$. Synthetic case-referent studies are typically conducted when there is a computerized database describing the entire population of M subjects, but the study requires the costly collection of additional data not in the database. See Silber et al. (2001) for an example in which the population is comprised of all Medicare patients in Pennsylvania. This sort of study does occur, but far more common are case-referent studies that do not use random sampling. Synthetic case-referent studies are easier to consider theoretically because the mechanism that selects subjects has known properties. The term "synthetic" was introduced by Mantel (1973), while the odds ratio property discussed in this section is due to Cornfield (1951).

Consider the data before cases and referents are sampled, as in §3.2.1, so the jth of the M subjects has observed covariate $\mathbf{x}_{[j]}$, treatment assignment $Z_{[j]}$, and observed binary response $R_{[j]}$, which equals $r_{T[j]}$ if j is given the treatment and $r_{C[j]}$ if j is given the control. Divide the M subjects in the population into strata based on $\mathbf{x}$, and abbreviate by $\sum_{\mathbf{x}}$ a sum over all subjects j with $\mathbf{x}_{[j]} = \mathbf{x}$; that is, write $\sum_{\mathbf{x}}$ for $\sum_{j:\mathbf{x}_{[j]}=\mathbf{x}}$. The subjects in the stratum with covariate value $\mathbf{x}$ are recorded in the contingency Table 3.2.

If there is no hidden bias, then $E(Z_{[j]}) = \lambda(\mathbf{x}_{[j]})$, so the entries in Table 3.2 have as expectations the values in Table 3.3.

The odds ratio or cross-product ratio in Table 3.3 is

$$\frac{\left(\lambda(\mathbf{x}) \sum_{\mathbf{x}} r_{T[j]}\right) \left(\{1 - \lambda(\mathbf{x})\} \sum_{\mathbf{x}} (1 - r_{C[j]})\right)}{\left(\lambda(\mathbf{x}) \sum_{\mathbf{x}} (1 - r_{T[j]})\right) \left(\{1 - \lambda(\mathbf{x})\} \sum_{\mathbf{x}} r_{C[j]}\right)} = \frac{\left(\sum_{\mathbf{x}} r_{T[j]}\right) \left(\sum_{\mathbf{x}} (1 - r_{C[j]})\right)}{\left(\sum_{\mathbf{x}} (1 - r_{T[j]})\right) \left(\sum_{\mathbf{x}} r_{C[j]}\right)}, \tag{3.5}$$

TABLE 3.4. Expected Counts in a Synthetic Case-Referent Study Absent Hidden Bias.

		Z	
		1	0
R	1	$k_{1\mathbf{x}}\lambda(\mathbf{x})\sum_{\mathbf{x}} r_{T[j]}$	$k_{1\mathbf{x}}\{1-\lambda(\mathbf{x})\}\sum_{\mathbf{x}} r_{C[j]}$
	0	$k_{0\mathbf{x}}\lambda(\mathbf{x})\sum_{\mathbf{x}}(1-r_{T[j]})$	$k_{0\mathbf{x}}\{1-\lambda(\mathbf{x})\}\sum_{\mathbf{x}}(1-r_{C[j]})$

and this odds ratio is a measure of the magnitude of a treatment effect. Notice that the odds ratio is one if there is no treatment effect in the stratum defined by $\mathbf{x}$, that is if $r_{T[j]} = r_{C[j]}$ for all $[j]$ with $\mathbf{x}_{[j]} = \mathbf{x}$, and the odds ratio is greater than one if the treatment has a positive effect in the stratum defined by $\mathbf{x}$, that is if $r_{T[j]} \geq r_{C[j]}$ for all j with $\mathbf{x}_{[j]} = \mathbf{x}$ and $r_{T[j]} \neq r_{C[j]}$ for some j with $\mathbf{x}_{[j]} = \mathbf{x}$. Hamilton (1979) discusses a wide variety of related measures under the model of a positive effect.

Tables 3.2 and 3.3 describe the initial population of M subjects. Consider a synthetic case-referent study formed from Table 3.2. Draw a random sample without replacement consisting of a fraction $k_{1\mathbf{x}}$ of the cases, the first row of the table, and a random sample consisting of a fraction $k_{0\mathbf{x}}$ from the referents, the second row, with $0 < k_{1\mathbf{x}} \leq 1$ and $0 < k_{0\mathbf{x}} \leq 1$. The resulting table of counts for the synthetic case-referent study has expectations shown in Table 3.4.

The key observation is that the odds ratio computed from Table 3.4 equals the odds ratio (3.5) before case-referent sampling. In other words, in the absence of hidden bias, the data from a synthetic case-referent study provide a direct estimate of the population odds ratio (3.5) at each $\mathbf{x}$.

As in §3.2, when attention shifts from the population of M subjects to the $N < M$ subjects included in the case-referent study, the notation in §3.2.3 is used; that is, the ith subject in the sth stratum defined by $\mathbf{x}$ has unbracketed subscript s, i. As in §3.2, let $m_s = \sum_i Z_{si}$ be the number of treated or exposed subjects in stratum s of the case-referent study. If the study were free of hidden bias, so (3.2) holds, and if the treatment had no effect, so $r_{Tsi} = r_{Csi}$ for each s, i, then after synthetic case-referent sampling, the conditional distribution of the treatment assignments $\mathbf{Z}$ given $\mathbf{m}$ is uniform on Ω. This means, for instance, that the Mantel–Haenszel statistic may be used to test the null hypothesis of no effect in a synthetic case-referent study that is free of hidden bias.

3.3.3 Selection Bias in Case-Referent Studies

Unlike the synthetic study in §3.3.2, most case-referent studies do not use random sampling. It is common to use all cases that are made available by some process, for instance, all new cases admitted to one or more hospitals

in a given time interval, together with referents selected by an ostensibly similar process, for instance, patients with some other illness admitted to the same hospitals at the same time, or neighbors or coworkers of the cases.

Nonrandom selections of cases and referents may distort the odds ratio in (3.5). For instance, if cases of lung cancer at a hospital were compared to referents who were selected as patients with cardiac disease in the same hospital, the odds ratio linking lung cancer with cigarette smoking would be too small, because smoking causes both lung cancer and cardiac disease. In this case, there is a *selection bias*, that is, a bias that was not the result of the manner in which subjects were assigned to treatment in the population, but rather a bias introduced by the nonrandom process of selecting cases and referents. Selection bias is discussed further in Chapter 8.

3.4 *Small Sample Inference with an Unknown Propensity Score

*3.4.1 *Conditional Inference Under a Logit Model for the Propensity Score*

In §3.2, in the absence of hidden bias, the distribution of treatment assignments $\text{prob}(\mathbf{Z} = \mathbf{z})$ in (3.2) was unknown because the propensity score $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)^{\mathrm{T}}$ was unknown. However, by conditioning on the number of treated subjects in each stratum $\mathbf{m}$, the conditional distribution of treatment assignments, $\text{prob}(\mathbf{Z} = \mathbf{z}|\mathbf{m})$, was known, and, in fact, was the distribution of treatment assignments in a uniform randomized experiment. Notice that the unknown parameter $\boldsymbol{\lambda}$ was eliminated by conditioning on a sufficient statistic $\mathbf{m}$. This line of reasoning generalizes.

Suppose that $\boldsymbol{\lambda}$ satisfies a logit model,

$$\log\left(\frac{\lambda_s}{1-\lambda_s}\right) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_s, \tag{3.6}$$

where $\boldsymbol{\beta}$ is an unknown parameter. Write

$$\bar{\mathbf{m}} = \sum_{s=1}^{S} m_s \mathbf{x}_s, \tag{3.7}$$

so $\bar{\mathbf{m}}$ is the sum of the $\mathbf{x}_s$ weighted by the number of treated subjects m_s in stratum s. Under the model (3.6), $\bar{\mathbf{m}}$ is sufficient for $\boldsymbol{\beta}$, so $\text{prob}(\mathbf{Z} = \mathbf{z}|\bar{\mathbf{m}})$ is a known distribution, free of the unknown parameter $\boldsymbol{\beta}$. See Cox (1970) for a detailed discussion of logit models, including a discussion of the sufficiency of $\bar{\mathbf{m}}$.

Let $\bar{\Omega}$ be the set containing all treatment assignments $\mathbf{z}$ that give rise to the same value of $\bar{\mathbf{m}}$, that is,

$$\bar{\Omega} = \left\{ \mathbf{z} : z_{si} \in \{0,1\},\ i = 1, \ldots, n_s, s = 1, \ldots, S,\ \bar{\mathbf{m}} = \sum_{s=1}^{S} \mathbf{x}_s \sum_{i=1}^{n_s} z_{si} \right\}.$$

Notice that $\bar{\Omega}$ is a larger set than Ω, in the sense that $\Omega \subseteq \bar{\Omega}$, so $\bar{\Omega}$ contains at least as many treatment assignments $\mathbf{z}$ as Ω. This is true because every $\mathbf{z} \in \Omega$ gives rise to the same $\mathbf{m}$, and hence also to the same $\bar{\mathbf{m}}$.

Under the logit model (3.6), the conditional distribution of treatment assignments given $\bar{\mathbf{m}}$ is constant on $\bar{\Omega}$,

$$\text{prob}(\mathbf{Z} = \mathbf{z} | \bar{\mathbf{m}}) = \frac{1}{|\bar{\Omega}|} \qquad \text{for each} \quad \mathbf{z} \in \bar{\Omega}. \tag{3.8}$$

As a result, in the absence of hidden bias, under the model (3.6), the known distribution (3.8) forms the basis for a permutation test; in particular, a test statistic $T = t(\mathbf{Z}, \mathbf{r})$ has significance level

$$\text{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq T | \bar{\mathbf{m}}\} = \frac{|\{\mathbf{z} \in \bar{\Omega} : t(\mathbf{z}, \mathbf{r}) \geq T\}|}{|\bar{\Omega}|}, \tag{3.9}$$

for testing the null hypothesis of no treatment effect. The significance level (3.9) is the proportion of treatment assignments $\mathbf{z}$ in $\bar{\Omega}$ giving a value of the test statistic $t(\mathbf{z}, \mathbf{r})$ at least equal to the observed value of T.

The procedure just described is useful in small samples when the strata are so thin that the set Ω is too small to be useful. For instance, if Ω contains fewer than 20 treatment assignments, then no pattern of responses can be significant at the 0.05 level using (2.5). Such a small Ω arises when the sample size N is small and this small sample size is thinly spread over many strata; for instance, strata with $m_s = 0$ or $m_s = n_s$ contribute nothing to (2.5). If Ω is too small to permit reasonable permutation inferences, the larger set $\bar{\Omega}$ may be used instead. The size $|\bar{\Omega}|$ of $\bar{\Omega}$ depends on how the covariates $\mathbf{x}$ are coded, and finding an $\bar{\Omega}$ of appropriate size may sometimes be accomplished by adjusting the coding of the covariates. An example of the method is given in the next section.

Exact stratification and matching in §3.2 are useful only if there are many strata or matched sets that contain at least $n_s \geq 2$ subjects with at least one treated subject and one control. In contrast, the method in the current section may be used when $n_s = 1$ for every s, so each unit may have a distinct $\mathbf{x}$.

The method of this section is a generalization of the method in §3.2. Suppose that $\mathbf{x}$ simply contains indicators coding the S strata. For instance, in one such coding, $\mathbf{x}$ has $S-1$ binary coordinates, with coordinate $x_{sij} = 1$ if $j = s$ and $x_{sij} = 0$ if $j \neq s$, for $j = 1, \ldots, S-1$. In this case, $\bar{\Omega} = \Omega$, and (3.9) equals (2.5).

TABLE 3.5. Fourteen Lung Cancer Patients from a Phase II Trial of Pacco.

ID	Tumor Response r_{si}	Treatment Z_{si}	Cell Type	Previous Treatment	Performance Status	s
1	0	0	Squamous	None	0	1
2	0	0	Large cell	None	1	2
3	0	0	Squamous	Radiation	1	3
4	0	0	Squamous	Radiation	1	3
5	0	0	Squamous	Radiation	2	4
6	1	1	Squamous	Radiation	1	3
7	0	1	Squamous	Radiation	1	3
8	0	1	Adeno.	Radiation	1	5
9	1	1	Squamous	None	1	6
10	0	1	Large cell	None	2	7
11	0	1	Squamous	Radiation and chemotherapy	2	8
12	0	1	Squamous	Chemotherapy	1	9
13	0	1	Squamous	None	0	1
14	2	1	Squamous	None	1	6

*3.4.2 *An Example: A Small Observational Study Embedded in a Clinical Trial*

Table 3.5 describes 14 patients taken from a clinical trial of the drug treatment combination PACCO in the treatment of nonsmall cell bronchogenic carcinoma (Whitehead, Rosenbaum, and Carbone 1984, Rosenbaum 1984b, §3). This phase II trial contained two minor variations of what was intended to be the same treatment; however, when the responses were tabulated, all of the patients who responded to therapy had received the same variation of the treatment. The question is whether this is evidence that the treatments differ, given the characteristics of the patients involved. As presented here, the example is adapted to illustrate the method.

The outcome is tumor response, where 0 signifies no response, 1 signifies a partial response, and 2 signifies a complete response. The covariate information describes the cell type, previous treatment, and performance status. In $\mathbf{x}_s$ previous treatment was coded as three binary variables, cell type as two binary variables, and performance status was taken as a single variable with three scored categories, so $\mathbf{x}_s$ has dimension six.

Table 3.6 describes the strata based on the covariates. The 14 patients are divided into nine strata, with most strata containing a single patient. The set Ω has $2 \times 1 \times 6 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 = 12$ treatment assignments, so no matter what responses are observed, the smallest possible significance

TABLE 3.6. Strata for the PACCO Trial.

s	Cell Type	Previous Treatment	Performance Status	n_s	m_s	$\binom{n_s}{m_s}$
1	Squamous	None	0	2	1	2
2	Large cell	None	1	1	0	1
3	Squamous	Radiation	1	4	2	6
4	Squamous	Radiation	2	1	0	1
5	Adeno.	Radiation	1	1	1	1
6	Squamous	None	1	2	2	1
7	Large cell	None	2	1	1	1
8	Squamous	Radiation and chemotherapy	2	1	1	1
9	Squamous	Chemotherapy	1	1	1	1

level is $1/12 = 0.083$. The set Ω is too small to be useful for a permutation test.

The set $\bar{\Omega}$ is somewhat larger, containing 28 treatment assignments. These are all the treatment assignments that give rise to the observed value of $\bar{\mathbf{m}}$. This means that a treatment assignment $\mathbf{z}$ must have the correct number of treated patients with each cell type, each previous treatment, and the correct average performance status. It contains all treatment assignments $\mathbf{z}$ such that:

(i) nine patients receive treatment one, and of those nine:

(ii) one has adenocarcinoma;

(iii) one has large cell carcinoma;

(iv) seven have squamous cell carcinoma;

(v) three have had only previous radiation therapy;

(vi) one has had only previous chemotherapy;

(vii) one has had both previous chemotherapy and previous radiation therapy; and

(viii) the average performance status is $10/9 = 1.1$.

In other words, these 28 treatment assignments resemble the observed treatment assignment in the sense that similar patients received the treatment. The test statistic is the total of the response scores for treated patients, $T = t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{r} = 2+1+1 = 4$. Of the 28 treatment assignments $\mathbf{z}$ in $\bar{\Omega}$, 8 have $\mathbf{z}^{\mathrm{T}}\mathbf{r} = 4$, 13 have $\mathbf{z}^{\mathrm{T}}\mathbf{r} = 3$, 4 have $\mathbf{z}^{\mathrm{T}}\mathbf{r} = 2$, and 3 have $\mathbf{z}^{\mathrm{T}}\mathbf{r} = 1$. Therefore, under the null hypothesis, the distribution (3.9) of

$\mathbf{Z}^{\mathrm{T}}\mathbf{r}$ assigns probability $8/28 = 0.29$ to $4, 13/28 = 0.46$ to $3, 4/28 = 0.14$ to 2, and $3/28 = 0.11$ to 1. In other words, it is not surprising to find that all the responses occurred in treatment group one—this would happen by chance 29% of the time if the treatments did not differ in their effects.

Algorithms for computations involving $\bar{\Omega}$ are discussed in Rosenbaum (1984b).

3.5 *Large Sample Inference with an Unknown Propensity Score

*3.5.1 *Covariance Adjustment of Randomization Tests*

This section describes a method of covariate adjustment for randomization tests. The method is simple to describe and to apply, and motivation follows the description. It turns out that the method is a large sample approximation to the exact test in §3.4.

As in §3.4, suppose a logit model accurately describes the propensity score,

$$\log\left(\frac{\lambda_s}{1-\lambda_s}\right) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_s, \tag{3.10}$$

and suppose the study is free of hidden bias, so that $\text{prob}(Z_{si} = 1) = \pi_{si} = \lambda(\mathbf{x}_s) = \lambda_s$. As in §3.4.1, each stratum may consist of a single unit, $n_s = 1$, so that each unit may have a distinct value of $\mathbf{x}$.

Under the null hypothesis of no treatment effect, $H_0 : r_{Tsi} = r_{Csi} = r_{si}$ for all s, i, the observed responses $\mathbf{R}$ equal a fixed vector $\mathbf{r}$ not varying with $\mathbf{Z}$. As in Chapter 2 and §3.2, when the null hypothesis of no effect is true, the fixed $\mathbf{R} = \mathbf{r}$ is observed no matter what $\mathbf{Z}$ is, so r_{si} does not predict Z_{si} at each $\mathbf{x}_s$. On the other hand, if the null hypothesis were false and instead $r_{Tsi} > r_{Csi}$ for each s, i, then the observed response, $R_{si} = Z_{si}r_{Tsi} + (1 - Z_{si})r_{Csi}$ would be positively related to Z_{si} at each $\mathbf{x}_s$. So the task is to check for a relationship between treatment Z_{si} and observed response R_{si} given $\mathbf{x}_s$, exploiting the assumed form (3.10).

Let $\mathbf{q}(\mathbf{R})$ be some way of scoring the observed responses, such as their ranks, so that under the null hypothesis of no treatment effect, $\mathbf{q}(\mathbf{R}) = \mathbf{q}(\mathbf{r}) = \mathbf{q}$, say, is fixed. Consider the model:

$$\log\left\{\frac{\text{prob}(Z_{si} = 1)}{\text{prob}(Z_{si} = 0)}\right\} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_s + \theta q_{si}. \tag{3.11}$$

By what has been said, under the null hypothesis of no treatment effect, in the absence of hidden bias, assuming model (3.10), it follows that $\theta = 0$ in (3.11).

There are several ways to test $H_0 : \theta = 0$ in (3.11). There is an exact, uniformly most powerful unbiased test of $H_0 : \theta = 0$ versus $H_A : \theta > 0$; see Cox (1970, §4.2). As it turns out, this test is precisely the test in §3.4 with significance level (3.9) providing $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^T\mathbf{q}$; see Rosenbaum (1984b, §4.2). In other words, so far, this section has not yielded a new procedure, but rather a new motivation for the procedure in §3.4. The most powerful unbiased test is not always practical, however. For instance, $\bar{\Omega}$ may be too large or too small to permit practical use of (3.9).

When N is large compared to the dimension of $\mathbf{x}$, it is more common to test a hypothesis about a coefficient in a logit model, such as $H_0 : \theta = 0$ in (3.11), using one of several large sample tests associated with maximum likelihood estimation of the model. These tests may also be applied here.

Suppose, again, that the study is free of hidden bias and (3.10) describes the propensity score, but instead of the null hypothesis of no effect, consider the model of an additive treatment effect, $H_0 : r_{Tsi} = r_{Csi} + \tau$. In this case, the observed response $R_{si} = r_{Csi} + \tau Z_{si}$ will be related to Z_{si} at each $\mathbf{x}_s$, positively related for $\tau > 0$, and negatively related for $\tau < 0$. Consider testing the hypothesis $H_0 : \tau = \tau_0$. The procedure is analogous to that in §2.6.1. Specifically, calculate the adjusted responses, $\mathbf{r} = \mathbf{R} - \tau_0\mathbf{Z}$, which are fixed when the null hypothesis is true, and apply the method above with $\mathbf{q} = \mathbf{q}(\mathbf{r}) = \mathbf{q}(\mathbf{R} - \tau_0\mathbf{Z})$. A confidence interval for τ is found by inverting the test.

*3.5.2 *Example: Benzene Exposure Among Shoe Workers*

Tunca and Egeli (1996) studied the effects of benzene exposure among shoe workers. The $58 = \sum m_s$ shoe workers, $Z = 1$, in Bursa, Turkey were exposed to benzene from glues. They were compared to $20 = \sum n_s - m_s$ controls, $Z = 0$, from the same region who were not exposed to benzene. The outcome, R, is one measure of chromosome damage, namely, the percentage of cells with breaks. Adjustment is made for a three-dimensional covariate, $\mathbf{x}$, giving age, alcohol intake (+ or −), and smoking in packs per day. Table 3.7 gives the data for 6 of the $78 = 58 + 20$ subjects, with data for all subjects given in Tunca and Egeli (1996). Casual examination of means shows the shoe workers have many more breaks than the controls, but they are also older (37.2 years versus 28.5 years), smoke more (0.8 packs per day versus .2 packs per day), and are more likely to drink alcohol (40% versus 5%).

To test the null hypothesis of no treatment effect, the model (3.11) was fit, predicting treatment Z from the three covariates and the ranks of the percentage of breaks R. When a logit model is fitted to independent binary trials by maximum likelihood, a common test of the hypothesis that a coefficient is zero uses the ratio of the coefficient to its approximate standard error, and compares that ratio to the standard Normal distribution. For age, alcohol, smoking, and the ranks of the percentages of breaks, these

TABLE 3.7. Chromosome Breaks Among Shoe Workers and Controls.

Treatment Z	Breaks R	Age	Alcohol	Smoking
0	0.00	24	−	0
0	0.00	31	−	0
⋮	⋮	⋮	⋮	⋮
0	5.00	23	−	0.1
⋮	⋮	⋮	⋮	⋮
1	11.11	26	−	1.0
⋮	⋮	⋮	⋮	⋮
1	0.00	50	+	1.5
1	9.09	40	−	0

ratios are, respectively, 2.96, 1.72, 1.59, and 2.18. Because $2.18 \geq 1.65$, and $1 - \Phi(1.65) = 0.05$, where $\Phi(\cdot)$ is the standard Normal cumulative distribution, the hypothesis $H_0 : \theta = 0$ may be rejected at the 0.05 level in a one-sided test. This is an appropriate large sample test of the null hypothesis of no treatment effect assuming the study is free of hidden bias and (3.10) accurately describes the propensity score.

The model of an additive treatment effect, $H_0 : r_{Tsi} = r_{Csi} + \tau$, is not useful—indeed, it is misleading—in this example. Many of the shoe workers and nearly all of the controls had $R = 0.00$ breaks, even though many other shoe workers had substantial numbers of breaks. As just noted, the hypothesis $H_0 : \tau = 0$ was rejected at the 0.05 level. However, the hypothesis $H_0 : \tau = \tau_0$ is not rejected in a one-sided 0.05 level test for *every* $\tau_0 > 0$, no matter how small τ_0 is. The reason is that if a positive quantity is subtracted from the many $R = 0.00$ values among shoe workers, these values become strictly smaller than all control responses. A better model for the treatment effect in data of this sort is discussed in §5.6.

3.6 *Inexact Matching Followed by Stratification

*3.6.1 *An Example: Vasectomy and Myocardial Infarction*

Walker et al. (1981) studied the possible effect of vasectomy on increased risk of myocardial infarction (MI), a possibility suggested by animal studies where increased risks were observed. The study contained 4830 pairs of men, one vasectomized and one control, matched for year of birth and calendar time of follow-up. The data were not matched for two other variables, obesity and smoking history recorded as binary traits, both of which are believed to be related to the risk of MI. This section describes a method for controlling covariates that were not controlled by the matching.

The outcome is a binary variable indicating whether an MI occurred during the follow-up period; however, the method discussed here may be used with outcomes of any kind. McNemar's test statistic is used; see §2.4.3 for discussion of this test. In most of the 4830 pairs, no MI occurred. Pairs containing no MI or two MIs are said to be concordant, and it is not difficult to verify that these pairs do not contribute to McNemar's test. There were 36 discordant pairs, that is, pairs in which one person had an MI and the other did not—only these affect the test. Walker's (1982) data for these 36 pairs of men are given in Table 3.8. The score q_{si} is 1 if the ith man in pair s had an MI and is 0 otherwise.

In each pair, there are two possibilities. A pair may be exactly matched for obesity and smoking, so the two matched men are the same on these variables, or else the men may differ. If they differ, then there are six ways they may differ; that is, there are six patterns of imbalance in the covariates. For instance, one possible imbalance occurs if one man in a pair is a nonsmoker who is not obese and the other is a smoker who is not obese; this is (os, oS) in Table 3.8. Notice that, in counting the patterns of imbalance in covariates, we consider only the covariates and not vasectomy or MI. The 36 pairs are grouped into seven classes, one class that is perfectly matched and six classes for the six patterns of imbalance.

*3.6.2 *Adjusting Inexactly Matched Pairs by Stratifying Pairs*

Suppose that an observational study is free of hidden bias, so (3.2) holds, and pairs of treated and control units are matched inexactly for $\mathbf{x}$, so $\mathbf{x}_{s1}$ may not equal $\mathbf{x}_{s2}$. How does one control imbalances in $\mathbf{x}$ that remain after matching? The method described in this section is useful when matching has failed to control a few coordinates of $\mathbf{x}$ containing discrete variables, as in §3.6.1. It involves grouping the matched pairs into classes so that the pairs in the same class have the same pattern of covariate imbalance.

Using (3.2), for any two distinct units j and k, $\text{prob}(Z_{[j]} = 1, Z_{[k]} = 0) = \lambda\left(\mathbf{x}_{[j]}\right)\left\{1 - \lambda\left(\mathbf{x}_{[k]}\right)\right\}$ and $\text{prob}(Z_{[j]} = 0, Z_{[k]} = 1) = \lambda\left(\mathbf{x}_{[k]}\right)\left\{1 - \lambda\left(\mathbf{x}_{[j]}\right)\right\}$. Pair matching selects units so that $m_s = Z_{s1} + Z_{s2} = 1$ for each s. Therefore, conditionally given that a pair contains exactly one treated unit, the chance that the first unit is the treated unit is

$$\text{prob}(Z_{s1} = 1 | Z_{s1} + Z_{s2} = 1) = \frac{\lambda(\mathbf{x}_{s1})\{1 - \lambda(\mathbf{x}_{s2})\}}{\lambda(\mathbf{x}_{s1})\{1 - \lambda(\mathbf{x}_{s2})\} + \lambda(\mathbf{x}_{s2})\{1 - \lambda(\mathbf{x}_{s1})\}}. \tag{3.12}$$

In many applications, some coordinates of $\mathbf{x}$ are matched and others are not. In the vasectomy and MI example, men were matched for year of birth and calendar time of follow-up, but not for smoking and obesity. In this situation, it seems natural to let the matching control the matched coordinates of $\mathbf{x}$ and to make additional adjustments only for the unmatched

TABLE 3.8. Vasectomy and Myocardial Infarction

Class c	Covariate imbalance $(\mathbf{x}_{s1}, \mathbf{x}_{s2})$	q_{s1}	q_{s2}	Z_{s1}	Walker's id#, s
0	No imbalance	1	0	0	4
		1	0	1	8
		1	0	1	10
		0	1	1	11
		1	0	0	12
		1	0	1	16
		0	1	0	17
		0	1	0	23
		1	0	0	26
		0	1	0	27
		1	0	1	30
		0	1	0	35
		1	0	1	36
1	(os, oS)	0	1	0	3
		1	0	1	6
		0	1	1	9
		0	1	1	14
		0	1	0	15
		0	1	1	20
		0	1	1	21
		1	0	0	22
		0	1	0	24
		0	1	1	29
		0	1	0	32
2	(os, Os)	0	1	1	1
		0	1	0	28
		1	0	0	33
3	(oS, OS)	0	1	0	19
		0	1	0	25
4	(oS, Os)	0	1	0	2
		0	1	0	7
		1	0	0	34
5	(oS, OS)	0	1	0	13
		1	0	0	18
		0	1	1	31
6	(Os, OS)	0	1	1	5

Key: O = obese, o = not obese, S = smoker, s = nonsmoker.

coordinates of $\mathbf{x}$. Write $\mathbf{x} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ where subjects are matched for $\overline{\mathbf{x}}$ but not for $\widetilde{\mathbf{x}}$. Then $\overline{\mathbf{x}}_{s1} = \overline{\mathbf{x}}_{s2}$ for every s, but $\widetilde{\mathbf{x}}_{s1} \neq \widetilde{\mathbf{x}}_{s2}$ for at least some s. Ideally, matched pairs could be grouped into classes based on the imbalances in the unmatched coordinates, $\widetilde{\mathbf{x}}_{s1}$ and $\widetilde{\mathbf{x}}_{s2}$, ignoring the matched coordinates, $\overline{\mathbf{x}}_{s1} = \overline{\mathbf{x}}_{s2}$, as was done in Table 3.12 where pairs are grouped based on obesity and smoking, ignoring year of birth and calendar time of follow-up. As intuition might suggest, we can control imbalances in $\overline{\mathbf{x}}$ by matching on $\overline{\mathbf{x}}$, then separately control the remaining imbalances in $\widetilde{\mathbf{x}}$ by classifying the pairs, only if $\overline{\mathbf{x}}$ and $\widetilde{\mathbf{x}}$ do not interact with each other in determining $\lambda(\mathbf{x})$. Specifically, consider the following additive logit model for the treatment assignment probabilities, $\lambda(\mathbf{x})$:

$$\lambda(\mathbf{x}) = \frac{\exp\{\xi(\overline{\mathbf{x}}) + \zeta(\widetilde{\mathbf{x}})\}}{1 + \exp\{\xi(\overline{\mathbf{x}}) + \zeta(\widetilde{\mathbf{x}})\}} \tag{3.13}$$

for some unknown functions $\xi(\cdot)$ and $\zeta(\cdot)$. Because $\overline{\mathbf{x}}_{s1} = \overline{\mathbf{x}}_{s2}$, it follows that $\xi(\overline{\mathbf{x}}_{s1}) = \xi(\overline{\mathbf{x}}_{s2})$, so that, substituting (3.13) into (3.12) and simplifying yields:

$$\text{prob}(Z_{s1} = 1 | Z_{s1} + Z_{s2} = 1) = \frac{\exp\{\zeta(\widetilde{\mathbf{x}}_{s1})\}}{\exp\{\zeta(\widetilde{\mathbf{x}}_{s1})\} + \exp\{\zeta(\widetilde{\mathbf{x}}_{s2})\}}, \tag{3.14}$$

which depends only on the unmatched coordinates, $\widetilde{\mathbf{x}}$, not on the matched coordinates, $\overline{\mathbf{x}}$. This sort of simplification is often possible with logit models; see Cox (1970). In other words, under model (3.13), matching on $\overline{\mathbf{x}}$ has removed all of the bias due to $\overline{\mathbf{x}}$, and only the bias due to $\widetilde{\mathbf{x}}$ remains. When model (3.13) holds, covariate imbalance refers to $\widetilde{\mathbf{x}}$ only, so the 11 pairs in Table 3.8 with a nonobese nonsmoker matched to a nonobese smoker, (os, oS), all have the same pattern of covariate imbalance, and all 11 pairs may be placed in the same class for adjustments. When model (3.13) does not hold, the method of this section may still be applied, but many more classes are needed. This is because pairs with different values of the matched covariates, $\overline{\mathbf{x}}$, must be placed in different classes, so the 11 pairs in Table 3.8 with a nonobese nonsmoker matched to a nonobese smoker, (os, oS), would have to be divided up into different classes based on year of birth and calendar time of follow-up. The discussion that follows applies to both cases, that is, whether or not model (3.13) holds; however, the definition of covariate imbalance and the classes that control it do depend on whether model (3.13) holds. Specifically, when model (3.13) holds, a pattern of covariate imbalance refers to unmatched coordinates $(\widetilde{\mathbf{x}}_{s1}, \widetilde{\mathbf{x}}_{s2})$ only, but when this model does not hold, a pattern of covariate imbalance refers to all coordinates $(\mathbf{x}_{s1}, \mathbf{x}_{s2})$ whether matched or not.

Divide the S pairs into $C + 1$ classes, where class 0 contains the exactly matched pairs with $\mathbf{x}_{s1} = \mathbf{x}_{s2}$, and the other C classes contain the C patterns of imbalance in $\mathbf{x}$. Let l_c be the set of pairs exhibiting imbalance c, so that $l_0 \cup l_1 \cup \cdots \cup l_C = \{1, \ldots, S\}$. In class $c = 0$, there is an

exact match; that is, $\mathbf{x}_{s1} = \mathbf{x}_{s2}$ for $s \in l_0$. In the other classes, there is an imbalance, $\mathbf{x}_{s1} \neq \mathbf{x}_{s2}$ for $s \in l_c$ for $1 \leq c \leq C$. Write $\tilde{n}_c$ for the number of pairs in l_c.

Renumber the two units in each pair so that every pair s in class c has the same value of $\tilde{\mathbf{x}}_{s1}$, and every pair in class c has the same value of $\tilde{\mathbf{x}}_{s2}$, as in Table 3.8. This notational change simplifies the appearance of various quantities but it does not change their values.

With this notation, (3.12) takes the same value for all pairs in the same class. Write ρ_c for the common value of (3.12) or (3.14) in class c; that is, $\text{prob}(Z_{s1} = 1 | Z_{s1} + Z_{s2} = 1) = \rho_c$ for all $s \in l_c$. Notice that $\rho_0 = \frac{1}{2}$, but the other ρ_c are unknown since $\lambda(\mathbf{x})$ is unknown.

For $c = 1, \ldots, C$, write $\tilde{m}_c$ for the number of pairs in class s in which the treated unit is the first unit,

$$\tilde{m}_c = \sum_{s \in l_c} Z_{s1} \qquad \text{and write} \quad \tilde{\mathbf{m}} = \begin{bmatrix} \tilde{m}_1 \\ \vdots \\ \tilde{m}_C \end{bmatrix}.$$

Then $\tilde{m}_c$, like m_s, is a random variable, since it depends on the Z's. Note the distinction between $\mathbf{m}$ and $\tilde{\mathbf{m}}$, both of which are used below. With this notation, the distribution of treatment assignments within pairs is

$$\begin{aligned} \text{pr}(\mathbf{Z} = \mathbf{z} | \mathbf{m}) &= \prod_{c=0}^{C} \prod_{s \in l_c} \rho_c^{z_{s1}} \{1 - \rho_c\}^{z_{s2}}, \\ &= \left(\frac{1}{2}\right)^{\tilde{n}_0} \prod_{c=1}^{C} \rho_c^{\tilde{m}_c} \{1 - \rho_c\}^{\tilde{n}_c - \tilde{m}_c}. \end{aligned} \tag{3.15}$$

Unlike the distribution of treatment assignments in §3.2.3 for exact matching, the distribution (3.15) for inexact matching involves unknown parameters, the ρ_c, reflecting the remaining imbalances in $\mathbf{x}$.

Consider the conditional distribution of the treatment assignments $\mathbf{Z}$ given both $\mathbf{m}$ and $\tilde{\mathbf{m}}$. As will now be seen, conditioning on both $\mathbf{m}$ and $\tilde{\mathbf{m}}$ yields a known distribution free of the unknown ρ_c. It is a distribution on a set $\tilde{\Omega} \subseteq \Omega$, where $\mathbf{z} \in \tilde{\Omega}$ if and only if:

(i) $\mathbf{z}$ is a treatment assignment for S matched pairs; that is, $z_{s1} + z_{s2} = m_s = 1$ or, equivalently, $\mathbf{z} \in \Omega$; and

(ii) $\mathbf{z}$ exhibits the same degree of imbalance as the observed data; that is, $\tilde{m}_c = \sum_{s \in l_c} z_{s1}$ for $c = 1, \ldots, C$.

The set $\tilde{\Omega}$ has

$$\tilde{K} = 2^{\tilde{n}_0} \prod_{c=1}^{C} \binom{\tilde{n}_c}{\tilde{m}_c}$$

elements, and $\text{pr}(\mathbf{Z} = \mathbf{z}|\mathbf{m}, \tilde{\mathbf{m}}) = 1/\tilde{K}$ for each $\mathbf{z} \in \tilde{\Omega}$. In class $c = 0$ containing $\tilde{n}_0$ pairs, all $2^{\tilde{n}_0}$ possible assignments are equally likely. In class $c \geq 1$ containing $\tilde{n}_c$ pairs, the assignments give the treatment to the first unit in exactly $\tilde{m}_c$ pairs, and there are $\binom{\tilde{n}_c}{\tilde{m}_c}$ such assignments.

Using the known distribution $\text{pr}(\mathbf{Z} = \mathbf{z}|\mathbf{m}, \tilde{\mathbf{m}})$, significance levels are obtained in a manner similar to that in §2.4.1. Under the null hypothesis of no treatment effect, the statistic $T = t(\mathbf{Z}, \mathbf{r})$ has significance level

$$\text{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq T|\mathbf{m}, \tilde{\mathbf{m}}\} = \frac{|\{\mathbf{z} \in \tilde{\Omega} : t(\mathbf{z}, \mathbf{r}) \geq T\}|}{\tilde{K}}, \tag{3.16}$$

which parallels (2.5) and is simply the proportion of treatment assignments in $\tilde{\Omega}$ giving values of the test statistic at least as large as the observed value.

If the test statistic is a sum statistic in the sense of §2.4.4, that is, if $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\text{T}}\mathbf{q}$, where $\mathbf{q}$ is a function of $\mathbf{r}$, then the null expectation and variance of the test statistic are given in the following proposition. The proposition assumes the study is free of hidden bias in the sense that (3.2) holds, and it concerns the conditional expectation and variance of a sum statistic given $\mathbf{m}$, $\tilde{\mathbf{m}}$, that is, the expectation and variance of $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\text{T}}\mathbf{q}$ over the distribution (3.16).

Proposition 11 *If the study is free of hidden bias, then the null expectation and variance of* $\mathbf{Z}^{\text{T}}\mathbf{q}$ *are*

$$\begin{aligned} E(\mathbf{Z}^{\text{T}}\mathbf{q}|\mathbf{m}, \tilde{\mathbf{m}}) &= \sum_{s \in l_0} \frac{q_{s1} + q_{s2}}{2} + \sum_{c=1}^{C} \tilde{m}_c \tilde{\mu}_{c1} + (\tilde{n}_c - \tilde{m}_c)\tilde{\mu}_{c2} \\ \text{var}(\mathbf{Z}^{\text{T}}\mathbf{q}|\mathbf{m}\tilde{\mathbf{m}}) &= \sum_{s \in l_0} \frac{(q_{s1} - q_{s2})^2}{4} + \sum_{c=1}^{C} \frac{\tilde{m}_c(\tilde{n}_c - \tilde{m}_c)\tilde{\sigma}_c^2}{\tilde{n}_c}, \end{aligned} \tag{3.17}$$

where

$$\tilde{\mu}_{ci} = \frac{1}{\tilde{n}_c} \sum_{s \in l_c} q_{si} \qquad \textit{for} \quad i = 1, 2$$

and

$$\tilde{\sigma}_c^2 = \frac{1}{\tilde{n}_c - 1} \sum_{s \in l_c} \{(q_{s1} - q_{s2}) - (\tilde{\mu}_{c1} - \tilde{\mu}_{c2})\}^2.$$

Proof. The proof makes use of several elementary observations. First, the conditional distribution of $\mathbf{Z}$ given $\mathbf{m}$, $\tilde{\mathbf{m}}$ is uniform on $\tilde{\Omega}$, and as a result, the Z_{si}'s in different classes are independent of each other. Second, write $\mathbf{Z}^{\text{T}}\mathbf{q}$ as

$$\mathbf{Z}^{\text{T}}\mathbf{q} = \sum_{c=0}^{C} \sum_{s \in l_c} Z_{s1}(q_{s1} - q_{s2}) + \sum_{c=0}^{C} \sum_{s \in l_c} q_{s2}, \tag{3.18}$$

where the second sum on the right is a constant since it does not involve $\mathbf{Z}$. In class $c = 0$, the Z_{s1}'s are independent of each other given $\mathbf{m}$, $\tilde{\mathbf{m}}$, and each Z_{s1} equals 1 or 0 with probability 1/2. For $c \geq 1$, the sum $\sum_{s \in l_c} Z_{s1}(q_{s1} - q_{s2})$ in (3.18) is the sum of $\tilde{m}_c$ of the $(q_{s1} - q_{s2})$'s randomly selected without replacement from among the $\tilde{n}_c$ pairs $s \in l_c$. The proposition then follows directly from standard facts about simple random sampling without replacement. ■

3.6.3 *Return to the Example of Vasectomy and Myocardial Infarction

The 36 pairs in Table 3.8 are divided into seven classes, numbered $0, 1, \ldots, 6$, based on the pattern of imbalance in two unmatched covariates, obesity and smoking. In class 0, the two men are exactly matched for obesity and smoking. There are $\tilde{n}_0 = 13$ exactly matched pairs in class 0. In class 1, labeled (os, oS), each pair contains two men who are not obese, of whom exactly one is a smoker, and there are $\tilde{n}_1 = 11$ pairs with this imbalance, and so on. Pair $s = 3$ is the first pair in class $c = 1$, labeled (os, oS), and in this pair, the second man is a nonobese smoker, oS, who had an MI, $q_{32} = 1$, and a vasectomy, $Z_{32} = 1 = 1 - Z_{31}$.

McNemar's statistic is the number of vasectomized men who had an MI, namely, $\sum_s \sum_i Z_{si} q_{si} = 20$ of a possible 36. Proposition 11 gives the moments of the statistic adjusting for obesity and smoking in addition to the variables used to form matched pairs. Table 3.9 gives a few intermediate calculations. Using these in (3.17) gives an expectation of 19.015 and a variance of 6.813 for the McNemar statistic under the null hypothesis of no treatment effect. In words, the number of vasectomized men who had an MI, namely, 20, is quite close to the expectation 19.015 in the absence of a treatment effect. The standardized deviate with continuity correction is $(20 - 19.015 - \frac{1}{2})/\sqrt{6.813} = 0.186$, and this is small when compared with the standard normal distribution, so there is no indication of an effect of vasectomy on the risk of MI. This would be a correct test if the study were free of hidden bias once adjustments had been made for the two matched and the two unmatched covariates.

Had smoking and obesity been ignored, the usual expectation for McNemar's statistic would have been $36/2 = 18$ vasectomized MIs rather than 19.015, so the adjustment for smoking and obesity moved the expected count closer to the observed 20. In other words, if vasectomy had no effect on the risk of MI, we would nonetheless have expected more than half of the vasectomized men to exhibit MIs, because both vasectomy and MI are related to smoking and obesity in the data shown in Table 3.9.

TABLE 3.9. Intermediate Calculations for Vasectomy and Myocardial Infarction.

c	$\tilde{n}_c$	$\tilde{m}_c$	$\tilde{\mu}_{c1}$	$\tilde{\mu}_{c2}$	$\tilde{\sigma}_c^2$
0	13	6			
1	11	6	0.182	0.818	0.655
2	3	1	0.333	0.667	1.333
3	2	0	0.000	1.000	0.000
4	3	0	0.333	0.667	1.333
5	3	1	0.333	0.667	1.333
6	1	1	0.000	1.000	0.000

3.7 Bibliographic Notes

Direct adjustment is surveyed by Bishop, Fienberg, and Holland (1975, §4.3), Cochran and Rubin (1973), Fleiss (1981, §14), Kitagawa (1964), and Mosteller and Tukey (1977, §11). The analysis of matched pairs and matched sets is surveyed by Breslow and Day (1980, §5), Cochran and Rubin (1973), Fleiss (1981, §8), Gastwirth (1988, §11), and Kleinbaum, Kupper, and Morgenstern (1982, §18). Statistical procedures that may be derived from an assumption of random assignment of treatments within strata have long been applied to observational studies; see, for instance, the influential paper by Mantel and Haenszel (1959). Cochran (1965, §3.2) viewed matching, subclassification, and model-based adjustments as different ways of doing the same thing. In an important paper, Rubin (1977) demonstrated that if treatments are randomly assigned on the basis of a covariate, then adjustments for the covariate produce appropriate estimates of a treatment effect. Rubin (1978) develops related ideas from a Bayesian view. Rosenbaum (1984b, 1987a) obtains known permutation distributions by conditioning on a sufficient statistic for unknown assignment probabilities, as in §3.2.

The propensity score is proposed in Rosenbaum and Rubin (1983), and its link in §3.2.5 to permutation inference is discussed by Rosenbaum (1984b). Joffe and Rosenbaum (1999) survey and extend propensity score methods, discussing in particular propensity scores for doses of treatment and propensity scores in case-cohort studies; see also Imbens (2000).

Cole (1979, p. 16) says the first true case-referent study was conducted in 1926. Cornfield (1951) showed that case-referent sampling did not alter the odds ratio. He also argued that if the disease is rare, as it is in most case-referent studies, then the odds ratio approximates another measure, the relative risk. See also Greenhouse (1982). Mantel's (1973) paper is a careful discussion of case-referent studies; see also Hamilton (1979), Rosenbaum (1987b), and Holland and Rubin (1988). Further references concerning case-referent studies are given in Chapter 7.

Conditional tests given a sufficient statistic for the propensity score in §3.4 are discussed in Rosenbaum (1984b), along with the large sample approximation in §3.5; see also Robins, Mark, and Newey (1992) and Robins and Ritov (1997). Inexact matching followed by stratification in §3.6 is discussed in Rosenbaum (1988); however, that paper mistakenly does not mention the need for assumption (3.13) in the presented analysis of Walker's data, where the classes were based only on the unmatched covariates.

3.8 Problems

1. Problems 1 and 2 consider a test statistic motivated by linear regression, but instead of referring the statistic to a theoretical distribution, these problems consider its permutation distribution. Write $\mathbf{X}$ for the matrix with N rows, numbered $(s,i), s = 1,\ldots,S, i = 1,\ldots,n_s$ where row (s,i) is $\mathbf{x}_s^{\mathrm{T}}$. Assume that $\mathbf{X}$ has more rows than columns and has rank equal to the number of columns. Consider the linear regression model $\mathbf{R} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\tau + \mathbf{e}$, where $\mathbf{e}$ is a vector of unobserved errors, but do not assume the model is correct. The least squares estimate of τ under this model is

$$t(\mathbf{Z},\mathbf{R}) = \frac{\mathbf{R}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\mathbf{Z}}{\mathbf{Z}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\mathbf{Z}}, \qquad \text{where} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}. \quad (3.19)$$

   In a uniform randomized experiment (or in an observational study free of hidden bias), consider the permutation distribution (2.5) of this $t(\mathbf{Z},\mathbf{R})$ under the null hypothesis of no treatment effect. Show that the significance level (2.5) for the covariance adjusted estimate $t(\mathbf{Z},\mathbf{R})$ equals the significance level (2.5) for the total response in the treated group, $\mathbf{Z}^{\mathrm{T}}\mathbf{r}$, which in turn equals the significance level for the difference in sample means (2.5). (Hint: How does $\mathbf{X}^{\mathrm{T}}\mathbf{z}$ vary as $\mathbf{z}$ ranges over Ω?)

2. Continuing Problem 1, show that the same conclusion holds if the significance level is based on (3.9) rather than (2.5). That is, show that the significance level (3.9) with $t(\mathbf{Z},\mathbf{R})$ given by the covariate adjusted estimate (3.19) equals the significance level (3.9) with $t(\mathbf{Z},\mathbf{R})$ given by $\mathbf{Z}^{\mathrm{T}}\mathbf{r}$. (Hint: How does $\mathbf{X}^{\mathrm{T}}\mathbf{z}$ vary as $\mathbf{z}$ ranges over $\bar{\Omega}$?) (Rosenbaum 1984b, §2.5)

3.9 References

Becker, H. S. (1996) The epistemology of qualitative research. In: *Ethnography and Human Development*, R. Jessor, A. Colby, and R. Shweder, eds., Chicago: University of Chicago Press, pp. 53–72.