CHAPTER 5

Matching

5.1 CONFOUNDING VARIABLES

When we plan to compare the mean responses y in two or more groups of subjects from populations exposed to different experiences or treatments, the distributions of y, ideally, should be the same in all populations, except for any effects produced by the treatments. In fact, the values of y are usually influenced by numerous other variables $x_1, x_2, \ldots$ which will be called *confounding variables*. They may be qualitative or ordered classifications, discrete or continuous.

Confounding variables may have two effects on the comparison of the response y in the two populations. First, since in an observational study the investigator has limited control over his choice of populations to be studied, the distributions of one or more of the confounding variables may differ systematically from population to population. As a result, the distributions of y may also differ systematically and the comparison of the sample mean values of y may be biased. Second, even if there is no danger of bias—the distributions of a confounding variable x being the same in different populations—variations in x contribute to variability in y and decrease the precision of comparisons of the sample means.

Two simple examples will be given to show how a confounding variable may produce bias and decrease the precision of a comparison $\bar{y}_1 - \bar{y}_2$. The examples also suggest the two principal techniques used in practice to control undesirable effects of a confounding variable.

In the first example there are two samples, treated (t) and control (c), and y has a linear regression on a confounding variable x, of the same form $\alpha + \beta x$ in each population. Hence

$$y_{ti} = \alpha + \delta + \beta x_{ti} + e_{ti}$$

and

$$y_{ci} = \alpha + \beta x_{ci} + e_{ci}$$

where δ is the effect of the treatment. As given previously, the sample mean difference $\bar{d}$ is

$$\bar{d} = \bar{y}_t - \bar{y}_c = \delta + \beta(\bar{x}_t - \bar{x}_c) + (\bar{e}_t - \bar{e}_c) \qquad (5.1.1)$$

and

$$E(\bar{d}\,) = E(\bar{y}_t - \bar{y}_c) = \delta + \beta(\mu_{tx} - \mu_{cx}) \qquad (5.1.2)$$

so that the bias is of amount $\beta(\mu_{tx} - \mu_{cx})$. Further, regardless of whether $\mu_{tx} = \mu_{cx}$, we have from (5.1.1), assuming the variances of x and e are the same in each population,

$$V(\bar{d}\,) = V(\bar{y}_t - \bar{y}_c) = \frac{2}{n}\left(\beta^2\sigma_x^2 + \sigma_e^2\right) = \frac{2}{n}\left(\rho^2\sigma_y^2 + (1 - \rho^2)\sigma_y^2\right)$$

It follows, as is well known, that if we could remove the effects of variations in x, we could reduce $V(\bar{d})$ from $2\sigma_y^2/n$ to $2(1 - \rho^2)\sigma_y^2/n$. This is the main reason for the use of the methods known as the *analysis of covariance* and *blocking* in controlled experiments. By the random assignment of subjects to treatment groups and by other precautions, the investigator in a simple controlled experiment hopes that he does not have to worry about bias in the comparison $\bar{d}$. It may still be worth trying to control confounding x variables for the potential increase in the precision of $\bar{d}$.

If the linear-regression model is correct, Eq. (5.1.1) suggests two alternative methods of removing the danger of bias and increasing the precision. The first, used at the planning stage, is to select the treatment and control samples so that $\bar{x}_t$ and $\bar{x}_c$ are equal, or nearly equal. In repeated samples of this type,

$$\bar{d} \doteq \delta + (\bar{e}_t - \bar{e}_c) \qquad (5.1.3)$$

giving $E(\bar{d}) \doteq \delta$ and $V(\bar{d}) \doteq 2\sigma_e^2/n = 2\sigma_y^2(1 - \rho^2)/n$

The second method is applied at the analysis stage. From (5.1.1),

$$\bar{d} = \delta + \beta(\bar{x}_t - \bar{x}_c) + (\bar{e}_t - \bar{e}_c)$$

Hence, we compute an estimate $\hat{\beta}$ of β and estimate δ by the adjusted mean difference

$$\bar{d}' = \bar{d} - \hat{\beta}(\bar{x}_t - \bar{x}_c) = \delta + (\bar{e}_t - \bar{e}_c) - (\hat{\beta} - \beta)(\bar{x}_t - \bar{x}_c)$$

If $\hat{\beta}$ is an unbiased estimate of β, then $E(\bar{d}') = \delta$. If the effects of sampling errors in $\hat{\beta}$ were negligible, we would have $V(\bar{d}') = 2\sigma_y^2(1 - \rho^2)/n$. These sampling errors increase this value, but only trivially in large samples.

To summarize for this example, we should plan to control an x variable either if there seems to be a danger of nonnegligible bias or if a substantial gain in precision may result. Using a linear regression, we do not begin to get a substantial reduction in $V(\bar{d})$ until ρ is at least 0.4. An x variable can be controlled either by the way in which the samples are selected in the planning stage or by recording the values of x and adjusting the estimate in the analysis stage.

As a second example, suppose that x is a two-class variate and y is a proportion calculated from a $0 - 1$ variate. For the treatment (t) and control (c) populations, the proportions and the means of y in each class are given in the list below.

	Population Proportion in		Population Mean of y		
Sample	x: Class 1	Class 2	x: Class 1	Class 2	x: Overall
Treatment	f_t	$(1 - f_t)$	$\delta + p_1$	$\delta + p_2$	$\delta + f_t p_1 + (1 - f_t)p_2$
Control	f_c	$(1 - f_c)$	p_1	p_2	$f_c p_1 + (1 - f_c)p_2$

The population means of y differ in the two classes, having values p_1 and p_2 for the control population. The true treatment effect is δ in each class. Each sample has total size n, but the expected proportions f_t and f_c that fall in class 1 have been made to differ for the treatment and control samples. This difference is the source of the trouble.

If the overall treatment and control samples are randomly drawn from the populations composed of classes 1 and 2, the sample proportions $\hat{p}_t$ and $\hat{p}_c$ have means as shown in the right-most column of the list. It follows from the list that

$$E(\bar{d}) = E(\hat{p}_t - \hat{p}_c) = \delta + (f_t - f_c)(p_1 - p_2)$$

Thus there is a bias of amount $(f_t - f_c)(p_1 - p_2)$. Note that a large bias requires both a large difference in the expected proportions f_t and f_c in class 1 and a large difference in the means p_1 and p_2 in the two classes. This explains why there is sometimes only a small bias in the estimated difference $\hat{p}_t - \hat{p}_c$, even if f_t and f_c differ widely.

This result also suggests two methods of controlling bias, analogous to the methods given in the first example. At the planning stage, we could draw samples subject to the restriction that $f_t = f_c$, but otherwise drawn at random. This technique is frequently referred to as "within-class matching" or "frequency matching." Alternatively, without this restriction, we could adopt a different estimate of δ at the analysis stage. Let $\hat{p}_{t1}$, $\hat{p}_{t2}$, $\hat{p}_{c1}$ and $\hat{p}_{c2}$ be the sample estimates of the treatment and control proportions in each class from random samples. Any weighted estimate of the form

$$\bar{d}' = W_1(p_{t1} - p_{c1}) + W_2(p_{t2} - p_{c2}) \qquad (W_1 + W_2 = 1)$$

is clearly an unbiased estimate of δ under this model. In practice, the particular choice of W_1 and W_2 has varied a good deal from study to study. Some investigators choose the weights W_1 and W_2 from a standard population that is of interest (perhaps the target population), and others choose W_1 and W_2 to minimize the variance of $\bar{d}'$.

In this example there is no danger of bias if either $p_1 = p_2$ or $f_t = f_c$. Do the methods of controlling bias increase the precision in the "no bias" situations, as they did in example 1? There is no increased precision if $p_1 = p_2$. If $p_1 \neq p_2$ there is a possible increase in precision if we make $f_t = f_c$, but this is small unless p_1 and p_2 differ greatly. In my opinion this is seldom worth any extra trouble in practice.

To summarize for this example, attempts to remove bias or increase precision may be made either in the planning or the analysis stages. However, when y is a proportion and there is no danger of bias, the increase in precision resulting from these attempts is usually small or moderate.

In handling the problem of confounding variables, the investigator should first list the principal confounding variables that he recognizes, in order of their importance in influencing y, inasmuch as this can be judged. A decision is made to exercise some control over an x variable either if the possibility of a nonnegligible bias exists in the y comparison or if a substantial gain in precision may result.

A second requirement about any x variable that we plan to control is that its value should not be influenced by the treatments to be compared. Suppose that in the first example the values of x are higher in the treatment than in the control population because the treatment affects x, and that x and y are positively correlated. In this case, subtracting $\hat{\beta}(\bar{x}_t - \bar{x}_c)$ from $\bar{d}$ removes part of the treatment effect on y.

This mistake is avoided when the x variables are measured before the introduction of the treatment, but the danger exists whenever the measurement is subsequent to the introduction of the treatment. For example, Stanley (1966) cites a study intended to measure the effect of brain damage existing at birth on the arithmetic-reasoning ability of 12-year-old boys. In comparing samples of brain-damaged and undamaged boys, the investigator might be inclined to use current measures of other kinds of ability, or parental socioeconomic status as confounding x variables whose effects are to be removed by matching or regression adjustment. Other kinds of ability at age 12 might obviously be affected by the brain damage, and as Stanley points out, even parental socioeconomic status might also be affected because of the strain and cost of medical care for the brain-damaged child.

In the subsequent discussion of specific techniques for handling confounding variables, it is necessary to keep in mind the scales of measurement of both x and y. With x, the principal distinction is between a classification and a discrete or continuous variable; with y the principal distinction is between a proportion derived from a $0-1$ variate and a continuous or discrete variable. This chapter discusses methods used at the planning stage for handling confounding variables. Adjustments in analysis are discussed in Chapter 6.

5.2 MATCHING

In matching, a confounding x variable is handled at the planning stage by the way in which the samples for different treatment groups are constructed. In some methods each member of a given treatment group has a match or partner in every other treatment group, where the partners are within defined limits in the values of all x variables included in the match. The number of x variables matched in applications may range from 1 to as many as 10 or 12.

From inspection of medical journals, Billewicz (1965) reports that the numbers of variables most often matched in medical studies were two or three. The idea of "matching" is the same as that known as "pairing" or "blocking" in experimentation. As a rule, matching is confined to smaller studies of simple structure—most commonly, two-group comparisons. The more complex the plan, the more difficult it will be to find matches. What is meant by a match? This depends on the nature of the confounding variable x.

x a Classification. A match usually means belonging to the same class. With three classified x variables having two, four, and five classes, respectively, a match on all three variables is another subject in the same cell of the $2 \times 4 \times 5 = 40$ cells created by this three-way classification.

x Discrete or Continuous. Two procedures are common: One is to change x into a classification variable (e.g., ages arranged in five-year classes) and as before regard a match as someone in the same class. This method is common if, say, two of three x variables are already classification variables, the third variable being originally continuous. This method will be called *within-class matching* (other terms used are "stratified matching" and "frequency matching").

With x discrete or continuous a second method is to call two values of x a match if their difference lies between defined limits $\pm a$. This method is called *caliper matching*, a name suggested by Donald Rubin (1970). With x continuous, within-class matching is more common than caliper matching, for which it is harder to find matches.

Two advantages of matching are that the idea is easy to grasp and the statistical analysis is simple. Perfect caliper matching on x removes any effects of an x variable, whatever the mathematical nature of the relation between y and x, provided that this relation is the same in the populations being compared. No assumption of a linear regression of y on x is required. To verify this, suppose that for the jth subject in population 1 the relation between y and x is of the general form

$$y_{1j} = \delta_1 + f(x_{1j}) + e_{1j}$$

In population 2,

$$y_{2j} = \delta_2 + f(x_{2j}) + e_{2j}$$

where $f(x)$ has any functional form and e_{1j} and e_{2j} have means zero.

Then if $x_{1j} = x_{2j}$ for all j,

$$\bar{d} = \bar{y}_1 - \bar{y}_2 = \delta_1 - \delta_2 + \bar{e}_1 - \bar{e}_2$$

and is unbiased whatever the nature of the function $f(x)$. If matching is not perfect but fairly tight, the hope is that for any continuous function $f(x)$ we will have $f(x_{1j}) \doteq f(x_{2j})$ because $x_{1j} \doteq x_{2j}$, and the remaining bias in $\bar{y}_1 - \bar{y}_2$ will be small.

Matching has some disadvantages. The long time taken to form matches, may hardly seem worthwhile if under the original matching rules no matches can be found for some members of one sample. Imperfect matching on the chosen variables or omission of important variables on which we failed to match can leave systematic differences between the members of a matched pair. Billewicz cites an example by Douglas (1960), in which children from premature births ($5\frac{1}{2}$ lb or less) were found to have inferior school performance at ages 8 and 11 to normal-birth children. The original sample size was 675; samples were matched with regard to sex, mother's age, social class, birth rank in the family, and degree of crowding in the home. It became evident, however, as the study progressed that despite the matching, systematic differences remained between the parents of premature- and normal-birth children regarding (1) social level, (2) maternal care, and (3) interest in school progress. Each matched pair of children was assigned a score (from $+3$ to -3) according to the extent to which these three variables favored the premature child. The mean differences in the exam results (of 11-year-old premature- and normal-birth children) appeared as follows when subclassified by this score. (Results for the exam of eight-year-olds were similar.)

Premature–Normal Scores On Confounding Variables	Exam Scores
+3	+6.0
+2	+0.4
+1	−0.6
0	−1.7
−1	−5.6
−2	−6.7
−3	−12.0

Clearly, the original matching did not guarantee that partners were equivalent on all important confounding variables, even after matching on five variables. (Firm interpretation of this finding rests, in part, on being certain that neither maternal care nor interest in school progress is affected by the comparison variables—premature versus normal birth.)

With x continuous, some results for two other matching methods will be presented later in this chaper. One method called *mean matching* (or "balancing") does not attempt to produce closely matched individual pairs, but instead concentrates on making $\bar{x}_1 - \bar{x}_2$ as small as possible. This method is not new. It is of course tied to the assumption that y has the same linear regression on x in each population. The second method, called *nearest available matching*, tries to produce well-matched pairs in difficult situations and is described later.

In studying the performance and properties of various matching procedures, we shall consistently use x_1, n_1, σ_1, and so forth, to relate to the group of observations *for* which matching observations are being sought. They are sought *from* reservoir 2, characterized by entities x_2, n_2, σ_2, and so forth, all bearing the subscript "2." This notational asymmetry is critical for correctly interpreting tables and formulas.

5.3 THE CONSTRUCTION OF MATCHES

Published reports of studies using matching illustrate that the practical difficulties in constructing matched samples vary greatly from study to study. Numerous factors are relevant. In order to form a matched sample of size n from two or more populations, we obviously need larger supplies or reservoirs of subjects from which matches may be sought. The sizes and accessibility of these reservoirs are important. The most difficult case is one in which the investigator has only n subjects available from one population and needs all of them. Unless the reservoir from a second population is much larger than n, the investigator may be unable to find matches for all n from population 1.

Another factor is the planned size of the sample to be compared. Matching is seldom used when this planned n exceeds say 500, presumably because of the labor and time needed to match. The difficulties of finding matches under fixed rules also mount rapidly with each increase in the number of x variables to be matched. In Section 5.2 the Douglas example included a reservoir of 12,000 normal births, where $n = 675$ with five matched variables.

Matching of samples from two populations becomes more difficult when the x distributions differ markedly in the two populations—the situation in which the risk of bias due to x is greatest. If most people in population 1 are older than those in population 2, it may be impossible to find good matches for the oldest members of a sample from population 1. This difficulty is sometimes handled by omitting sample members who cannot be matched by the original rules, rather than by relaxing the matching rules. The full consequences of the possible alternatives have not been investigated. If the regression of y on x is the same function in both populations, omission may be the better procedure, though it means that one of the samples is badly distorted at one end.

The time taken to create matched samples depends much on the ease with which we can locate sample members whose x values are in the desired range. Sometimes, the available records are kept in a form that facilitates this search, as might happen if a match for a newborn baby is as follows: a child of the same sex, born in the same hospital during the same week with no complications of delivery. When one has to seek matches by going through a reservoir case by case, chance plays a major role in determining how long the job takes. For illustration, suppose that x is a five-class variable and that the sample from population 1 has only $n = 100$ cases available, distributed as follows:

Class	1	2	3	4	5	Total
Number	10	25	30	25	10	100

In population 2 the proportions falling in these five classes are assumed as follows:

Class	1	2	3	4	5	Total
Proportions	0.038	0.149	0.269	0.326	0.218	1.000

In classes 4 and 5, it appears that a search of less than 100 cases from population 2 should provide the needed 25 and 10 matches, but more than 100 seem necessary, on the average, for the other three classes. The most difficult case is class 1, where we need 10 matches but expect only 3.8 from a sample of 100. The number of reservoir cases needed to find these 10 matches is a random variable following a simple waiting-time distribution [Feller (1957)]. The mean number needed is $10/0.038 = 263$, or more generally m/p, where m is the number required and p is the proportion. This number has a large variance $m(1 - p)/p^2 = 3835$ in this example. The consequence is that the upper 95% point of the waiting-time distribution exceeds 350. Thus we might be lucky and find the 10 matches needed in class 1 in 150 cases, or we might be unlucky and have to search over 350 cases. This uncertainty makes the case-by-case construction of matches from random samples frustrating, particularly when there are several x variables.

For this reason, matching is inadvisable if potential sample members from the different populations become available at the rate of only a few per week, for example, subjects entering an agency for a service of some kind. There may be an indefinite delay while waiting for matches for certain subjects. This point is discussed more fully by Billewicz (1965).

Computers should be able to perform much of the detailed labor of finding matches if the values of the x variables in the available reservoirs are in a form suitable for input into computers. The easiest case is within-class matching when all the x variables, three for example, are already in classified form. For each reservoir, simple instructions will arrange and list in a printout the sample in each cell of the three-way classification. We learn, for instance, that the first reservoir has 19 cases in cell 2; the second reservoir has 28 cases. For any desired sample size up to 19 from this cell, the partners can be drawn at random by the computer from the 19 or 28 cases available. If two or three x variables to be matched are already in classified form, while the third is a continuous variable that is to be classified for within-class matching, the computer will perform this classification, arrange each reservoir in cells, and list as before.

With a single discrete or continuous x for which caliper matching is desired, the computer can rank and list the values of x in each reservoir from lowest to highest. From these lists caliper matches can be made quickly if available. I do not know how best to extend this method to two or three continuous x's where a caliper match is needed for each. A partial help is to have the computer classify each x into one of 2^m classes by binary splitting. Thus with three x's and four classes per variable, the computer arranges the trinomial distribution into 64 cells and lists. Since the values of x are not strictly in rank order within a cell, such lists are less convenient, but still a considerable help in searching for caliper matches on all three x's.

In matching samples from two populations with one continuous x, the investigator sometimes needs to use *all* cases in the reservoir from population 1. A sample of at least 100 cases is needed and there are only 100 cases from population 1. The reservoir from population 2 has say 272 cases. In this case it is not clear that caliper matching of each case in sample 1 can be performed with a prechosen fixed $\pm a$ value. The investigator does not want to reject any cases, since this reduces the sample size below 100. For this problem, Donald Rubin (1970) has developed a method called *nearest available pair matching*, performed entirely by computer, that attempts to do the best job of matching subjects to the restriction that every member of sample 1 must be matched.

The computer first arranges sample 1 in random order. For the first member of sample 1 it picks out the member of the reservoir for sample 2 that is nearest to it and lays this pair aside as the first match. The process is repeated for the second member of sample 1 with respect to the 271 items remaining in the reservoir, and so forth. Thus all matches are found, though, of course, the difference $|x_{1j} - x_{2j}|$ will differ from pair to pair.

Two variants of this method that might be better were also examined by Rubin. Instead of arranging sample 1 in random order the computer first ranks the sample 1 members from lowest x_1 to highest x_n. In variant 1 we seek matches from reservoir 2 in the order $x_n, x_{n-1}, \ldots$ (high–low). In variant 2 the order is $x_1, x_2, \ldots$ (low–high).

In mean matching, the objective is to make $\bar{x}_1 - \bar{x}_2$ as small as possible for any x. If all members of sample 1 must be used, their mean $\bar{x}_1$ is first found. The computer selects from reservoir 2 the value x_{21} nearest to $\bar{x}_1$. Then x_{22} is chosen such that $(x_{21} + x_{22})/2$ is nearest to $\bar{x}_1$ and so forth.

5.4 EFFECT OF WITHIN-CLASS MATCHING ON x

The problem of the effects of matching is complex, and not nearly enough is known about it. As we have mentioned, the purposes in matching are (1) to protect against bias in $\bar{y}_1 - \bar{y}_2$ that might arise from differences between the x distributions in different populations to be compared and (2) to increase the precision of the comparison of the y means. This section discusses the effect of within-class matching on x. We must first consider the nature of the x variable. There are three possibilities:

1. *An "Ideal" Classification.* This term is used for classifications in which two members of the same class are identical with regard to x. Within-class matching therefore gives perfect matching, which as previously noted removes bias in y for any functional form of the relationship between y and x that is the same in both populations. Unfortunately, it is not clear how often ideal classifications occur in practice. This might be so with a qualitative classification like the O, A, B, AB blood types in which two subjects with the same blood type are identical with regard to any effect of the relevant genes on y.

Sex (male, female) might be an ideal classification for some types of response y, but not for other types. For instance, in traits related to behavior or attitudes, it is natural to think of some women as more feminine than other women, and some men as more masculine than other men; then the male–female classification would have classes of nonidentical members.

2. *A Classification with Any Underlying Distribution.* Numerous examples can be given of classifications in which members of the same class need not be identical with regard to the variate which x is designed to measure. Consider an urban–rural classification. Many aspects of urban–rural living that are likely to influence y, are themselves affected by the fact that some people in the urban class have a more typically city environment than others in the urban class; likewise, some people in the rural class have a more typically country environment then others in the rural class. The same is true, for some responses, of a classification by religion into Catholic, Jewish, and Protestant. Some subjects of a given religion are much more heavily committed to religious beliefs and activities than are other subjects.

Ordered classifications such as socioeconomic level or degree of interest (none, little, much) in some topic are a more-obvious example. One can sense an underlying continuous x variable that has been divided into a small number of ordered classes. Indeed, ordered classifications are often used when we recognize that a correctly measured x would be continuous, but can measure only crudely, so that an ordered classification seems all that the measuring instrument will justify.

In examining the effect of within-class matching on bias when there is an underlying distribution, I assume that the correct x (the value that influences y) is continuous and that the observed classified x represents a grouping of the correct x into ordered classes. Consider how matching affects $\mu_1 - \mu_2$, the mean of $\bar{x}_1 - \bar{x}_2$. If $\mu_1 > \mu_2$ it seems plausible that within most classes the mean μ_{1j} will exceed μ_{2j}, where j stands for the class. This is true for common unimodal distributions such as two normals or two t variates with different means. In the matched samples the mean of $\bar{x}_1 - \bar{x}_2$ is $\Sigma W_j(\mu_{1j} - \mu_{2j})$, where W_j are the relative numbers in the classes. Thus it seems likely that even after matching, the mean of $\bar{x}_1 - \bar{x}_2$ will tend to be positive, though its size should be limited by the width of the classes.

3. *x Discrete or Continuous.* Since we are discussing within-class matching, we assume that discrete or continuous x's are grouped into a limited number of classes before matching. Consequently, mathematical study of the effect of matching on $\bar{x}_1 - \bar{x}_2$ follows the method just indicated, except that we are now interested also in the optimum choice of class boundaries and in the effects of different numbers of classes, since these are under our control when we create the classified x variable.

With x distributed as $N(B, 1)$ in population 1 and as $N(0, 1)$ in population 2, the percent bias removed by matching was first calculated for $0 \leqslant B \leqslant 1$ for specified division points $x_0, x_1, \ldots, x_c$ (with c classes). The value $B = 1$ was considered a larger initial bias than would be typical in practice. The percent bias removed is

$$100[1 - (\mu_1 - \mu_2)_m/B]$$

where $(\mu_1 - \mu_2)_m$ is the mean of $\bar{x}_1 - \bar{x}_2$ from the matched samples. The percent bias removed was found to be practically constant in the range $0 \leqslant B \leqslant 1$ and therefore could be approximated by calculus methods for B small [Cochran (1968)].

Let the distribution of x be $f(x)$ in population 1 and $f(x - B)$ in population 2, where $f(x)$ has unit SD (standard deviation). By the calculus method, the percent reduction in bias was found to be

$$100 \sum_{j=1}^{c} M_j\left[f(x_{j-1}) - f(x_j)\right] \qquad (5.4.1)$$

where M_j is the mean value of x from $f(x)$ in the interval (x_{j-1}, x_j). In particular, for x normal,

$$M_j P_j = \frac{1}{\sqrt{2\pi}} \int_{x_{j-1}}^{x_j} x e^{-x^2/2}\, dx = f(x_{j-1}) - f(x_j)$$

where P_j is the total frequency in class j. Thus, for the normal, (5.4.1) gives

$$\text{Percent reduction in bias} = 100 \sum_{j=1}^{c} \frac{[f(x_{j-1}) - f(x_j)]^2}{P_j} \quad (5.4.2)$$

It happens that for the normal distribution, (5.4.2) is also the percent reduction in the *variance* of $\bar{x}_1 - \bar{x}_2$ due to matching, a result that follows immediately from results given in other cases by Ogawa (1951) and D. R. Cox (1957). For numbers of classes between 2 and 10, these authors also determined the optimum boundaries (shown here as the optimum relative class sizes), the corresponding maximum percent reductions in bias and variance for x normal, and the percent reductions when the classes are made equal in relative frequency. These values are given in Table 5.4.1.

With the optimum boundaries at least five classes are necessary to remove 90% or more of an initial bias in $\bar{x}_1 - \bar{x}_2$. Only 64% is removed with two classes and 81% with three classes. This is disappointing because matching with three classes is not uncommon, sometimes because only three classes are given in an ordered classification. With the optimum boundaries the central classes are larger, in terms of frequency, than the extreme classes. It is noteworthy, however, that with equal-sized classes the percentage reductions in bias and variance are only around 2% less than the maximum reductions. The choice of class boundaries and resultant sizes is not critical.

For equal-sized classes, some investigations of nonnormal distributions (Cochran, 1968) found that the percent reductions in bias agreed quite well with those for the normal—running about 2% higher in some cases. In variance, the percent reductions tended to fall below those for the normal as skewness and kurtosis increase.

Unequal variances were also examined where x follows $N(B, 1)$ in population 1 and $N(0, \sigma_2^2)$ in population 2. For values of σ_2^2 between $\frac{1}{2}$ and 2, it appeared that the percent reductions in the bias of $\bar{x}_1 - \bar{x}_2$ differed little from the values given for $\sigma_2^2 = 1$.

The effect of within-class matching on the bias of $\bar{x}_1 - \bar{x}_2$ are presented in Table 5.4.1. Results for the effects of caliper matching, nearest-neighbor matching, and mean matching on $\bar{x}_1 - \bar{x}_2$ will be given in Sections 5.5–5.7; the situation with respect to the bias of $(\bar{y}_1 - \bar{y}_2)$ will be discussed in Section 5.8.

Table 5.4.1. Optimum Sizes of Classes and Percent Reductions in Bias and Variance of $\bar{x}_1 - \bar{x}_2$ Due to Within-Class Matching (x Normal)

Number of Classes	Optimum Class Frequencies (%)[a]	Percent Reductions	
		Maximum	Equal Classes
2	50	63.7	63.7
3	27, (46)	81.0	79.3
4	16, 34	88.2	86.1
5	11, 24, (30)	92.0	89.7
6	7, 18, 25	94.2	91.9
7	5.5, 14, 20, (21)	95.6	93.4
8	4, 11, 16, 19	96.7	94.5
9	3, 8, 13, 17, (18)	97.2	95.4
10	2, 7, 11, 14, 16	97.6	95.9

[a] Since the distribution is symmetrical, only the lower half is shown, starting with the lowest class. Thus for $c = 4$, the frequencies are 16, 34, 34, 16 in percentages; for $c = 5$, they are 11, 24, 30, 24, 11.

5.5 EFFECT OF CALIPER MATCHING ON x

Caliper matching is a tighter and more-efficient method than within-class matching and can be used when x is continuous or discrete. For x continuous, there is a certain inconsistency in within-class matching. For example, when we search for a match for a subject whose true x value is 59.4, we may reject a subject whose x value is 60.2 because this subject is in the next-higher class, but we may accept a subject whose value is 42.1 because this subject is in the same class.

With caliper matching to within $\pm a$, the frequency functions of x in the two groups (populations 1 and 2) were assumed to be $N(B, 1)$ and $N(0, 1)$. As with within-class matching, the percent of the bias removed in $\bar{x}_1 - \bar{x}_2$ is fairly constant for values of an initial bias B which are typical of those values that occur in practice. For a given $f(x)$, the percent depends primarily on a or, more generally, on the ratio a/σ.

The amount of bias removed also depends, to some extent, on how the caliper matching is done. One method starts with a sample from population 1 and finds matches for all its members. Thus in the matched pairs, the members from sample 1 still represent an undistorted sample from population 1. However, if say $\mu_1 > \mu_2$, the matches selected will make the members from population 2 a selected sample that is biased upwards. If instead we start with ample reservoirs from both populations and search for the matches that can be found most quickly, we will tend to select members on the low-bias side from population 1 and on the high-bias side from population 2, since these are the easiest to match. An extreme form of this approach is to assume that we start with a random sample of the differences $x_{1j} - x_{2j}$ and go through this sample selecting the pairs that are caliper-matched. This approach results in smaller values of $|x_{1j} - x_{2j}|$ for matched pairs because of the distortion of *both* initial samples.

Table 5.5.1. Percent Reduction in Bias of $\bar{x}_1 - \bar{x}_2$ with Caliper Matching to Within $\pm a$ (Normal Distribution). In (1) one Sample is Random and in (2) Matches are Made from $x_{1j} - x_{2j}$

$\pm a$	Percent Reduction (1)	Percent Reduction (2)
0.2	99	99
0.3	97	98
0.4	95	97
0.5	93	96
0.6	90	94
0.7	87	92
0.8	84	90
0.9	80	87
1.0	76	84

For x normal, Table 5.5.1 shows the percent bias removed when (1) one sample is random and (2) matches are made from random differences. These are intended to indicate the range of performance in applications.

By comparison with Table 5.4.1, which gives percent reductions in bias due to within-class matching, caliper matching to within $\pm 0.9\sigma$, which seems quite loose, should be as good as within-class matching with three classes, and removes 80% or more of the bias. Caliper matching to within $\pm 0.4\sigma$ is as good as within-class matching with nine classes, and removes 95% of the bias. However, the benefits of caliper matching have a cost—caliper matching, in general, requires much larger reservoirs and more time.

When there is no bias, the percent reductions in the variance of $\bar{x}_1 - \bar{x}_2$ with caliper matching approximate the higher values for bias removed, that is, the percent (2) values presented in Table 5.5.1.

With unequal variances in the two populations and x having frequency functions $N(B, 1)$ in population 1 and $N(0, \sigma_2^2)$ in population 2, caliper matching for a given a does a little better than the values presented in Table 5.5.1 when $\sigma_2^2 > 1$, and somewhat worse when $\sigma_2^2 < 1$.

5.6 EFFECT OF "NEAREST AVAILABLE" MATCHING ON x

Rubin's (1970) results for the effects on bias will be quoted for x normal. [For a more extensive treatment see Rubin (1973, a, b).] It is assumed that all n_1 available subjects from population 1 are to be matched, and that the

reservoir from population 2 has N_2 subjects. The effect of the technique naturally depends on the ratio N_2/n_1. It also depends on the size of the initial bias B, unlike the previous methods, because with B positive ($\mu_1 > \mu_2$) and substantial, we can expect only "poor" nearest neighbors for the highest members of sample 1.

Table 5.6.1 gives the percent reductions in bias of $\bar{x}_1 - \bar{x}_2$ for $n_1 = 50$, 100; $N_2/n_1 = 2, 3, 4$; and $B/\sigma = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$, with σ assumed the same in both populations. My opinion is that $B/\sigma = \frac{1}{4}, \frac{1}{2}$ is more representative of the sizes of initial bias that occur in practioe than is $B/\sigma = \frac{3}{4}, 1$, which seems unusually large, although initial age biases for cigar and pipe smokers in the studies on smoking (Section 2.4) were around 0.7σ.

The results suggest that with a reservoir in population 2 four times as large as the sample from population 1, the method should remove nearly all the bias for initial biases up to $\frac{3}{4}\sigma$. For moderate biases (up to $\frac{1}{2}\sigma$) a 2 : 1 ratio of reservoir to sample may be expected to remove around 90% of an initial bias.

Any difference in the population standard deviations is also relevant. The percent-bias-removed values are higher than those shown in the table if $\sigma_1 < \sigma_2$ and lower if $\sigma_1 > \sigma_2$, again because of the problem of matching the highest members of sample 1.

Rubin also investigated the two variants of "nearest available" matching: (1) High–low, in which the members of sample 1 are ranked from high to low instead of at random—the highest member of sample 1 paired first, and so forth. (2) Low–high, in which the pairing proceeds from the lowest to the highest member of sample 1. For $\mu_1 > \mu_2$ Rubin found the "low–high" method best, the random method second best, and the "high–low" method third best, although the differences in performance were not great.

As with caliper matching, "nearest available" matching removed a somewhat higher percentage of the initial bias when $\sigma_2^2 > \sigma_1^2$ and removed a

Table 5.6.1. Percent Reduction in Bias of $\bar{x}_1 - \bar{x}_2$ with "Nearest Available" Matching

n_1	$\frac{N_2}{n_1}$	$\frac{B}{\sigma} = \frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
50	2	92	86	77	67
	3	96	94	90	84
	4	98	97	96	89
100	2	94	89	79	68
	3	98	96	92	85
	4	99	98	96	91

lower percentage when $\sigma_2^2 < \sigma_1^2$. With $\sigma_2^2 < \sigma_1^2$, the matches for the highest members of sample 1 tend to be too low, since population 1 has both a higher mean and a higher variance than population 2.

5.7 EFFECT OF MEAN MATCHING ON x

Table 5.7.1, which has the same format as Table 5.6.1, shows that the computer method of mean matching is highly successful in removing the bias of $\bar{x}_1 - \bar{x}_2$, as might be expected. The results in Tables 5.6.1 and 5.7.1 were obtained by computer simulation with x normal.

5.8 EFFECTS ON BIAS OF $\bar{y}_1 - \bar{y}_2$

The preceding results on the effects of different matching techniques on a confounding x variable were given as a step toward examining the effects of matching on the response variable y. These effects depend on the nature of the regression of y on x. First, consider bias with two populations. Several cases may be distinguished:

1. *Linear Regression—The Same in Both Populations.* With a single x and y continuous, the model is

$$y_{1j} = \alpha + \delta + \beta x_{1j} + e_{1j}; \qquad y_{2j} = \alpha + \beta x_{2j} + e_{2j}$$

where the constants α and β that define the regression are the same in both populations, and δ represents a constant effect of the difference in treatment. Since

$$E(\bar{y}_1 - \bar{y}_2) = \delta + \beta(\mu_1 - \mu_2)$$

Table 5.7.1. Percent Reduction in Bias of $\bar{x}_1 - \bar{x}_2$ with Mean Matching

n_1	$\dfrac{N_2}{n_1}$	$\dfrac{B}{\sigma} = \dfrac{1}{4}$	$\dfrac{1}{2}$	$\dfrac{3}{4}$	1
50	2	100	99	91	77
	3	100	100	99	96
	4	100	100	100	100
100	2	100	100	96	80
	3	100	100	100	98
	4	100	100	100	100

it is clear that under this model the percent reduction in the bias of y equals that in x.

Next, suppose that y has a multiple linear regression on k variables, $x_1, x_2, \ldots, x_k$, to which matching has been applied. In this case

$$E(\bar{y}_1 - \bar{y}_2) = \delta + \sum_{i=1}^{k} \beta_i(\mu_{1i} - \mu_{2i})$$

The ratio of the final to the initial bias in y may be written

$$\frac{\sum_{i=1}^{k} \beta_i(\mu_{1i} - \mu_{2i})_m}{\sum_{i=1}^{k} \beta_i(\mu_{1i} - \mu_{2i})} \tag{5.8.1}$$

where $(\mu_{1i} - \mu_{2i})_m$ represents the difference in means for the ith x variable after matching. If the matching technique that is applied to each x produces the same percent reduction in bias, this is also the percent reduction in bias of y.

If the matching technique produces different percent reductions in bias for different x variables, the effect on y under multiple linear regression can be more complex. When all the terms $\beta_i(\mu_{1i} - \mu_{2i})$ have the same sign, the percentage of bias removed by matching lies between the least and greatest percentages for the individual x's, as follows from (5.8.1). However, if the terms $\beta_i(\mu_{1i} - \mu_{2i})$ have different signs, it is easy to construct cases in which the initial bias in y is small, because of cancellation of signs, and is increased by matching.

2. *Nonlinear Regression—the Same in Both Populations.* If the regression of y on x is $\phi(x)$, we are concerned with the effect of matching methods on the mean of $\phi(x_{1j}) - \phi(x_{2j})$. Some investigation has been made of monotone, moderately curved regressions such as $c_1x + c_2x^2$, for c_1, c_2, and x all positive and $e^{x/2}$ or $e^{-x/2}$.

In these cases the condition $\sigma_1 = \sigma_2$ (the variance of x is the same in the two populations) becomes important. With $\sigma_1 = \sigma_2 = \sigma$, consider $\phi(x) = c_1x + c_2x^2$, where x follows $N(\mu + \frac{1}{2}, 1)$ in population 1 and $N(\mu, 1)$ in population 2, with $\mu \geqslant 4$ so that negative values of x are rare. With either within-class or caliper matching, the percent reductions in $E[\phi(x_{1j})] - E[\phi(x_{2j})]$ are close to those in $E(x_{1j}) - E(x_{2j})$. These results are also suggested by the following algebraic argument. Let

$$y_{ij} = c_0 + c_1x_{ij} + c_2x_{ij}^2 + e_{ij} \qquad (i = 1, 2)$$

Then, apart from any treatment effect, the initial bias in $\bar{d} = \bar{y}_1 - \bar{y}_2$ is

$$E(\bar{d}) = c_1(\mu_1 - \mu_2) + c_2(\mu_1^2 + \sigma_1^2 - \mu_2^2 - \sigma_2^2) \tag{5.8.2}$$

where x_{ij} has mean μ_i and variance σ_i^2. With $\sigma_1^2 = \sigma_2^2$ this becomes

$$E(\bar{d}) = (\mu_1 - \mu_2)[c_1 + c_2(\mu_1 + \mu_2)]$$

The proportionate effect of matching on $E(\bar{d})$ will therefore equal its effect on $\mu_1 - \mu_2$, except that in (5.8.2) the value of $\mu_1 + \mu_2$ is slightly altered by matching and the variances σ_{1m}^2 and σ_{2m}^2 will not be exactly equal in the population created by matching.

However, when $\sigma_1 \neq \sigma_2$, the percentage reduction in bias due to matching depends on the size and sign of the term $c_2(\sigma_1^2 - \sigma_2^2)$ in the initial bias, and on how matching affects population variances. Taking $y = x^2$, with $\mu_1 = 4.5$ and $\mu_2 = 4.0$, within-class matching with two classes reduces the bias by 63% when $\sigma_1^2 = \sigma_2^2 = 1$; by 70% when $\sigma_1^2 = \frac{2}{3}$ and $\sigma_2^2 = \frac{4}{3}$; and by only 52% when $\sigma_1^2 = \frac{4}{3}$ and $\sigma_2^2 = \frac{2}{3}$. Some results for within-class matching, with $y = e^{x/2}, y = e^{-x/2}$, and $\mu_1 - \mu_2 = 0.5$, are given in Table 5.8.1.

When $\sigma_1 = \sigma_2$, results for the effect of within-class matching on x can apparently be used as a rough, if slightly optimistic, guide to its effect on monotone, moderately curved regressions. This is not so when there are substantial differences in variances. Rubin's results for "nearest available" matching will be given in Chapter 6 for comparison with regression adjustments.

Mean matching is highly successful in diminishing $\mu_1 - \mu_2$ in the matched samples and is essentially intended to cope with a linear regression of y on x. Its performance under nonlinear regressions will depend both on the nature of the regression and on the specific method of mean matching. Rubin's method, for instance, will tend to make σ_2^2 less than σ_1^2 in the matched populations even if they are initially equal, since it chooses values of x_{2j} near to $\bar{x}_1$. Rubin's method is likely to do poorly on regressions like $e^{x/2}$, even if $\sigma_1 = \sigma_2$ initially, and should be avoided if nonlinear regressions are suspected; other matching methods are preferable to nearest available pair matching.

Table 5.8.1. Percent Reductions in Bias of y for Within-Class Matching When $y = x, y = e^{x/2}$ and $y = e^{-x/2}$

$E(y\|x)$	x	$e^{x/2}$			$e^{-x/2}$		
$(\sigma_{1x}^2, \sigma_{2x}^2)$	(1, 1)	(1, 1)	$(\frac{2}{3}, \frac{4}{3})$	$(\frac{4}{3}, \frac{2}{3})$	(1, 1)	$(\frac{2}{3}, \frac{4}{3})$	$(\frac{4}{3}, \frac{2}{3})$
Two Classes	64	61	94	44	61	47	88
Three Classes	79	76	105[a]	60	76	64	110[a]
Four Classes	86	84	108[a]	69	84	73	107[a]

[a] Entry 105 denotes that remaining bias is 5% of the original bias, but of opposite sign. Other entries exceeding 100 are similarly interpreted.

3. *Regression—Different in the Two Populations.* The concept of matching is geared to the assumption that the regression of y on x is the same in the populations being compared. Suppose that there are different linear regressions in the two populations, namely,

$$y_{1j} = \alpha_1 + \delta + \beta_1 x_{1j} + e_{1j}; \qquad y_{2j} = \alpha_2 + \beta_2 x_{2j} + e_{2j} \tag{5.8.3}$$

where, as usual, δ is the effect of the difference in treatment. Then

$$E(\bar{y}_1 - \bar{y}_2) = \delta + (\alpha_1 - \alpha_2) + \beta_1\mu_1 - \beta_2\mu_2 \tag{5.8.4}$$

Matching will not affect the term $\alpha_1 - \alpha_2$, which represents a constant bias. Further, even if we succeed in making $\mu_1 = \mu_2 = \mu$, a bias $(\beta_1 - \beta_2)\mu$ remains after matching. It is best to avoid matching in this case.

Unfortunately, it is possible only in a before–after study to detect from the data that this situation exists before deciding whether to match. In such a study, y and x are measured in two populations *before* any difference in treatment has occurred, and also *after* a period of exposure to the two treatments. From the "before" data, the regressions of y on x in the two populations can be estimated and compared at a time when the populations have no difference in treatment. The existence of a model like (5.8.3) can thus be detected before a decision on matching is made. In an "after only" study, the measurement of y is usually postponed until after the samples have been selected (i.e., the decision to match or not to match has already been made).

In the final results in an "after only" study, the finding of different linear regressions in the two populations has another possible interpretation. Assuming $\alpha_1 = \alpha_2$, from (5.8.3) it follows that for a given value of x

$$y_1 - y_2 = \delta + (\beta_1 - \beta_2)x + e_1 - e_2 \tag{5.8.5}$$

This relation might hold because the effect of the difference in treatments is $\delta + (\beta_1 - \beta_2)x$, varying with the level of x. If we have matched on x, a linear regression of $\bar{y}_1 - \bar{y}_2$ on x would reveal this situation. In fact, in many studies the investigator expects the effect of the difference in treatments to vary with x. Of course the investigator could be misled in this interpretation if the regressions actually have different slopes in the two populations.

5.9 EFFECT OF MATCHING ON THE VARIANCE OF $\bar{y}_1 - \bar{y}_2$

Let us now consider how matching increases precision when there is no danger of bias. Consider a linear regression of y on x, which is the same in both populations, with $\mu_1 = \mu_2$. For this,

$$V(\bar{y}_1 - \bar{y}_2) = \beta^2 V(\bar{x}_1 - \bar{x}_2) + V(\bar{e}_1 - \bar{e}_2) \tag{5.9.1}$$

Table 5.9.1. Percent Reduction ($f\rho^2$) in $V(\bar{y}_1 - \bar{y}_2)$ due to Matching in the "No Bias" Situation (Linear Regression)

Number of					ρ			
Classes	f	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.64	6	10	16	23	31	41	52
3	0.81	7	13	20	29	40	52	66
4	0.88	8	14	22	32	43	56	71
5	0.92	8	15	23	33	45	59	75
∞	1.00	9	16	25	36	49	64	81

With random (unmatched) samples of size n, the two terms on the right-hand side can be written in terms of the correlation ρ between y and x.

$$V(\bar{y}_1 - \bar{y}_2) = \frac{2}{n}\left[\rho^2\sigma_y^2 + (1 - \rho^2)\sigma_y^2\right]$$

If the fractional reduction in $V(\bar{x}_1 - \bar{x}_2)$ due to matching is f, this reduction affects the component $\rho^2\sigma_y^2$, but not the residual component $(1 - \rho^2)\sigma_y^2$. Hence the fractional reduction in $V(\bar{y}_1 - \bar{y}_2)$ due to matching is $f\rho^2$, and can be calculated for a given ρ from the values of f in preceding tables. Table 5.9.1 shows the percent reductions from within-class matching for the smaller numbers of classes.

With three or more classes, the percent reductions are determined primarily by the value of ρ rather than by the number of classes. Matching for increased precision when there is no danger of bias, does not begin to pay substantial dividends until ρ is 0.5 or greater.

When a decision about matching is to be made, either for protection against bias or for increased precision, an important question to consider is "What are the alternatives to matching? The principal alternatives—adjustments during the statistical analysis—are the subject of Chapter 6, which includes comparisons with matching where available.

5.10 INTRODUCTION TO STATISTICAL ANALYSIS OF PAIR-MATCHED SAMPLES

In this section we introduce methods of statistical analysis for pair-matched (caliper or "nearest available") and mean-matched samples. For within-class matching the methods of analysis are essentially the same as those for the adjustment of unmatched samples, and will be discussed in Chapter 6.

Pair Matching: y Continuous

The data form a two-way classification (treatments and pairs or matched groups). Under the additive model

$$y_{ij} = \mu + \tau_i + \gamma_j + e_{ij}$$

with $V(e_{ij}) = \sigma^2$, the usual two-way analysis of variance provides an estimate of the standard error of any comparison among the treatment means. With only two treatments, we may equivalently analyze the column of differences $d_j = y_{1j} - y_{2j}$, between the members of a pair, in order to estimate the standard error of $\bar{d} = \bar{y}_1 - \bar{y}_2$. Note that the analysis with two treatments does not require the assumption that $\sigma_1^2 = \sigma_2^2$. Similarly, if the variances $V(e_{ij}) = \sigma^2$ are thought to change from treatment to treatment, valid standard errors and t tests for any comparison $\Sigma\lambda_i\bar{y}_i$ are obtained by analyzing the column of values $\Sigma\lambda_i y_{ij}$.

The analysis of matched pairs is usually directed at estimation and testing of the mean difference $\bar{d} = \bar{y}_1 - \bar{y}_2$. However, with tight matching it is also possible to examine whether $d_j = y_{1j} - y_{2j}$ varies with the level of x. One approach is to let $x_j = (x_{1j} + x_{2j})/2$ and compute the linear regression of d_j on x_j, which constitutes 1 d.f. (degree of freedom) from the $(n - 1)$ d.f. for the variation of d_j from pair to pair. Higher polynomial regression terms may be added if appropriate, or multivariate regression of d_j may be used on different x variables than were used in matching.

Pair Matching: y (0, 1)

If y represents a two-way classification with two pair-matched treatments, a member of any pair can only have the y values 0 or 1. Thus the pairs have only the four y values (1, 1), (1, 0), (0, 1), and (0, 0), where the first number refers to treatment 1 (T_1) and the second number to treatment 2 (T_2). The data may be summarized as follow:

T_1	T_2	Number of Pairs
1	1	n_{11}
1	0	n_{10}
0	1	n_{01}
0	0	n_{00}
Total		n

The proportions of "ones" for the two treatments are $\hat{p}_1 = (n_{11} + n_{10})/n$ and $\hat{p}_2 = (n_{11} + n_{01})/n$. Thus $\bar{d} = (n_{10} - n_{01})/n$. As McNemar (1947) and other investigators have shown, the null hypothesis $p_1 = p_2$ is tested by regarding n_{10} and n_{01} as binomial successes and failures from $n_{10} + n_{01}$ trials, with probability of success $\frac{1}{2}$ on the null hypothesis. An exact test can be made from the binomial tables. For an approximate test, the value of χ^2 corrected for continuity with 1 d.f. is

$$\frac{(|n_{10} - n_{01}| - 1)^2}{n_{10} + n_{01}}$$

Stuart (1957) gives the estimated standard error of $\bar{d} = \hat{p}_1 - \hat{p}_2$ as

$$\left(\frac{\hat{p}_{10} + \hat{p}_{01} - (\hat{p}_{10} - \hat{p}_{01})^2}{n} \right)^{1/2}$$

Billewicz (1965) reports that in 9 out of 20 examples of matching in the field of medicine, the analysis used was (incorrectly) that appropriate to independent rather than matched samples. This mistake overestimates the standard error of $\bar{d}$ and underestimates χ^2. [See Cochran (1950) for an extension of the pair-sample methods when more than two treatments are used.]

With two treatments, let p_{1j} and p_{2j} be the true probabilities of success in the jth pair. In pairing, we presumably expect the probabilities of success to vary from pair to pair. In seeking a model that describes how p_{1j} and p_{2j} vary from pair to pair, many authors [writing q for $(1 - p)$] have used the following relations:

$$\frac{p_{1j}}{q_{1j}} = \psi\lambda_j; \qquad \frac{p_{2j}}{q_{2j}} = \lambda_j \tag{5.10.1}$$

where λ_j represents the level of the jth pair and ψ measures the disparity between the effects of the treatments. In model (5.10.1), the quantity that is regarded as constant from pair to pair is $\psi = p_{1j}q_{2j}/p_{2j}q_{1j}$, sometimes called the *odds ratio*, rather than $\delta = p_{1j} - p_{2j}$. The model assumes that the effects of the treatment and the pair are additive on the scale of $\log(p_{ij}/q_{ij})$, called the "logit of p_{ij}." An additive model on the scale of p_{ij} itself has the logical difficulty that p_{ij} must lie between 0 and 1.

With model (5.10.1) the quantity to be estimated is the odds ratio ψ. If the λ_j are regarded as nuisance parameters, D. R. Cox (1958) has shown that (1) an optimum estimate of ψ uses the $(1, 0)$ and $(0, 1)$ pairs only, and that (2) the ratio $n_{10}/(n_{10} + n_{01})$ is a binomial estimate of $\theta = \psi/(1 + \psi)$, based on a sample of size $n_{10} + n_{01}$. This result provides an estimate of θ and hence of $\psi = \theta/(1 - \theta)$. Similarly, confidence limits for ψ can be obtained from binomial confidence limits for θ as shown in Hald's (1952) tables or the Fisher and Yates tables (1953).

5.11 ANALYSIS WITH MEAN MATCHING; y CONTINUOUS

So far as I know, analysis with mean matching is seldom used. It has merit when the regression of y on x is linear with the same slope in both populations. The following analysis makes these assumptions. Let

$$y_{1u} = \mu + \tau_1 + \beta x_{1u} + e_{1u}; \qquad y_{2v} = \mu + \tau_2 + \beta x_{2v} + e_{2v}$$

Hence

$$\bar{y}_1 - \bar{y}_2 = \tau_1 - \tau_2 + \beta(\bar{x}_1 - \bar{x}_2) + \bar{e}_1 - \bar{e}_2$$

An effective mean matching makes $\bar{x}_1 - \bar{x}_2$ so close to zero that $\bar{y}_1 - \bar{y}_2$ serves as the estimate of $\tau_1 - \tau_2$, with variance

$$V(\bar{y}_1 - \bar{y}_2) = \frac{\sigma_{1e}^2 + \sigma_{2e}^2}{n}$$

In order to estimate σ_{1e}^2 and σ_{2e}^2 and hence $V(\bar{y}_1 - \bar{y}_2)$, the effects of the linear regression of y on x must be removed from variations in y. Let $(yy)_1$ denote $\Sigma(y_{1u} - \bar{y}_1)^2$, etc. Then

$$\hat{\sigma}_{1e}^2 = \frac{(yy)_1 - (yx)_1^2/(xx)_1}{n - 2}$$

with a similar expression for $\hat{\sigma}_{2e}^2$.

A large-sample $1 - \alpha$ confidence interval for $\tau_1 - \tau_2$, with effective mean matching, where y has a linear regression on x with the same slope in both populations is thus

$$\bar{y}_1 - \bar{y}_2 - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_{1e}^2 + \hat{\sigma}_{2e}^2}{n}} < \tau_1 - \tau_2 < \bar{y}_1 - \bar{y}_2 + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_{1e}^2 + \hat{\sigma}_{2e}^2}{n}}$$

5.12 SUMMARY

In an observational study systematic differences between the populations from which different treatment groups are drawn can have two effects on comparisons $\bar{y}_1 - \bar{y}_2$ between the response means for samples exposed to different treatments. These differences may create a bias in these comparisons and may decrease precision of comparisons. In attempting to avoid these consequences the first step is to list the principal variables x that influence y and are not themselves affected by the treatments. Such variables are called *confounding variables* (sometimes called "covariates" or "control variables"). They may be classifications, or discrete or continuous variables.

Matching is a common method of handling such confounding x variables at the planning stage by making the samples for different treatments resemble each other in certain respects. The matching method used depends on the nature of the x distributions.

If the x's are classifications, the cells that are created by this multiple classification are formed. In *within-class* or *frequency matching*, each member of any sample has a partner in any other sample belonging to the same cell, so that in the ith cell all samples for different treatments contain the same number of members n_i. This method is often used with discrete or classified x variables (e.g., number of children, age) by first grouping the values of the variable into classes.

With continuous or discrete x variables *caliper matching* requires the x values for partners in different samples to agree within prescribed limits $\pm a$. *Mean Matching* concentrates on making the means $\bar{x}_{ti}$ for different treatments agree as closely as possible for the ith x variable.

The idea of matching is simple to understand. Its objective is to free the comparisons of the means $\bar{y}_t$ from the effects of differences among the x distributions in different treatment groups. Perfect matching removes from these comparisons the effects of any shape of relationship between y and the x's, provided that this shape is the same in all populations being compared. Hence the statistical analysis of matched samples is relative simple.

The primary disadvantage is the time and effort required to construct matched samples. The degree of difficulty depends on the desired sample size, the sizes of the reservoirs available for seeking matches, the number of x's to be matched, the tightness of the matching rules (caliper matching, in general is more difficult than within-class matching) and the sizes of the differences between the x distributions in different populations. A case-by-case search for paired matches on say four x variables can be tedious and may require several relaxations of the original matching rules in order to find matches. One consequence of matching whose effects are more difficult to assess is that the sample-population relationship is disturbed. In matched sampling every sample may be a nonrandom sample from its own population.

In finding matches, computers should be able to take over much of the work if the x's in the reservoirs are in a form suitable for data entry. For instance, using within-class matching on three x variables having c_1, c_2, and c_3 classes, the computer can arrange and print the data in the $c_1 \times c_2 \times c_3$ cells from which matches are easily selected. A similar method is a considerable help in seeking caliper matches.

Sometimes an investigator needs all n cases available from one population and has a limited reservoir from the second population. With x continuous it is not clear whether caliper matches with a prescribed $\pm a$ can be found for all n. For this problem D. Rubin has developed a method of computer matching, called *nearest available* matching, that guarantees that every case is matched.

In comparing two samples, a primary objective of matching is, of course, to remove bias in $\bar{y}_1 - \bar{y}_2$ due to differences in the x distributions. If y has the same linear regression on x in both populations, the percentage reduction in the bias of $\bar{y}_1 - \bar{y}_2$ due to matching equals the percentage reduction in the bias of $\bar{x}_1 - \bar{x}_2$. Consequently, the effect of matching on the bias of $\bar{x}_1 - \bar{x}_2$ is examined first.

In within-class matching, some types of classified x's are such that two members of the same class are identical with regard to x. In this event, within-class matching completely removes any initial bias in $\bar{x}_1 - \bar{x}_2$. But many classified x's (e.g., social level, degree of aggressiveness) more nearly represent a grouping of an underlying continuous x, as is the case when a continuous x is deliberately grouped in order to use within-class matching. In this situation, matching with two, three, four and five classes removes approximately 64%, 80%, 87%, and 91% of an initial bias in $\bar{x}_1 - \bar{x}_2$.

With x continuous, the effect of caliper matching, to within $\pm a$ units, depends primarily on the ratio a/σ_x and to some extent on the way in which the caliper matches are constructed. "Loose" caliper matching to within $\pm 0.9\sigma_x$ removes slightly more than 80% of an initial bias (as effective as within-class matching with three classes). Caliper matching to within $\pm 0.4\sigma_x$ removes 95–97% of the initial bias.

Mean matching is highly successful in removing bias in $\bar{x}_1 - \bar{x}_2$ even if all n members of sample 1 must be used and the sample 2 reservoir is only of size $2n$.

The effect of "nearest available" matching depends on the size of the initial bias $\bar{x}_1 - \bar{x}_2$ and on the size of the reservoir in population 2. With a reservoir of size $4n$ this method removes nearly all the initial bias, unless this bias is exceptionally large. Even a reservoir of size $2n$ should remove around 90% of a moderate-sized initial bias.

Suppose that the regression of y on x is the same in two populations but is nonlinear, being monotone and moderately curved as represented by a quadratic regression or by $y = e^{\pm x/2}$. Some evidence indicates that for within-class and caliper matching, the percent reduction in the bias of

$\bar{y}_1 - \bar{y}_2$ is only slightly less than that in the linear case, provided x has the same variance in the two populations. When σ_{1x}^2 and σ_{2x}^2 are unequal this result does not hold; the percentage reduction in bias is sometimes much greater and sometimes much less than when $\sigma_{1x}^2 = \sigma_{2x}^2$. Mean matching should be avoided when the regression of y on x is curved.

Matching methods are only partially successful to varying degrees in removing an initial bias in $\bar{y}_1 - \bar{y}_2$ due to confounding x variables. Under a linear regression of y on x, removal of over 90% of an initial bias requires five classes in within-class matching, or caliper matching to within $\pm 0.4\sigma_x$. Matching is not suitable when the regression of y on x is of a different form in the two populations.

Matching is also used to increase the precision of the comparison $\bar{y}_1 - \bar{y}_2$ when the x distribution is thought to be the same in the two populations, that is, where there is no danger of bias. For within-class and caliper matching, the percent reduction f in $V(\bar{x}_1 - \bar{x}_2)$ is similar to that in the bias of $\bar{x}_1 - \bar{x}_2$. Under a linear regression the percent reduction in $V(\bar{y}_1 - \bar{y}_2)$ is $f\rho^2$, where ρ is the correlation between y and x. Thus the reduction in $V(\bar{y}_1 - \bar{y}_2)$ does not become substantial until $|\rho|$ exceeds 0.5.

REFERENCES

Billewicz, W. Z. (1965). The efficiency of matched samples: an empirical investigation. *Biometrics*, **21**, 623–644.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, **37**, 256–266 [Collected Works #43].

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

Cox, D. R. (1957). Note on grouping. *J. Am. Statist. Assoc.*, **52**, 543–547.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, **45**, 562–565.

Douglas, J. W. B. (1960). 'Premature' children at primary schools. *Br. Med. J.*, **1**, 1008–1013.

Feller, W. (1957). *An Introduction to Probability Theory and its Applications* (2nd ed.). Wiley, New York.

Fisher, R. A. and F. Yates (1953). *Statistical Tables for Biological, Agriculture and Medical Research* (4th ed.). Oliver and Boyd, Edinburgh, Scotland.

Hald. A. (1952). *Statistical Tables and Formulas*. Table XI. Wiley, New York.

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

Ogawa, J. (1951). Contributions to the theory of systematic statistics. I. *Osaka Math. J.*, **3**, 175–213.

Rubin, D. B. (1970). The Use of Matched Sampling and Regression Adjustment in Observational Studies. Ph.D. Thesis, Harvard University, Cambridge, Mass.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, **29**, 159–183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.

Stanley, J. C. (1966). A common class of pseudo-experiments. *Am. Educational Res. J.*, **3**, 79–87.

Stuart, A. (1957). The comparison of frequencies in matched samples. *Br. J. Statist. Psychol.*, **10**, 29–32.