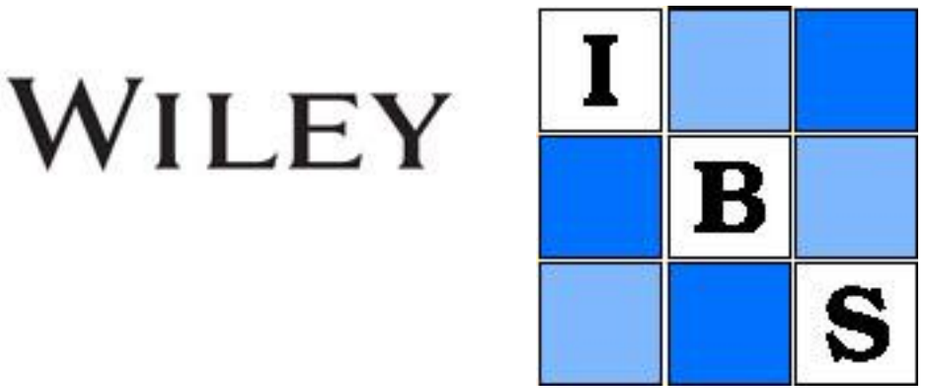

THE EFFECTIVENESS OF ADJUSTMENT BY SUBCLASSIFICATION IN REMOVING BIAS IN OBSERVATIONAL STUDIES

W. G. COCHRAN

Harvard University, Cambridge, Mass., U. S. A.

SUMMARY

In some investigations, comparison of the means of a variate y in two study groups may be biased because y is related to a variable x whose distribution differs in the two groups. A frequently used device for trying to remove this bias is adjustment by subclassification. The range of x is divided into c subclasses. Weighted means of the subclass means of y are compared, using the same weights for each study group. The effectiveness of this procedure in removing bias depends on several factors, but for monotonic relations between y and x, an analytical approach suggests that for $c = 2, 3, 4, 5$, and 6 the percentages of bias removed are roughly 64%, 79%, 86%, 90%, and 92%, respectively. These figures should also serve as a guide when x is an ordered classification (e.g. none, slight, moderate, severe) that can be regarded as a grouping of an underlying continuous variable. The extent to which adjustment reduces the sampling error of the estimated difference between the y means is also examined. An interesting side result is that for x normal, the percentage reduction in the bias of $\bar{x}_2 - \bar{x}_1$ due to adjustment equals the percentage reduction in its variance.

Under a simple mathematical model, errors of measurement in x reduce the amount of bias removed to a fraction $1/(1 + h)$ of its value, where h is the ratio of the variance of the errors of measurement to the variance of the correct measurements. Since ordered classifications are often used because x is difficult to measure, h may be substantial in such cases, though more information is needed on the values of h that are typical in practice.

1. INTRODUCTION

Examples of the type of observational study considered in this paper are comparisons of the death rates of men with different smoking habits and Kinsey's comparisons of the frequencies of a specific type of sexual behavior among men of different socioeconomic levels. The investigator studies a response variable y in two or more groups of people who differ with respect to some characteristic (smoking, air pollution, socioeconomic level). He realizes, however, that part or all of the observed differences between the mean values of y in the groups may be due to other variables x_1 , x_2 , $\cdots$ in which the groups differ, rather than to the specific characteristic that he is interested in studying.

In planning such studies, it is therefore good practice to note any x variable that is known or suspected to have an important relationship with y. Whenever feasible, steps are taken to measure x for the people in each group. The investigator can then examine whether the frequency distributions of x differ in the different groups. If so, he tries to adjust the mean values of y so as to remove any biases that may have arisen because of these differences. Incidentally, x must not be a variable that is itself causally affected by the characteristic (smoking, socioeconomic level) that is under study. If it is, the adjustments remove part of the difference in response that the investigator wants to measure.

In this paper we study the effectiveness of a frequently used method, which will be called adjustment by subclassification. The distribution of x is broken up into 2, 3, or more subclasses. For each group of subjects, the mean value of y is calculated separately within each subclass. Then a weighted mean of these subclass means is calculated for each group, using the same weights for every group. The actual weights employed depend on the judgment of the investigator. Sometimes they are proportional to the numbers of subjects in the subclasses of one group that is regarded as the standard of comparison, sometimes to the combined numbers of subjects in the subclasses over all groups. They may be derived from an extraneous standard population or from least squares theory.

The rationale of this method of adjustment is that when there are numerous subclasses, each with a restricted range of x, the within-class distributions of x cannot differ much from one group to another. Comparisons among the subclass means of y for the different groups should therefore be almost free from bias due to x. Since the subclass weights are the same for every group, the same remark should apply to the overall adjusted means of y.

We are particularly interested in the case in which the number of subclasses is small. Adjustments using 2–5 subclasses per variable are common in practice, especially when there are several x variables for which adjustment is advisable. Further, when x is difficult to measure accurately, the x variable often takes the form of an ordered classification with a limited number of classes, e.g. none, mild, moderate, severe; or for, neutral, against. It is sometimes reasonable to regard an ordered classification as representing a grouping of an underlying continuous x variable. Our results should therefore give an indication of the effectiveness of subclassification when x can be measured only as an ordered classification.

This type of adjustment by subclassification also resembles another

procedure known as *frequency matching* or *stratified matching*. In frequency matching with respect to a single variable x, the range of x is divided into a number of subclasses. The groups of subjects to be compared are selected in such a way that every group has the same size of sample, say n_i , in subclass i. It follows that the unweighted means $\bar{y}$ of the groups are all weighted means of the subclass means of y, with the same weights n_i/n for every group. Consequently, results about the effectiveness of adjustment by subclassification will apply also to a frequency matching that uses the same set of subclasses.

2. AN ILLUSTRATION

The following data were selected from data supplied to the U. S. Surgeon General's Committee from three of the studies in which comparisons of the death rates of men with different smoking habits were made. In each study there are three groups of men: non-smokers, smokers of cigarettes only, and smokers of cigars or pipes or both. The studies are the Canadian study of Best and Walker, with six years experience, the British Doctors' study by Doll and Hill, with five years experience, and the second American Cancer Society study by Hammond, with about 20 months experience. These data were chosen solely for illustration in this paper. They are not the latest reports from these studies, nor do these investigators have any connection with remarks made here.

Table 1 shows for each group the *unadjusted* death rates per 1,000 person-years of experience. These rates were obtained by dividing the total deaths that had occurred in each smoking group by the total number of person-years of exposure ($\times 10^3$) of the men in that group.

These studies in three different countries agree well in the relative death rates of the three groups of smokers. To the naive, the conclusion seems clear. We should urge the cigar and pipe smokers to give up smoking. If they lack the strength of will to do so, they should switch

TABLE 1

DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Study		
	Canadian	British	U. S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	20.5	14.1	13.5
Cigars, pipes	35.5	20.7	17.4

to cigarettes only. (In the British study the increase in death rates from 11.3 for non-smokers to 14.1 for cigarette smokers is statistically significant by a crude test, but the Canadian and U. S. studies show no corresponding increase.)

Before lending any credence to the comparisons in Table 1, we should ask: are there other variables in which the three groups of smokers may differ, that (i) are related to the probability of dying and (ii) are clearly *not* themselves affected by smoking habits. For men under 40, I suppose the answer is that there are no known variables with a consistent and strong relationship. For men over 40 a variable of this type that becomes of overwhelming importance is of course age. The regression of probability of dying on age for men over 40 is a concave upwards curve, the slope rising more and more steeply as age advances. The mean ages for each group in Table 1 are as follows.

TABLE 2

MEAN AGES, YEARS

Smoking group	Study		
	Canadian	British	U. S.
Non-smokers	54.9	49.1	57.0
Cigarettes only	50.5	49.8	53.2
Cigars and/or pipe	65.9	55.7	59.7

The high mean ages for the cigar-pipe smokers are notable and not unexpected. Adjustment to remove the effects of these age differences is essential. Incidentally, even if the mean ages for the three smoking groups had been identical within a study, this would not automatically eliminate age as a possible source of bias in death rates unless the regression of death rate on age were linear with the same slope in each group.

Table 3 shows the adjusted death rates obtained when the age distributions were divided into 2 subclasses, 3 subclasses, and the maximum number of subclasses that the available data allow, these being 12, 9, and 11, respectively. The subclass boundaries were chosen to give subclasses of approximately equal size (this is not necessarily the optimum choice of boundaries). The weights used were those for the non-smokers, so that the non-smoker death rates remain the same for the different numbers of classes.

With the maximum numbers of subclasses, the adjusted cigarette

TABLE 3

ADJUSTED DEATH RATES USING 2, 3, AND 9–11 SUBCLASSES

Number of subclasses	Canadian			British			U. S.		
	N. S.*	C.+	CP′	N. S.	C	CP	N. S.	C	CP
1	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
2	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
3	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
9–11	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

*Non-smokers, +Cigarettes only, ′Cigars, Pipes

death rates now show substantial increases over the non-smoker death rates in all three studies. The adjusted cigar-pipe death rates, on the other hand, exhibit no elevation over those for non-smokers.

Looking down the columns we see that with 2 subclasses substantially over half the effect of the age bias has been removed in most cases, while with 3 subclasses, a little more is removed. The only exception occurs for cigarette smokers in the British study. In this case the mean ages of non-smokers and cigarette smokers agree closely, 49.1 against 49.8 years, so that the adjustments are removing the effects of differences in the shapes of the age distributions rather than in the mean ages.

3. TOPICS TO BE DISCUSSED

These topics will be described under three headings, as follows.

1. As already mentioned, results are presented on the effectiveness of adjustment by subclassification in controlling bias when y and x are continuous and the number of subclasses ranges from 2 to 6. In order to introduce some notation, suppose that we are comparing two populations, denoted by the subscripts 1 and 2. Independent random samples have been drawn from each population. Let $u(x)$ represent the population regression of y on x. If y_{1i} , y_{2k} are random members of the two populations, the model is

$$y_{1i} = \alpha_1 + u(x_{1i}) + e_{1i} , \qquad y_{2k} = \alpha_2 + u(x_{2k}) + e_{2k} ,$$

where e_{1i} , e_{2k} are random residuals with zero means in the respective populations. The quantity to be estimated is $(\alpha_2 - \alpha_1)$. For the unadjusted means of y in the two groups, it follows that

$$E(\bar{y}_1) = \alpha_1 + \bar{u}_1 , \qquad E(\bar{y}_2) = \alpha_2 + \bar{u}_2 ,$$

where

$$\bar{u}_1 = \int u(x)\phi_1(x)\,dx, \qquad \bar{u}_2 = \int u(x)\phi_2(x)\,dx, \tag{3.1}$$

and $\phi_1(x)$, $\phi_2(x)$ are the frequency functions of x in the two populations. Hence if no adjustment is made, the initial bias due to x is $\bar{u}_2 - \bar{u}_1$.

As implied above, the regression function $u(x)$ is assumed to be the same in the two populations. In observational studies this assumption does not necessarily hold. When the u's differ in the two populations the meaning of the adjustment and the best method of making it require further study.

In the ith subclass, let the boundaries of x be x_{i-1} and x_i and let the sample means of y be $\bar{y}_{1i}$ and $\bar{y}_{2i}$. We have

$$E(\bar{y}_{1i}) = \alpha_1 + \bar{u}_{1i} ,$$

where

$$\bar{u}_{1i} = \int_{x_{i-1}}^{x_i} u(x)\phi_1(x)\,dx \Big/ \int_{x_{i-1}}^{x_i} \phi_1(x)\,dx. \tag{3.2}$$

After adjustment, the remaining bias due to x is

$$\sum w_i(\bar{u}_{2i} - \bar{u}_{1i}), \tag{3.3}$$

where w_i is the weight assigned to subclass i. The proportion of the initial bias that is removed by the adjustment is therefore

$$1 - \sum w_i(\bar{u}_{2i} - \bar{u}_{1i})/(\bar{u}_2 - \bar{u}_1). \tag{3.4}$$

From the nature of expressions (3.1), (3.2), and (3.3), it appears that this proportion will depend on the following quantities: the mathematical form of the regression function $u(x)$, the shapes of the frequency functions $\phi_1(x)$ and $\phi_2(x)$, the number of subclasses, the division points x_i , and the choice of weights w_i .

In view of the numerous variables or functions involved, a thorough investigation may require extensive experimental sampling. Instead, we have attempted to use an analytical approach by restricting the scope of the problem in two ways.

(i) If $u(x)$ is monotone and differentiable, as is presumably the case, for example, in a function designed to represent the regression on age of probability of dying at ages over 40, we can replace x by u in a theoretical investigation, so that the regression becomes linear. The functions $\phi_1(x)$ and $\phi_2(x)$ are replaced by the corresponding frequency functions $f_1(u)$ and $f_2(u)$ into which this transformation converts them. Reverting to x, we therefore assume that the regression of y on x is linear and that x has frequency functions $f_1(x)$ and $f_2(x)$ in the

two populations. An important consequence of a linear regression of y on x is that the percentage reduction in the bias of $\bar{y}_2 - \bar{y}_1$ equals that in $\bar{x}_2 - \bar{x}_1$.

(ii) If $f_1(x)$ and $f_2(x)$ differ only in the value of a single parameter θ that enters into the specification of each, then, as will be shown, the proportion of the initial bias in x or y that is removed by subclassification can be expressed by a calculus formula when θ is small. In particular, if $f_1(x) = f(x)$, $f_2(x) = f(x - \theta)$, so that the distributions differ only by a translation, this proportion becomes

$$\sum_{i=1}^{c} M_i(f_{i-1} - f_i).$$

In this expression, f_{i-1} and f_i are the ordinates of $f(x)$ at the boundaries x_{i-1} and x_i of the ith subclass, and M_i is the mean value of x in the ith subclass. Numerical values of these proportions will be presented for c = number of subclasses = 2(1)6 and for $f(x)$ having the normal, χ^2, and t distributions, as well as for some beta distributions and a distribution simulating the age distribution of males.

2. Some results are given to show the effect of subclassification on the *precision* of the comparison of the two population means, as measured by the ratio of the variance of the adjusted difference $\bar{y}_{2w} - \bar{y}_{1w}$ to that of the initial difference $\bar{y}_2 - \bar{y}_1$. Adjustment by subclassification is often used in situations in which the investigator does not suspect any danger of bias, believing that $f_1(x)$ and $f_2(x)$ are closely similar. Instead, he regards the variance of y as due in part to its regression on x. By controlling on x through the adjustment process, he hopes to reduce the variance of $\bar{y}_{2w} - \bar{y}_{1w}$. The objective is the same as in a standard analysis of covariance.

3. The effects of errors of measurement of x on the performance of the adjustments will be examined under a simple mathematical model. This topic is particularly relevant when x is an ordered classification, since these classifications are often employed because x is difficult to measure accurately, so that errors of measurement are anticipated.

4. PERCENT REDUCTION IN BIAS WITH A LINEAR REGRESSION

Before considering a calculus approach, some preliminary calculations of the percentage reductions in the bias of $\bar{x}_2 - \bar{x}_1$ were made on a computer, using formula (3.4), for the case in which $f_1(x)$ is the normal distribution $N(0, \sigma^2)$, while $f_2(x)$ is $N(\theta, \sigma^2)$. Equal-sized subclasses in population 1 and equal weights were chosen. For θ/σ = 1, $\frac{1}{2}$, $\frac{1}{4}$ and for 2, 3, 4, 5, and 6 subclasses, the percent reductions in the bias of $\bar{x}_2 - \bar{x}_1$ appear in Table 4.

TABLE 4

PERCENT REDUCTIONS IN BIAS: LINEAR REGRESSION, x NORMAL

θ/σ	Number of subclasses 2	3	4	5	6
1	61.8	78.2	85.3	89.1	91.5
$\frac{1}{2}$	63.2	79.1	85.9	89.6	91.8
$\frac{1}{4}$	63.6	79.3	86.0	89.7	91.9

As would be expected, the percentage of the initial bias that is removed by the adjustment increases steadily as the number of subclasses is increased. Table 4 also indicates that for initial biases which are not too large ($\theta/\sigma \leq \frac{1}{2}$), the percent bias removed may be almost independent of the value of θ/σ. This observation suggests that results applicable to any continuous $f(x)$ might be obtainable by a calculus approach in which θ/σ is assumed small. It may be noted that a bias for which $\theta/\sigma = \frac{1}{2}$ is by no means negligible: with two samples of size n, the ratio of $\bar{y}_2 - \bar{y}_1$ to its standard error has a bias $\frac{1}{2}\sqrt{n/2}$. With $n = 32$, for example, the probability of finding a significant difference in $\bar{y}_2 - \bar{y}_1$ that is entirely due to this bias is about 0.5 in a two-tailed test at the 5% level.

Let $f(x)$ depend on a parameter θ that has the value 0 in population 1 and the value θ in population 2. For the adjustments, the range of x is divided into c subclasses by division points $x_0 , x_1 , \cdots , x_c$. In the ith subclass let $P_i(\theta)$ denote the proportion of the population and $M_i(\theta)$ the mean value of x. The weights used may be the $P_i(0)$, the $P_i(\theta)$ or a combination of the two. Since θ tends to zero in this approach, these different choices of weights become identical. We assume that the $P_i(0)$ are used.

If $M(\theta)$ denotes the overall mean of x, the initial bias, $M(\theta) - M(0)$ may be written

$$\sum_{i=1}^{c} [P_i(\theta)M_i(\theta) - P_i(0)M_i(0)] \doteq \theta \sum_{i=1}^{c} \left[P_i \frac{dM_i}{d\theta} + M_i \frac{dP_i}{d\theta} \right] = \theta \frac{dM}{d\theta} \tag{4.1}$$

assuming θ small, where the derivatives are taken at $\theta = 0$. After adjustment, the bias remaining is

$$\sum_{i=1}^{c} P_i(0)[M_i(\theta) - M_i(0)] \doteq \theta \sum_{i=1}^{c} P_i \frac{dM_i}{d\theta}. \tag{4.2}$$

Consequently, the proportion of the initial bias that is removed is approximately

$$\sum_{i=1}^{c} M_i \frac{dP_i}{d\theta} \Big/ \frac{dM}{d\theta}. \tag{4.3}$$

The utility of this expression depends, of course, on whether the functions that enter into (4.3) are easily found analytically. If $f_1(x) = f(x)$, $f_2(x) = f(x - \theta)$, the denominator of (4.3) becomes 1, since the initial bias in (4.1) is θ. Further,

$$P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x - \theta)\, dx = \int_{x_{i-1}-\theta}^{x_i-\theta} f(x)\, dx$$

so that at $\theta = 0$,

$$dP_i/d\theta = f(x_{i-1}) - f(x_i). \tag{4.4}$$

Consequently, the proportional reduction in bias from (4.3) becomes,

$$\sum_{i=1}^{c} M_i(f_{i-1} - f_i), \tag{4.5}$$

where $f_i = f(x_i)$.

For the unit normal distribution,

$$M_i P_i = \frac{1}{\sqrt{2\pi}} \int_{x_{i-1}}^{x_i} x e^{-\frac{1}{2}x^2}\, dx = f_{i-1} - f_i \tag{4.6}$$

so that the proportional reduction becomes

$$\sum_{i=1}^{c} (f_{i-1} - f_i)^2 / P_i\ . \tag{4.7}$$

Expression (4.7) has turned up previously in several papers. In particular, D. R. Cox [1957] obtained it as 1 minus the ratio of the average within-subclass variance of x to the original variance of x when x is normal. The connection between his results and (4.7) will be discussed in section 6. Table 5, taken from his computations, shows the optimum P_i that maximize the percentage reduction in bias, the corresponding reductions, and the reductions obtained when the P_i are all equal.

With the optimum sets of boundaries the central subclass is the largest, the subclasses becoming steadily smaller towards both ends of the range of x. Table 5 shows also that the choice of boundaries is not crucial: use of subclasses of equal size diminishes the percent reductions by only about 2% as compared with the optimum boundaries. The percent reductions with equal P_i agree closely with those given in Table 4 for $\theta/\sigma = \frac{1}{4}$ and $\frac{1}{2}$, as had been anticipated.

TABLE 5

OPTIMUM CHOICE OF P_i AND PERCENT REDUCTIONS IN BIAS GIVEN BY OPTIMUM AND EQUAL P_i WHEN x IS NORMAL

Number of subclasses, c	Optimum P_i (%)	Maximum reduction	Reduction with equal P_i
2	50%; 50%	63.7%	63.7%
3	27%; 46%; 27%	81.0%	79.3%
4	16%; 34%; 34%; 16%	88.2%	86.1%
5	11%; 24%; 30%; 24%; 11%	92.0%	89.7%
6	7%; 18%; 25%; 25%; 18%; 7%	94.2%	91.9%

5. PERCENT REDUCTION IN BIAS FOR SOME NON-NORMAL DISTRIBUTIONS

Three non-normal distributions were also investigated—χ^2 and t with a range of numbers of degrees of freedom, and some beta distributions. With χ^2 and t it was assumed that population 2 differs from population 1 only in a translation θ, formula (4.5) being used to compute the proportional reduction in bias when θ is small.

The beta distributions were $B(m, 3)$, for which the frequency function is proportional to $x^{m-1}(1 - x)^2$, where $0 \leq x \leq 1$. The mean of this distribution is $m/(m + 3)$. The sets of pairs $m = (2, 3); (3, 4.5); (4.5, 7)$; and $(7, 12)$ in the two populations give means $(0.4, 0.5); (0.5, 0.6); (0.6, 0.7)$; and $(0.7, 0.8)$, respectively. The beta distribution provides an interesting variant. When m changes, the shape of the distribution changes as well as the mean; further, the distribution has a finite range. The proportional reductions in bias for this distribution were computed directly without using the calculus approximation (4.3).

Two examples simulating situations with non-linear regressions were also worked. The variate x was age, the distribution in population 1 being that of the 1963 U. S. population aged 45–75 by single years. The frequency function is not unlike a trapezoid with a negative slope. The variate y was the death rate, approximately a cubic function of x. Two distributions were tried for population 2—a rectangular distribution and the reversed age distribution. Before adjustment for age the mean death rate in population 2 was 22% higher than that in population 1 for the rectangular and 42% higher for the reversed distribution. Table 6 shows the results.

In all these cases the P_i (subclass sizes) were made equal. With χ^2 the percentage reductions differ only trivially from those for the

TABLE 6

PERCENT REDUCTIONS IN BIAS FOR NON-NORMAL DISTRIBUTIONS WITH EQUAL-SIZED SUBCLASSES

Distribution of x		c = number of subclasses				
Pop. 1	Pop. 2	2	3	4	5	6
χ^2_4	$\chi^2_4 + \theta$	65.8	79.8	85.7	89.4	91.5
χ^2_6	$\chi^2_6 + \theta$	64.9	79.6	85.9	89.4	91.6
χ^2_{10}	$\chi^2_{10} + \theta$	64.9	79.4	85.9	89.5	91.7
χ^2_{20}	$\chi^2_{20} + \theta$	64.0	79.3	85.9	89.7	91.8
t_3	$t_3 + \theta$	81.0	97.6	102.7	104.9	105.7
t_5	$t_5 + \theta$	72.1	88.1	93.8	96.8	98.3
t_{10}	$t_{10} + \theta$	67.3	88.3	89.4	92.9	94.9
t_{20}	$t_{20} + \theta$	65.4	81.4	87.8	91.5	93.7
$B(2, 3)$	$B(3, 3)$	66.4	82.2	88.7	92.0	94.0
$B(3, 3)$	$B(4.5, 3)$	64.5	80.5	87.3	90.9	93.1
$B(4.5, 3)$	$B(7, 3)$	62.5	78.9	85.9	89.7	92.0
$B(7, 3)$	$B(12, 3)$	60.4	77.1	84.4	88.4	90.9
Age	Rect.	64.1	84.5	89.4	93.0	95.0
Age	Reversed	67.0	86.2	92.3	94.0	96.5
$N(0, 1)$	$N(\theta, 1)$	63.7	79.3	86.1	89.7	91.9

With χ^2 and t, the subscripts are the numbers of D.F.

normal distribution even when χ^2 has only 4 D.F. and is strongly skew. With t, adjustment by subclassification does substantially better than with the normal distribution when the degrees of freedom are small. The reductions of over 100% that appear when t has 3 D.F. may look erroneous—one cannot remove more bias than there is. What happens is that owing to the nature of the tails of t with small D.F., population 2 has a *lower* mean than population 1 in the highest subclass, the net result being to produce small overall negative biases in place of the initial positive bias.

From occasional calculations with actual populations having a finite range, my impression was that the adjustments perform better than the normal model suggests. The percentages shown in Table 6 for the beta distribution, however, all lie within about $\pm 2\%$ of the corresponding normal values and the averages over the four beta sets are close to the normal values although a little higher for $c \geq 3$. For the two age distribution examples, where the regression of y on x is markedly non-linear, the percentage reductions are consistently higher

than with the normal. These increases may be due in part to two factors: equal-sized classes are near the optimum for the trapezoidal age distribution, and the maximum number of classes that the data allow is 30, so that we are comparing 6 subclasses with 30 instead of 6 with an infinite number.

So far as they go, these calculations suggest that the normal values may serve as a working guide to the percentages of bias removed in practice with from 2 to 6 subclasses. There is a hint that the normal values may underestimate the effectiveness of adjustment for some types of distribution of x.

6. EFFECT OF ADJUSTMENT ON THE PRECISION OF THE COMPARISONS

An extensive study of the relative effectiveness of matching and of adjustment by covariance in increasing the precision of the comparison of the mean values of y in the two groups has been made by Billewicz [1965]. He used a series of experimental samplings designed to simulate the range of models and conditions under which this problem arises in practice. The present analytical approach is a minor supplement to his work.

We assume that x has the same distribution $f(x)$ in the two groups and that the regression of y on x is linear. The formulas are simplest in the case of stratified matching, which is considered first. If independent samples of size n are drawn in the two groups, with no matching or adjustment, the variance of $\bar{y}_2 - \bar{y}_1$ is $V(\bar{y}_2 - \bar{y}_1) = 2\sigma_y^2/n$. Of this, a part $2\beta^2\sigma_x^2/n = 2\rho^2\sigma_y^2/n$ is due to variations in x and a part $2(1-\rho^2)\sigma_y^2/n$ to other sources of variability, where β is the regression coefficient of y on x and ρ the correlation between y and x.

With stratified matching on x, the range of x is divided into c subclasses. In group 1 let P_i be the proportion of the population and n_i the sample number of observations that fall into the ith subclass. Group 2 is constrained to have n_i values of x in this subclass also. Consequently, if σ_i^2 is the variance of x within the ith subclass, the variance of $\bar{x}_{2i} - \bar{x}_{1i}$ is $2\sigma_i^2/n_i$. The variance of the overall mean difference $\sum n_i(\bar{x}_{2i} - \bar{x}_{1i})/n$ is $2\sum n_i\sigma_i^2/n^2$. Since n_i is a binomial estimate of nP_i , the average value of this variance is $2\sum P_i\sigma_i^2/n$.

The effect of stratified matching is therefore that the contribution of variations in x to $V(\bar{y}_2 - \bar{y}_1)$ is reduced from $2\beta^2\sigma_x^2/n$ to $2\beta^2\sum P_i\sigma_i^2/n$. The contribution $2(1 - \rho^2)\sigma_y^2/n$ of other sources of variation is unaffected by the matching. If $g = \sum P_i\sigma_i^2/\sigma_x^2$, the net result is to reduce $V(\bar{y}_2 - \bar{y}_1)$ from $2\sigma_y^2/n$ to

$$2\sigma_y^2[g\rho^2 + (1 - \rho^2)]/n = 2\sigma_y^2[1 - (1 - g)\rho^2]/n. \qquad (6.1)$$

A similar formula holds for adjustment by subclassification, except that in moderate-sized samples the factor g is larger than $\sum P_i\sigma_i^2/\sigma_x^2$ owing to dissimilarities in the sample distributions of x in the two groups. If the samples are of sizes n_1 , n_2 , the adjusted mean difference $\sum w_i(\bar{x}_{2i} - \bar{x}_{1i})$ has variance

$$\sum w_i^2\sigma_i^2[(1/n_{2i}) + (1/n_{1i})].$$

For given n_1 , n_2 , the average value of this variance depends, of course, on the choice of the weights w_i . For specified w_i this average can usually be expanded in a series of powers of $1/n_1$ and $1/n_2$, since n_{1i} and n_{2i} are binomial estimates of n_1P_i and n_2P_i , respectively. For instance, if $w_i = n_{1i}/n$, the two leading terms in the average variance can be shown to be

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sum P_i\sigma_i^2 + \frac{1}{n_2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sum Q_i\sigma_i^2 , \tag{6.2}$$

where $Q_i = 1 - P_i$. If the P_i are known and used as weights, the two leading terms are

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sum P_i\sigma_i^2 + \left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right) \sum Q_i\sigma_i^2 , \tag{6.3}$$

where $Q_i = 1 - P_i$. Since the variance of the unadjusted difference $\bar{x}_2 - \bar{x}_1$ is $\sigma_x^2(1/n_1 + 1/n_2)$, the first term in the above expressions leads to the factor g. The second term becomes negligible relative to the first as n_1 and n_2 increase, but can have an appreciable effect in moderate-sized samples. For instance, with $P_i = 1/c$, $Q_i = (c - 1)/c$, the second term amounts to increasing the variance by a factor $1 + (c - 1)/n_2$ in (6.2) and a factor $1 + (c - 1)(n_1^2 + n_2^2)/n_1n_2(n_1 + n_2)$ in (6.3).

We now consider $g = \sum P_i\sigma_i^2/\sigma^2$. An alternative expression can be given for g. By the well-known result,

$$\sigma^2 = \sum P_i\sigma_i^2 + \sum P_i M_i^2 - \mu^2$$

we have

$$1 - g = \sum P_iM_i^2/\sigma^2 - \mu^2/\sigma^2. \tag{6.4}$$

Two incidental results are of interest. For the normal distribution, $N(0, 1)$, $M_iP_i = f_{i-1} - f_i$, by (4.6). Hence, for this distribution, putting $\mu = 0$, $\sigma = 1$ in (6.4),

$$1 - g = \sum M_i(f_{i-1} - f_i). \tag{6.5}$$

The quantity $1 - g$ is the proportional reduction in the variance of $\bar{x}_2 - \bar{x}_1$ due to adjustment by subclassification. But by (4.5), the same expression holds for the proportional reduction in the bias of $\bar{x}_2 - \bar{x}_1$.

This equivalence of the proportional reductions in variance and bias appears to hold only for the normal distribution.

Secondly, the expression $\sum P_i\sigma_i^2/n$ is the variance of the mean of a stratified sample of size n with proportional stratification (ignoring the correction for finite population), where P_i and σ_i^2 refer to the ith stratum. Dalenius [1957] has shown that the set of class boundaries x_i which minimize this expression satisfy the simple relations

$$x_i = \tfrac{1}{2}(M_i + M_{i+1}). \tag{6.6}$$

Although these equations still have to be solved by iteration, they are a valuable guide to the iteration that leads to a minimum g.

For comparison with the normal, Table 7 shows the percent reductions $100(1 - g)$ with equal-sized classes, for a series of χ^2 and t distributions. In both cases the percent reductions are substantially less than with the normal when the degrees of freedom are small. The reductions for optimum class boundaries (not shown) are somewhat closer to the normal optima but are still inferior.

To come finally to the objective of this section: Table 8 presents the percentage reductions in the variance of $\bar{y}_2 - \bar{y}_1$ that result from stratified matching on x, the formula being $100(1 - g)\rho^2$ by (6.1). The main conclusions are: (i) subclassification with a small number of subclasses suffers most, in comparison with an infinite number of classes, when the potential gains in precision are large (ρ high); (ii) on the other hand, with $\rho < 0.7$, as occurs in most applications, the use of at least 4 subclasses realizes nearly all the potential gain.

If x is continuous, we can compare stratified matching with a

TABLE 7

PERCENT REDUCTIONS IN VARIANCE FOR NON-NORMAL DISTRIBUTIONS, WITH EQUAL-SIZED SUBCLASSES

	c = Number of subclasses				
Distribution	2	3	4	5	6
χ_4^2	55.5	71.5	79.7	83.9	86.8
χ_6^2	58.2	73.8	81.4	85.7	88.2
χ_{10}^2	60.3	76.3	83.5	86.8	89.6
χ_{20}^2	62.0	77.4	84.7	88.4	90.6
t_3	40.5	52.3	57.8	61.7	64.4
t_5	54.0	68.9	75.4	79.8	82.7
t_{10}	59.8	75.6	82.3	86.5	89.2
t_{20}	61.9	77.9	84.6	88.7	91.2
Normal	63.7	79.3	86.1	89.7	91.9

TABLE 8

PERCENT REDUCTIONS IN $V(\bar{y}_2 - \bar{y}_1)$ RESULTING FROM STRATIFIED MATCHING ON x, WITH EQUAL-SIZED SUBCLASSES AND x NORMAL

ρ	c = Number of subclasses 2	3	4	5	6	∞
0.2	2.5	3.2	3.4	3.6	3.7	4.0
0.3	5.8	7.1	7.7	8.1	8.3	9.0
0.4	10.2	12.7	13.8	14.4	14.7	16.0
0.5	15.9	19.8	21.5	22.4	23.0	25.0
0.6	22.9	28.5	31.0	32.3	33.1	36.0
0.7	31.2	38.9	42.2	44.0	45.0	49.0
0.8	40.8	50.8	55.1	57.4	58.9	64.0
0.9	51.6	64.2	69.7	72.7	74.4	81.0

linear covariance adjustment applied to two independent samples of size n. Covariance gives reductions of approximately $[R_\infty - (100 - R_\infty)/2n]$, where R_∞ is the reduction in the column $c = \infty$ in Table 8, the term $(100 - R_\infty)/2n$ being the usual allowance for the sampling error of the regression coefficient. A few calculations will show that for samples of any reasonable size, covariance gives greater gains than stratified matching, as Billewicz also concluded.

Table 8 applies to stratified matching with x normal. Earlier results in this section suggest that reductions will be somewhat less (i) under adjustment by subclassification and (ii) when x is non-normal, at least if the results for χ^2 and t are typical of those with non-normal x.

7. EFFECT OF ERRORS OF MEASUREMENT IN x

We return to the consideration of bias in y.

As before, we assume that y has a linear regression on the correct measurement x. However, the measurement actually used for adjustments or for the construction of an ordered classification is X, which also has a linear regression on x in each population of the form

$$X = \gamma + \beta x + d. \tag{7.1}$$

The term γ represents a constant bias in measurement, while if β differs from 1, this implies a bias that changes as x changes. The variable d is the random component of the error of measurement, with mean 0, variance σ_d^2. If x and d are normally and independently distributed, standard bivariate normal theory shows that the regression of x

on X is linear, with regression coefficient cov $(x, X)/\sigma_X^2 = \beta\sigma_x^2/(\beta^2\sigma_x^2 + \sigma_d^2)$.

Suppose that x has mean μ in population 1 and mean $\mu + \theta$ in population 2. On account of relation (7.1), the means of X in the two populations will differ by an amount $\beta\theta$. Since the adjustments are made by means of X, their effect on X is to reduce its bias by an amount $\beta\theta\lambda$, where λ is the proportional reduction that was shown, as a percentage, in Table 6. Because of the linear regression of x on X, this reduction of $\beta\theta\lambda$ in X produces a reduction of amount

$$(\beta\theta\lambda)\cdot\frac{\beta\sigma_x^2}{\beta^2\sigma_x^2 + \sigma_d^2} = \frac{\theta\lambda}{1 + h}$$

in x, where $h = \sigma_d^2/\beta^2\sigma_x^2$. Thus the proportional reduction in x, which equals that in y on account of the linear regression of y on x, is $\lambda/(1 + h)$. The quantity h can be regarded as the ratio of the variance of the random errors of measurement in x to the variance of the correct measurements x, because if X is divided by β so that a given change in x provides the same expected change in X, the random error of measurement becomes d/β, with variance σ_d^2/β^2.

To sum up for an ordered classification with a linear regression of y on x, the adjustment fails to remove all the bias in y for two reasons. Even if there are no random errors of measurement, the ordered classification being constructed from x itself or equivalently from a linear function $\gamma + \beta x$, the proportional reduction in the bias of y is not unity but γ as given in Table 6. Secondly, there may be errors of measurement in the variate X by which the ordered classification is made. These errors introduce some misclassifications so far as the correct measurement x is concerned, and reduce λ to $\lambda/(1 + h)$.

Consequently, some idea of the value of h is essential in forming a judgment on the effectiveness of adjustment. Although interest in the study of errors of measurement has increased in recent years, not much seems to be known about the values of h. In practice, three methods are commonly used to estimate h: (i) independent measurements of the same objects by the fallible and the correct method. This approach is feasible only when a correct method exists—usually in areas where the technique of measurement is far advanced; (ii) repeated independent measurements by the fallible (X) method or by competing fallible methods; (iii) repeated measurement by a superior but not perfect instrument. With methods (ii) and (iii), h can be estimated for continuous data as shown by Grubbs [1948] and for ordered classifications by unpublished methods reported to me by F. Mosteller. Estimation of h for a binomial variate has been discussed by Hansen *et al.* [1964].

From such results as I have seen, it is evident that h can vary in practice from a trivial amount to values substantially over 1 in difficult measurements. Thus, when errors of measurement are taken into account, the proportions in Table 6 may be changed only negligibly or may be reduced to figures less than half those shown.

With ordered classifications, the value of h is related to the percentage of wrong classifications by the fallible measurement X in method (i) above, and to the percentage of disagreements in classification by two independent readings X_1 , X_2 of the fallible measurement in method (ii). From the model (7.1), cov $(x, X) = \beta\sigma_x^2$ and cov $(X_1 , X_2) = \beta^2\sigma_x^2$, so that x and X follow a bivariate normal with correlation coefficient $1/\sqrt{1 + h}$, while X_1 and X_2 follow a bivariate normal with correlation coefficient $1/(1 + h)$. Hence, if specimens are being assigned to one of two classes of equal size, the value of h corresponding to a given proportion p of wrong classifications may be read from the tables of the bivariate normal, by noting the value of p needed to make the combined area in the two quadrants $(x > 0, X < 0)$ and $(x < 0, X > 0)$ equal to p.

As an illustration, Table 9 shows these percentages for a series of values of h. Also shown are the percentage reductions $100h/(1 + h)$ caused by errors in measurement in the amount of bias removed by subclassification, and the approximate percentages $63.7/(1 + h)$ of bias that would actually be removed by adjustment on X. For measurements of reasonably high precision with $h < 0.10$, the percentage of wrong classifications agrees closely with the percentage reduction in the bias removed. As the percentage of wrong classifications increases, the effectiveness of adjustment on X drops more sharply. With 15% of wrong classifications, only about half the bias is removed, and with 25%, less than one-third. Disagreements of up to 10% between two fallible measurements correspond to satisfactorily high precision, the reduction in bias removed being at most 5%.

The regression model (7.1) used here is likely to be an oversimplification of the relation between X and x in many applications. The relation may not be linear, and the variance σ_d^2 may depend on the value of x. Some investigation was made of a model in which d is correlated with x, but this did not provide any essential generalization, since d can be split into its regression on x and a random component independent of x, bringing us back to an ordinary regression model. As already mentioned, what seems most needed in order to appraise the seriousness of errors of measurement are data from which the values of h in different measurement problems can be estimated. Further, this paper has considered only adjustments on a single variable x. The effectiveness

TABLE 9

PERCENTAGE OF WRONG CLASSIFICATIONS AND OF DISAGREEMENTS BETWEEN REPEATED MEASUREMENTS (2 CLASSES) FOR SPECIFIED RATIOS h OF ERROR VARIANCE TO TRUE VARIANCE

h	Percentage of: wrong classifications	disagreements	reduction in bias removed	bias removed
0.00	0.0	0.0	0.0	63.7
0.05	4.8	9.7	4.8	60.7
0.10	9.5	13.6	9.1	57.9
0.20	13.3	18.7	16.7	53.1
0.30	15.9	22.0	23.1	49.0
0.50	19.5	26.8	33.3	42.5
1.00	25.0	33.3	50.0	31.8
1.50	28.2	36.9	60.0	25.5
2.00	30.4	39.2	66.7	21.2

of two-way and three-way classifications used to adjust simultaneously for two and three x-variables also merits investigation.

ACKNOWLEDGEMENTS

This work was assisted by Contract Nonr 1866(37) with the Office of Naval Research, Navy Department, and by Grant GS-341 from the National Science Foundation.

L'EFFICACITE DE L'AJUSTEMENT PAR SOUS-CLASSIFICATION POUR LA SUPPRESSION DES BIAIS DANS LES ETUDES D'OBSERVATION

RESUME

Dans certaines recherches la comparaison des moyennes d'une variable y entre deux groupes étudiés peut être biaisée parce que y est lié à une variable x dont la distribution diffère dans les deux groupes. Un procédé fréquemment utilisé pour essayer de supprimer ce biais est l'ajustement par sous-classification. L'étendue x est divisée en c sous-classes. Les moyennes pondérées des moyennes de sous-classes de y sont comparées en utilisant les mêmes poids pour chaque groupe étudié. L'efficacité de cette procédure pour supprimer le biais dépend de plusieurs facteurs, mais, pour des relations monotones entre y et x, une approche analytique suggère que pour $c = 2, 3, 4, 5$, et 6 les pourcentages de biais supprimés sont grossièrement respectivement 64%, 79%, 86%, 90%, et 92%. Ces chiffres devraient également servir de guide lorsque x est d'une classification ordonnée, (par exemple: aucun, léger, modéré, grave) qui peut être considérée comme un regroupement d'une variable sous-jacente continue. Le degré avec lequel l'ajustement réduit l'erreur d'échan-

tillonnage de la différence estimée entre les moyennes de y est également examiné. On peut noter au passage un résultat intéressant: pour une variable x normale, la réduction en pourcentage du biais de $\bar{x}_2 - \bar{x}_1$ dû à l'ajustement est égale à la réduction en pourcentage de sa variance.

Sous un modèle mathématique simple, les erreurs de mesure sur x réduisent le montant du biais supprimé d'une fraction $1/(1 + h)$ de sa valeur où h est le rapport de la variance des erreurs de mesure à la variance des mesures correctes. Les classifications ordonnées étant souvent utilisées du fait que x est difficile à mesurer, h peut être assez substantiel dans de tels cas; cependant des renseignements complémentaires sont nécessaires sur les valeurs de h qui sont typiques en pratique.

REFERENCES

Billewicz, W. Z. [1965]. The efficiency of matched samples: An empirical investigation. *Biometrics 21*, 623–43.

Cox, D. R. [1957]. Note on grouping. *J. Amer. Statist. Ass. 52*, 543–7.

Dalenius, T. [1957]. *Sampling in Sweden.* Almqvist and Wicksell, Stockholm.

Grubbs, F. E. [1948]. On estimating precision of measuring instruments and product variability. *J. Amer. Statist. Ass. 43*, 243–64.

Hansen, M. H., Hurwitz, W. N., and Pritzker, L. [1964]. Measurement errors in censuses and surveys. In *Contributions to Statistics.* Statistical Publ. Company, Calcutta, India.