Rubin, D. B. (1976b), Multivariate Matching Methods That Are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Samples Sizes, *Biometrics*, **32**(1), 121–132, 955.

Rubin, D. B. (1979), Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies, *Journal of the American Statistical Association*, **74**, 318–328.

Smith, A. H., Kark, J. D., Cassel, J. C., and Spears, G. F. S. (1977), Analysis of Prospective Epidemiologic Studies by Minimum Distance Case-Control Matching, *American Journal of Epidemiology*, **105**(6), 567–574.

Snedecor, G. W., and Cochran, W. G. (1967), *Statistical Methods*, 6th ed., Ames, IA: Iowa State University Press.

Ury, H. K. (1975), Efficiency of Case-Control Studies with Multiple Controls per Case: Continuous or Dichotomous Data, *Biometrics*, **31**, 643–649.

Yinger, J. M., Ikeda, K., and Laycock, F. (1967), Treating Matching as a Variable in a Sociological Experiment, *American Sociological Review*, **23**, 801–812.

CHAPTER 7

Standardization and Stratification

7.1 Standardization—Example and Basic Information 115
7.2 Choice of Standard Population 118
7.3 Choice of Standardization Procedure 119
7.4 Statistical Considerations for Standardization 120
7.4.1 Bias 120
7.4.2 Precision 121
7.5 Extension of Standardization to Case-Control Studies 121
7.6 Stratification 122
7.6.1 Estimators of the Odds Ratio 122
7.6.2 Comparisons of the Odds Ratio Estimators 124
7.7 Standardization and Stratification for Numerical Outcome Variables 126
7.8 Extension to More Than One Confounding Factor 127
7.9 Hypothesis Testing 128
Appendix 7A Mathematical Details of Standardization 128
7A.1 Notation 128
7A.2 Computation of Directly and Indirectly Standardized Rates 129
7A.3 Bias of Indirect Standardization 130
7A.3.1 Relative Risk 130
7A.3.2 Difference of Rates 131
7A.4 Equivalence of Direct and Indirect Standardization 132
Appendix 7B Stratified Estimators of the Odds Ratio 134
7B.1 Maximum Likelihood and Conditional Likelihood Estimators 134
7B.2 Woolf and Modified Woolf Estimators 136
7B.3 Mantel–Haenszel Estimator 137
References 138

Standardization and stratification are related adjustment procedures that are applicable when the confounding factor is categorical. The outcome and risk factors may be either categorical or numerical, except that for application to cohort studies, the risk factor must be categorical, and for case-control studies, the outcome must be categorical.

The goal of any adjustment procedure is to correct for differences in the confounding factor distributions between the treatment groups. Standardization does this by estimating what would have been observed had the confounding factor distributions been the same in the two groups being compared. In the data to be presented in Section 7.1, for example, the rates of death due to breast cancer are compared for two groups of women. The two groups of women differ with respect to age, an important confounding factor when comparing death rates. Standardization in this case estimates death rates based on a common age distribution. This common age distribution, or more generally, the common confounding factor distribution is taken from some other group, known as the standard population; hence the term "*standard*ization."

The term "stratification," when applied to adjustment procedures, can be used in two ways. The first and more general use is to describe any adjustment procedure that divides the study population into groups (strata) based on the values of the confounding factors and then combines information across groups to provide an estimate of the treatment effect. In this general sense, standardization is a stratified procedure where the standard population provides the basis for combining information across strata.

In this book, however, the terms "stratification" or "stratified analysis" will only be used in a second, more restrictive, way. This second usage is consistent with the first in that the study population is divided into strata and information is combined across groups. The restriction is that the basis for combining across groups be some statistical criterion, such as maximum likelihood or minimum variance, without reference to any standard population.

Standardization and stratification are employed for two purposes: (a) to provide summary statistics for comparing different populations with respect to such items as mortality, price levels, or accident rates; and (b) to yield estimators of the difference in rates or means between two populations or of the relative risk (θ) or odds ratio (ψ) that are unbiased, or at least approximately so.

In Section 7.1 we present the principles of standardization for the simple case of a cohort study with dichotomous risk and outcome factors. Some considerations in the choice of a standard population and standardization procedure are given in Sections 7.2 and 7.3, respectively; the bias and precision of standardized estimators are discussed in Section 7.4; and the extension to case-control studies is considered in Section 7.5. General formulas for direct and indirect standardization and more detailed bias considerations are given in Appendix 7A.

Stratification is introduced in Section 7.6 with emphasis on estimators of the odds ratio. The odds ratio estimators are considered in greater mathematical detail in Appendix 7B. The extension of standardization and stratification to numerical outcomes and multiple confounding factors is presented in Sections 7.7 and 7.8.

If the confounding variable is numerical, standardization and stratification can be applied by first categorizing the confounding variable (as in Table 7.1 for the numerical confounding variable, age). The effects of this categorization on the bias of the stratified estimators are discussed in Chapter 13. For now, it is sufficient to note that the estimators will always be biased, even in large samples. Comments in this chapter on bias are for a categorical confounding variable (except when considering McKinlay's work in Section 7.6.2). For the case of a numerical confounding variable, Cochran's work (1968) for frequency matching (Section 6.7) gives some guidance for choosing the number and sizes of the strata. Logit analysis (Chapter 9) and analysis of covariance (Chapter 8) are alternative procedures that do not require stratifying a numerical confounding variable.

7.1 STANDARDIZATION—EXAMPLE AND BASIC INFORMATION

The data in Table 7.1 on breast cancer death rates among females aged 25 or older is based on work by Herring (1936, Tables I and II). In this case, the risk factor is marital status with two categories, single (never married) and married (including widows and divorcees), and the outcome is death due to breast cancer. The breast cancer death rates, per 100,000 population, of 15.2 for single women and 32.3 for women who were ever married are called *crude* (or unadjusted) *death rates* because they have not been adjusted for any possible confounding. Crude rates are calculated by simply ignoring any possible confounding factors. For example, the crude rate for single women is found by dividing the average annual number of deaths due to breast cancer (1438) by the total population of single women (94.73). The crude relative risk of death due to breast cancer is 2.1 = 32.3 ÷ 15.2, purporting to indicate that women who marry are more than twice as likely to die of breast cancer as are women who never marry.

An examination of the death rates by age—the *age-specific rates*—indicates that age is an important factor. Age is a confounding factor in this circumstance because death rates increase with age and the ages of married and single women differ; the married women tend to be older (80% of the single women are younger than 35 as compared to only 35% of the women who had married). (The definition of a confounding factor is given in Section 2.1.) A fair comparison of the cancer death rates requires that age, at least, be adjusted for.

One approach to adjustment is *direct standardization.* This approach asks

Table 7.1 Breast Cancer Mortality in Females in the United States (1929–1931)

	Single Women		Ever-Married Women		All Women	
Age (yr)	1930 Population (100,000's)	Average Annual Breast Cancer death rate (per 100,000 population)	1930 Population (100,000's)	Average Annual Breast Cancer death rate (per 100,000 population)	1930 Population (100,000's)	Average Annual Breast Cancer death rate (per 100,000 population)
15–34	76.15	0.6	89.57	2.5	165.72	1.6
35–44	7.59	24.9	61.65	17.9	69.24	18.7
45–54	5.22	74.7	46.67	43.1	51.89	46.2
55–64	3.43	119.7	31.11	70.7	34.54	75.5
65–74	1.88	139.4	18.14	89.4	20.02	94.1
≥75	0.45	303.8	7.80	137.7	8.25	146.8
Total	94.73	15.2	254.94	32.3	349.67	27.6
Average annual number of deaths	1438		8228		9666	

how many cancer deaths would have occurred if the age distribution for both the married and single women had been the same as in some standard population, but the age-specific rates were the same as observed? In this example, a natural standard is the 1930 age distribution of all women in the United States. The directly standardized cancer mortality rates are (per 100,000 population):

Single women:

$$\frac{(165.72)\,(0.6) + \cdots + (8.25)\,(303.8)}{349.67} = \frac{15{,}131.27}{349.67} = 43.3$$

Married women:

$$\frac{(165.72)\,(2.5) + \cdots + (8.25)\,(137.7)}{349.67} = \frac{9257.95}{349.67} = 26.5.$$

Single women have a higher age-adjusted mortality rate, and the directly standardized relative risk is $\hat{\theta}^{\mathrm{D}} = 26.5/43.3 = 0.6$, indicating that, after correcting for age differences, women who marry actually have a lower risk of dying of breast cancer than do women who remain single. Note that $\hat{\theta}^{\mathrm{D}}$ is the ratio of the expected numbers of cancer deaths in the standard population based on the age-specific death rates for married women (9257.95) to the expected number of deaths in the standard population based on the rates for single women (15,131.27).

An alternative approach is *indirect standardization.* This approach asks how many cancer deaths would have occurred among single women if the age distributions for the single and married women were the same as observed, but the age-specific mortality rates had been the same as in some standard population? As we will see in Section 7.3, when applying indirect standardization, it is best to select the "standard population" to be one of the two groups being compared.

The results of indirect standardization are often quoted as *standard mortality* (or morbidity) *ratios,* which are ratios of the observed deaths for each category of the risk factor to the expected deaths given the standard age-specific rates. The indirectly standardized rate for each risk factor category is then found by multiplying the standard mortality ratio for that category by the crude rate for the standard population. (The rationale for this round about calculation of a standardized rate is given in Section 7A.2.) Indirectly standardized relative risks, $\hat{\theta}^{I}$, are found as ratios of the indirectly standardized rates. In the special case where the standard rates are taken to be those corresponding to one of the two groups being compared, the standard mortality ratio itself turns out to be a relative risk.

For our example, taking the standard rates to be the age-specific breast cancer death rates for married women, the expected deaths for single women are

$$(76.15)\,(2.5) + \cdots + (0.45)\,(137.7) = 1023.8.$$

The standard mortality ratio is then 1.40 = 1438 ÷ 1023.8, indicating that there were 40% more deaths among single women than would have been expected if their age-specific rates had been the same as for married women. The indirectly standardized breast cancer death rate for single women is found by multiplying the standard mortality ratio by the crude mortality rate for the married women:

$$(1.40)(32.3) = 45.2.$$

The indirectly standardized relative risk comparing married to single women is then

$$\hat{\theta}^{\mathrm{I}} = \frac{32.3}{45.2} = \frac{1}{1.40} = 0.71.$$

The standard mortality ratio of 1.40 is the inverse relative risk, that is, of remaining single compared to getting married.

It is common for different standardization procedures to yield different standardized rates and estimates of relative risk as occurred here. This emphasizes that the primary purpose of standardization is to provide a single summary statistic (mortality rates here) for each category of the risk factor so that the categories may be compared. The numbers that are most meaningful are the age-specific mortality rates, and standardization is not a substitute for reporting the specific rates.

To emphasize this point further, look again at the specific rates in Table 7.1. For women under the age of 35, married women have a slightly higher breast cancer death rate than do single women; for women age 35 or over, single women have the higher breast cancer mortality rate, and the difference and ratio of the rates varies with age. This is an example of interaction between the confounding factor age and the treatment effect (Section 3.3). This interaction is an important fact that is not conveyed by the reporting of a single relative risk, standardized or not. One consequence of this is that a different choice of a standard population could have resulted in standardized rates that, like the crude rates, were higher for married women than for single. The choice of standard thus becomes an important issue. Some guidelines for choosing the standard will be discussed in Section 7.2. Spiegelman and Marks (1966) and Keyfitz (1966) give examples of how different choices for the standard population can affect the standardized rates and relative risks. Keyfitz compares the 1963 female mortality rates in 11 countries using three different standard age distributions and finds that the ranking of the countries depends on the choice of standard.

The direct and indirect methods are the most important standardization procedures, but not the only ones. There are many standardized indices that have been developed for particular fields. Kitagawa (1964, 1966) discusses many of these alternatives, particularly with reference to demography.

7.2 CHOICE OF STANDARD POPULATION

The choice of standard population is, in general, a contextual decision. When standardization is being employed for comparison purposes, there are two commonsense guidelines for choosing the standard population. The first is to use the data for the entire population that the study subjects are chosen from. This was done for the breast cancer mortality example by using the data for all women in the United States in 1930 as the standard for direct standardization. If the population data are not available, an alternative is to combine all the risk factor groups (i.e., take the standard population to be the entire sample being studied). The rationale behind this alternative is that summing over the risk factor should yield a "population" that approximates the real population of interest. This approach will approximate the population well if the sampling fractions in the risk factor groups are equal, or nearly so; the approximation will be poor, if, for example, the subjects in one category of the risk factor are all persons known to have been exposed to the risk factor and the subjects in the second category are only a portion of those not exposed.

The second guideline is applicable in cases where all but one category of the risk factor correspond to a treatment of some sort, and the remaining category corresponds to the absence of a treatment. Then, the nontreatment category is a reasonable choice of standard population. For example, when Cochran (1968) standardized lung cancer rates for age, he chose the nonsmokers to be the standard population. The cigarette smokers and cigar and/or pipe smokers were two "treatment" groups.

As discussed in Section 7.1, the choice of standard can make a difference in the comparison of risk factor groups. Therefore, it is important to report the specific rates. If a single summary statistic is still required, the standard should be picked to resemble the risk factor groups as much as possible, so as to preserve, to the extent possible, the meaningfulness of the standardized comparison. For the data of Table 7.1, for example, the 1960 age distribution of males in Mexico would be an inappropriate choice of standard.

7.3 CHOICE OF STANDARDIZATION PROCEDURE

The choice of standardization procedure can depend on many considerations. If the total sample size and specific rates for the categories of the risk factor are known but the numbers of individuals at each level of the confounding factor are not known, then directly standardized rates can be calculated but indirectly standardized rates cannot. Conversely, indirectly standardized rates can be calculated if the specific rates are not known but the total number of deaths (outcomes) and the specific rates for the standard are known.

There is one important caution regarding indirect standardization. It is possible to have two categories of the risk factor with identical specific rates but different indirectly standardized rates. To see this, consider the artificial data in Table 7.2, with three risk factor groups and two categories in the confounding factor. The specific rates for the first two risk factor categories are identical so a proper adjustment procedure should yield a relative risk of 1.0. The crude rates are 0.82 for the first category and 0.18 for the second, so the crude relative risk is 4.56, reflecting the very different confounding factor distributions in the two groups being compared.

Now consider the direct and indirect standardized relative risks with the total of the three risk factor groups as the standard population. Following the procedures of Section 7.1, the directly standardized rates for the first two risk factor categories are both 0.50, so the directly standardized relative risk is 1.0. The indirectly standardized rates are 0.62 for the first risk factor category and 0.26

Table 7.2 Artificial Data to Demonstrate Comparability Problem of Indirect Standardization

	Risk Factor Category							
	1		2		3		Total	
Confounding Factor Category	Sample Size	Rate	Sample Size	Rate	Sample Size	Rate	Sample Size	Rate
1	900	0.9	100	0.9	1000	0.5	2000	0.7
2	100	0.1	900	0.1	1000	0.5	2000	0.3

for the second, yielding an indirectly standardized relative risk of 2.38. This result is more reasonable than the crude relative risk of 4.56, but still not good. The indirect method would only have worked in this example if the specific rates in the standard population happened to be the same as for categories 1 and 2 of the risk factor.

Indirect standardization is best used only for comparing two groups when one of those groups is the standard. In that case, the two methods of standardization are equivalent, in the sense that equal estimates of θ can be obtained by particular choices of the standard for each method. In addition, the indirectly standardized rates will be equal if all the specific rates are equal. Mathematical details are given in Sections 7A.3 and 7A.4.

7.4 STATISTICAL CONSIDERATIONS FOR STANDARDIZATION

As with other adjustment techniques, standardized estimation of treatment effects is most meaningful when the treatment effect is constant over the confounding factor strata [i.e., when there is no interaction (Section 3.3)]. In this section we will consider the bias and precision (variance) of the standardized estimators of the constant treatment effect, whether relative risk or difference of rates.

7.4.1 Bias

If the difference in risk factor rates is the same for each category of the confounding factor, direct standardization yields unbiased estimates of the difference between the risk factor rates. For estimating the relative risk, if the sample relative risks are the same, say $\hat{\theta}$, within each category of the confounding factor, then the directly standardized estimate of relative risk is also equal to θ (as demonstrated in Section 7A.3.1). This implies that the directly standardized relative risk will be approximately unbiased in large samples.

In general, the indirectly standardized estimators of both parameters are biased. An exception occurs when comparing two groups by means of relative risk and one of these groups is the standard (see Section 7A.4). Indirect standardization does not yield unbiased estimates of the difference between the rates because the standard mortality ratio is a ratio.

7.4.2 Precision

Indirect standardization has been found to be more precise than direct standardization for estimating rates (Bishop, 1967) and relative risks (Goldman, 1971). Goldman further showed that the precision of directly standardized relative risks could be improved by first applying the log-linear model technique (Chapter 10) and then directly standardizing using the fitted rates. Details are presented in Bishop (1967), Goldman (1971), and Bishop et al. (1974, Sec. 4.3).

7.5 EXTENSION OF STANDARDIZATION TO CASE-CONTROL STUDIES

Since we cannot estimate rates directly (see Section 3.1), much of the previous material is not applicable to case-control studies. We must instead ask how to obtain a standardized estimate of the odds ratio. Miettinen (1972) developed a procedure motivated by the idea, presented in Section 7.1, that the standardized relative risks are ratios of expected numbers of deaths (or other dichotomous outcomes). For case-control studies, Miettinen proposed using the ratio of the expected numbers of cases in the two risk factor groups, where the expectation is based on a standard distribution of numbers of controls. Letting C_k, a_{rk}, and c_{rk} denote standard numbers of controls, observed numbers of cases, and observed numbers of controls, respectively, where k denotes the categories of the confounding factor and r the categories of the risk factor ($r = 1$ if the risk factor is present; $r = 0$ if not), the standardized estimator of the odds ratio is

$$\hat{\psi}^M = \frac{\sum_{k=1}^{K} C_k(a_{1k}/c_{1k})}{\sum_{k=1}^{K} C_k(a_{0k}/c_{0k})}.$$

To see that $\hat{\psi}^M$ is the ratio of the "expected" number of cases in the two risk factor groups, let A_{1k} be the number of cases corresponding to C_k controls in the risk factor present group. To find A_{1k}, set the ratio of expected numbers of cases to controls equal to the observed ratio

$$\frac{A_{1k}}{C_k} = \frac{a_{1k}}{c_{1k}}$$

and solve for A_{1k}:

$$A_{1k} = \frac{C_k a_{1k}}{c_{1k}}.$$

Similarly,

$$A_{0k} = \frac{C_k a_{0k}}{c_{0k}}.$$

Summing A_{1k} over the K confounding factor strata yields the total number of expected cases, the numerator of $\hat{\psi}^M$. The ratio $\hat{\psi}^M$ then compares the number of cases expected in the risk-factor-present group to the number expected in the risk-factor-absent group, based on the same standard distribution of controls (the C_k). Considerations in the choice of standard, discussed in Section 7.2, apply here as well.

Miettinen shows that $\hat{\psi}^M$ can be written as a weighted average of the odds ratios from each of the K confounding factor strata (the specific odds ratios). Therefore, if the odds ratios, ψ_k, are constant over all categories of the confounding factor, $\hat{\psi}^M$ will be an approximately unbiased estimator of ψ in large samples (within each stratum).

7.6 STRATIFICATION

The method of stratification differs from standardization in that a statistical criterion, such as minimum variance, rather than a standard population, is the basis for combining across confounding factor strata. In this section we will cover the best-studied case of stratification, that of estimating the odds ratio for dichotomous risk and outcome factors in either cohort or case-control studies. Throughout Section 7.6 it will be assumed that the odds ratio is the same in all the confounding factor strata. The formulas for the various estimators are presented in Appendix 7B. Stratified estimation of the difference of means is covered in Section 7.7

7.6.1 Estimators of the Odds Ratio

A large number of estimators of the odds ratio have been proposed. Gart (1962) presented the maximum likelihood estimator and (1966) a modification to Woolf's (1955) estimator (the "modified Woolf" estimator). Birch's (1964) and Gart's (1970) estimators are approximations to yet another estimator, the conditional maximum likelihood estimator (Gart, 1970). Goodman (1969) proposed approximations to the maximum likelihood and conditional maximum likelihood estimators. A well-known estimator is that of Mantel and Haenszel (1959).

There are two different maximum likelihood estimators, usual and conditional, because there are two different sampling situations to be considered. As the theoretical properties of the various estimators depend on which sampling situation is appropriate, we must begin with an explanation of the two cases, specifically emphasizing what is meant by a large sample in each of the two cases.

Consider the situation where, within each confounding factor stratum there is some number of subjects in each of the two study groups, and suppose tha we wish to add more subjects so as to increase the total sample size. Then, there are two choices: more subjects can be added to the existing strata; or new strata can be added with corresponding, additional subjects.

The first case is the most commonly considered. Often the number of strata is fixed by the nature of the situation. For example, if the confounding factor was sex, the number of strata are fixed at two and the sample size can be increased only by adding more males and females. In such a situation, "large sample" means that the sample sizes in each study group within each strata are large, regardless of the number of strata.

Consider now a study that is conducted cooperatively in many institutions, and suppose that institution is the confounding variable of interest. Each stratum will then consist of the subjects from a particular institution. In this study a larger sample could be obtained in two ways. The first would be as above, namely adding subjects from each currently participating institution. The second is to add more institutions and select subjects from the new institutions. For every additional institution there will be an additional stratum for the confounding factor. "Large sample" in this second case means that the number of confounding factor strata is large, regardless of the sample sizes within each stratum.

To summarize, there are two definitions of large sample. In the first, the sample sizes within each stratum are large; in the second, the number of strata are large. This distinction is important because estimators can behave differently in the two cases. In particular, the properties of the (usual) maximum likelihood estimator apply only in the first case. The second case requires a different estimator, the conditional maximum likelihood estimator. Each maximum likelihood estimator will be approximately unbiased and normally distributed in large samples as defined for the appropriate sampling scheme (Gart, 1962; Andersen, 1970).

The numerical difficulties in solving for the conditional maximum likelihood estimator led Birch (1964) and Goodman (1969) to propose approximations that are easier to calculate. The Birch and Goodman estimators are unbiased in large samples only if the odds ratio is 1. In terms of bias considerations, the conditional maximum likelihood estimator is therefore preferable. Gart's (1970) approximation to the conditional maximum likelihood estimator is applicable if the sample sizes within each stratum are large. This approximation will be approximately unbiased when both the number of strata and sample sizes within each stratum are large. Although the unbiasedness holds for any value of the odds ratio, requiring both the number of strata and the sample sizes to be large is very restrictive.

The (usual) maximum likelihood estimator that is appropriate when the sample sizes are large within each stratum is the basis of comparison for the remaining estimators. In the appropriate large samples, this estimator is approximately unbiased and no other unbiased estimator has a lower variance. In large samples, then, this is a good choice of estimator.

Woolf's (1955) estimator is equivalent, in the sense of having the same large-sample distribution, to the maximum likelihood estimator. However, as will be shown in detail in Section 7B.2, this estimator cannot be calculated if either of the observed proportions in *any* stratum is 0 or 1. In practice this means the Woolf estimator will be virtually useless when dealing with rare outcomes in cohort studies or rare risk factors in case-control studies.

Gart (1966) suggested a modification to the Woolf estimator that does not suffer from this problem while retaining the large-sample equivalence to the maximum likelihood estimator. There are thus three estimators that are equivalent in large samples: the maximum likelihood, Woolf, and modified Woolf Estimators. What is known about their small-sample properties is considered in Section 7.6.2.

The Mantel–Haenszel (1959) estimator is also approximately unbiased and normally distributed in large samples (large within each stratum), but its variance is larger than that of the maximum likelihood estimator unless the odds ratio is 1 (Hauck, 1979). In large samples, then, one of the three equivalent estimators noted above would be preferable to the Mantel–Haenszel estimator.

7.6.2 Comparisons of Odds Ratio Estimators

Using simulation, McKinlay (1975) compared the bias, precision, and mean squared error of the modified Woolf, Mantel–Haenszel, and Birch estimators for the case of a numerical confounding variable that is stratified into various numbers of strata. As mentioned earlier, all the standardized and stratified estimators will be biased in such a case, even in large samples. McKinlay's work is discussed in greater detail in Section 12.2.2.

Based on the mean squared errors of the estimators, McKinlay recommended the modified Woolf estimator, but with some reservations. The modified Woolf estimator has a smaller variance than the Mantel–Haenszel estimator, and this is reflected in the mean squared errors. However, the bias of the modified Woolf estimator increases with increasing number of strata. McKinlay noted that "only Mantel and Haenszel's estimator consistently removed bias in all the simulated situations considered—a property which is masked in this investigation by the relatively large variance" (p. 863). In terms of bias removal, then, the Mantel–Haenszel estimator is to be preferred. In addition, the difference in mean squared errors between the Mantel–Haenszel and modified Woolf estimators became negligible for the large samples (total sample size of 600) in McKinlay's study.

In an unpublished study, W. Hauck, F. Leahy, and S. Anderson addressed the question of whether McKinlay's 1975 results are applicable to the case of a categorical confounding factor where the estimators would be approximately unbiased in large samples. This study was patterned after McKinlay's 1975 study and compared the modified Woolf, Mantel–Haenszel, and (usual) maximum likelihood estimators. In terms of bias, the modified Woolf was least and maximum likelihood most biased, except for increasing values of the odds ratio and large sample sizes where the Mantel–Haenszel estimator was the least biased and the modified Woolf the most. However, the bias was small for all three estimators for the sample sizes and number of strata considered by McKinlay. In terms of variance and mean squared error, the modified Woolf was more precise, sometimes considerably so, than the other two and the maximum likelihood estimator least precise.

In another simulation study with a numerical confounding factor and large samples (total sample sizes of 200 to 1000), McKinlay (1978) compared the Mantel–Haenszel estimator, Gart's (1970) asymptotic approximation to the conditional maximum likelihood estimator, and the modified Woolf estimator. The results for the modified Woolf and Mantel–Haenszel estimators were similar to her 1975 results, namely that the Woolf estimator was usually most precise but that its bias tended to increase with increasing numbers of strata from 2 to 10, while the Mantel–Haenszel estimator was preferable in terms of bias, particularly for the larger number of strata. The Gart estimator, a close approximation to the conditional maximum likelihood estimator in the cases considered by McKinlay, was never better than the Mantel–Haenszel estimator in terms of either bias or precision.

The three studies agree that on purely bias considerations, the Mantel–Haenszel estimator is best, selected over the modified Woolf estimator on the grounds of consistency. If precision is taken into account by considering mean squared error, then, for the cases considered, the modified Woolf estimator is best.

This is an example of a common statistical problem of making a trade-off between bias and precision. Since the modified Woolf estimator is sometimes considerably more precise, and since the biases of all the estimators considered are not large, this would seem to be the estimator of choice. What is of concern, however, is the tendency for the bias of the modified Woolf estimator to increase with increasing number of strata for a given total sample size and with increasing distance of the odds ratio from the null value of 1. This implies that the modified Woolf estimator is more sensitive to the sample sizes within each stratum than is the Mantel–Haenszel estimator. Consequently, the modified Woolf estimator can be clearly preferred only for a small number of strata with large sample sizes within each stratum; otherwise, the Mantel–Haenszel estimator is a good choice.

7.7 STANDARDIZATION AND STRATIFICATION FOR NUMERICAL OUTCOME VARIABLES

The standardization results of Sections 7.1 to 7.3 and Appendix 7A apply to numerical outcome variables with mean responses replacing the rates. The principal difference in using a numerical outcome is that interest generally shifts to differences, such as the difference in means, instead of ratios, such as the relative risk. If the mean treatment difference $\tau = \alpha_1 - \alpha_0$ is constant over all levels of the confounding factor, direct standardization will yield-unbiased estimates of the treatment effect. As with the difference of rates, indirect standardization yields biased estimates.

Stratified estimators of the mean treatment difference have the form of a weighted combination of the difference of means within each of the confounding factor strata:

$$\frac{\sum_{k=1}^{K} v_k(\overline{Y}_{1k} - \overline{Y}_{0k})}{\sum_{k=1}^{K} v_k}.$$

To minimize the variance of the stratified estimator, the weights, the v_k, are chosen, where possible, inversely proportional to the variance of $\overline{Y}_{1k} - \overline{Y}_{0k}$. In the simplest case, the variance of each observation is constant in both risk factor groups and in all confounding factor strata. Then

$$v_k = \left(\frac{1}{n_{0k}} + \frac{1}{n_{1k}}\right)^{-1}. \qquad (7.1)$$

Kalton (1968) discusses the choice of the weights in detail, including the use of estimated variances. If the weights are constants, such as in (7.1), the stratified estimator will be unbiased. If the weights depend on the sample data, as by the use of estimated variances, the stratified estimators are biased, but the bias will become negligible in large samples.

7.8 EXTENSION TO MORE THAN ONE CONFOUNDING FACTOR

More than one categorical confounding factor can easily be handled by treating them together as one confounding factor. For example, two dichotomous confounding factors can be combined to form a single four-category confounding factor. However, as the number of confounding factors increases, the number of categories in the combined confounding factor can grow very quickly. This leads to the problem of small numbers in each category of the confounding factor and consequently, poorly determined specific rates or means.

There are really only two solutions to this problem. The first is to be selective in choosing confounding factors to adjust for. The second is to first apply the log-linear model technique (see Section 7.4.2). An extreme situation is that the number of categories in the combined confounding factor may be so large that some of the sample sizes on which the specific rates would be based are zero. Direct standardization cannot then be applied. Application of log-linear analysis eliminates the zeros.

Indirect standardization is frequently advocated because it is more precise than direct standardization, particularly in the presence of small numbers. This is true because indirect standardization does not use the specific rates, which will be poorly determined in small samples. As elaborated upon earlier, the general use of indirect standardization is not recommended. However, if there are too many confounding factor categories and many zeros, precluding the use of direct standardization, indirect standardization can still be applied and be better than the crude rates.

For purposes of stratification, Miettinen (1976) has proposed a method for reducing a set of confounding factors, whether numerical or discrete, to a single numerical confounding factor. The resulting confounding factor, "confounder score" in Miettinen's terminology, can then be categorized and the procedures of Section 7.1 or 7.6 applied. This method may be applied to either cohort or case-control studies, as long as both the outcome and risk factors are dichotomous.

The basis of Miettinen's proposal, as for discriminant matching (Section 6.10.4), is to use a discriminant function to distinguish (discriminate) between cases and noncases. This discriminant function depends on the value of the risk factor and the confounding factors. (This is stated for cohort studies; for case-control studies, distinguish instead between the risk-factor-present and risk-factor-absent groups. The discriminant function will then depend on the values of the outcome and the confounding factors.) The confounder score for each individual is obtained by evaluating the discriminant function for that person, assuming that the person is in the risk-factor-absent group, regardless of which risk factor group he or she is actually in. (For case-control studies, the function is evaluated assuming the person to be a control.) The motivation for this method is that the confounder score is a single variable that may be interpreted as a risk score that takes into account all variables except the risk factor.

7.9 HYPOTHESIS TESTING

In conjunction with the estimation problem, it is frequently desired to test the hypothesis that the risk factor has no effect on the outcome. Tests of $\theta = 1$ based on standardized relative risks can be done by using a standard normal distribution test. The necessary standard error formulas are given by Chiang (1961) and Keyfitz (1966).

Tests for the odds ratio related to stratified estimators are due to Mantel and Haenszel (1959) and Gart (1962), the latter being related to Woolf's estimator, and for the difference of rates due to Cochran (1954). The odds ratio procedures allow us to test whether the odds ratio is the same in all confounding factor strata (i.e., test whether the no interaction assumption is valid), and then whether the common value of the odds ratio differs from 1. Cochran's procedure does the same for the difference of rates, testing whether the common value of the difference differs from zero. Alternatives are likelihood ratio tests in log-linear analysis (Chapter 10). Much of this material is reviewed by Gart (1971), whose paper contains an extensive bibliography, and by Fleiss (1973, Chap. 10).

APPENDIX 7A MATHEMATICAL DETAILS OF STANDARDIZATION

At this point, the assumption of a dichotomous risk factor will be loosened to allow a general categorical risk factor. It will still be assumed that the response is dichotomous, so that the discussion will be in terms of rates, and that the data were obtained from a cohort study. No assumption is made regarding the choice of the standard population. It may, for example, correspond to one of the categories of the risk factor.

7A.1 Notation

Let R and K denote the number of categories in the risk factor and confounding factor, respectively. Lowercase letters, r and k, will be used as the corresponding indices. Let p_{rk} denote the observed rate based on n_{rk} individuals for the rth category of the risk factor and the kth category of the confounding factor. For the example in Table 7.1, there are $R = 2$ categories of the risk factor

marital status, $K = 6$ categories of the confounding factor age, and, for example, $n_{11} = 76.15 \times 10^5$ women in the group corresponding to the first category of marital status and the first category of age, and $p_{26} = 137.7 \times 10^{-5}$ is the breast cancer death rate for women in the second marital status category and sixth age category. The standard population has N_k individuals in the kth category of the confounding factor, with a corresponding rate of P_k. If an index is replaced by a dot, it indicates summation over that index. For example,

$$n_{r.} = \sum_{k=1}^{K} n_{rk}.$$

The crude rate for the rth category of the risk factor is then found as

$$p_r^{\mathrm{C}} = \frac{1}{n_{r.}} \sum_{k=1}^{K} n_{rk} p_{rk},$$

and for the standard population it is

$$P = \frac{1}{N.} \sum_{k=1}^{K} N_k P_k.$$

7A.2 Computation of Directly and Indirectly Standardized Rates

The directly standardized rate for the rth category of the risk factor is

$$p_r^{\mathrm{D}} = \frac{1}{N.} \sum_{k=1}^{K} N_k p_{rk}.$$

The standard mortality ratio (SMR) is

$$\mathrm{SMR}_r = \frac{\sum_{k=1}^{K} n_{rk} p_{rk}}{\sum_{k=1}^{K} n_{rk} P_k} = \frac{p_r^{\mathrm{C}}}{(1/n_{r.}) \sum_{k=1}^{K} n_{rk} P_k},$$

and the indirectly standardized rate is

$$p_r^{\mathrm{I}} = \mathrm{SMR}_r \times P.$$

We can see here why the roundabout calculation of indirectly standardized rates, beginning with the computation of the standard mortality ratio, is necessary. The straightforward analog of direct standardization would be to use

$$\frac{1}{n_{r.}} \sum_{k=1}^{K} n_{rk} P_k$$

as the indirectly standardized rate. This, however, is the rate for the standard population directly standardized to the confounding factor distribution in the rth category of the risk factor and so is not a rate that reflects the influence of the risk factor category. The standard mortality ratio acts as a correction factor to the standard population rate, P. The standard mortality ratio is the ratio of observed to expected deaths in the rth risk factor group, where the expectation is with respect to the standard population specific rates. The standard mortality ratio, then, indicates how much P should be changed to reflect the specific rates in the risk factor group.

7A.3 Bias of Indirect Standardization

The bias and consequent interpretability problems of indirect standardization are sufficiently important to be elaborated further. Estimation of the relative risk, θ, and difference of rates, Δ, are considered separately. In Section 7A.4, the one case where the bias of the indirectly standardized relative risk can be eliminated is given.

7A.3.1 Relative Risk. We will show that when the relative risk is constant and equal to θ within each category of the confounding factor category, direct standardization will be unbiased in large samples. On the other hand, the indirectly standardized relative risk can remain biased, no matter how large the samples.*

Consider a two-category risk factor—present ($r = 1$) and absent ($r = 0$)—and an arbitrary standard population. From Section 7A.2 the directly standardized relative risk is

$$\hat{\theta}^{\mathrm{D}} = \frac{p_1^{\mathrm{D}}}{p_0^{\mathrm{D}}} = \frac{\sum_{k=1}^{K} N_k p_{1k}}{\sum_{k=1}^{K} N_k p_{0k}}, \tag{7.2}$$

and the indirectly standardized relative risk is

$$\hat{\theta}^{\mathrm{I}} = \frac{p_1^{\mathrm{I}}}{p_0^{\mathrm{I}}} = \frac{\mathrm{SMR}_1}{\mathrm{SMR}_0} = \frac{\sum_{k=1}^{K} n_{1k} p_{1k} \Big/ \sum_{k=1}^{K} n_{1k} P_k}{\sum_{k=1}^{K} n_{0k} p_{0k} \Big/ \sum_{k=1}^{K} n_{0k} P_k}. \tag{7.3}$$

* To be precise, direct standardization yields a consistent estimate of θ; indirect standardization does not.

Suppose that the sample relative risks within each category of the confounding factor are all equal to θ, that is, $p_{1k} = \theta p_{0k}$ for all k. (This will be approximately the case in large samples within each stratum.) Then we have, from (7.2),

$$\hat{\theta}^{\mathrm{D}} = \frac{\sum_{k=1}^{K} N_k \theta p_{0k}}{\sum_{k=1}^{K} N_k p_{0k}} = \theta,$$

regardless of the choice of standard population. Direct standardization is doing the right thing by yielding the common value, θ, as the standardized relative risk. For the indirectly standardized relative risk, on the other hand, we have from (7.3):

$$\hat{\theta}^{\mathrm{I}} = \theta \left(\frac{\sum_{k=1}^{K} n_{1k} p_{0k} \Big/ \sum_{k=1}^{K} n_{1k} P_k}{\sum_{k=1}^{K} n_{0k} p_{0k} \Big/ \sum_{k=1}^{K} n_{0k} P_k} \right),$$

which is not, in general, equal to θ. If, instead, we take the standard population to be one of the risk factor groups, say $r = 0$, so that $P_k = p_{0k}$ for all k, we have

$$\hat{\theta}^{\mathrm{I}} = \frac{\sum_{k=1}^{K} n_{1k} p_{1k}}{\sum_{k=1}^{K} n_{1k} p_{0k}} = \frac{\sum_{k=1}^{K} n_{1k} p_{0k} \theta}{\sum_{k=1}^{K} n_{1k} p_{0k}} = \theta.$$

The result of this section, taken together with the result to be presented in Section 7A.4, says that there is only one case where indirect standardization does the right thing in terms of properly estimating the relative risk, but in that case the same answer can be obtained by direct standardization. From a bias point of view, there is thus no reason for choosing indirect standardization.

7A.3.2 Difference of Rates. Now consider direct and indirect standardization as estimators of the difference of rates. For direct standardization,

$$p_1^{\mathrm{D}} - p_0^{\mathrm{D}} = \frac{1}{N.} \sum_{k=1}^{K} N_k (p_{1k} - p_{0k}).$$

If the expected value of $p_{1k} - p_{0k}$ is some constant Δ for each category of the confounding factor, then the difference of the directly standardized rates is an unbiased estimator of Δ for any sample size.

For indirect standardization, the difference of standardized rates is

$$p_1^{\mathrm{I}} - p_0^{\mathrm{I}} = P(\mathrm{SMR}_1 - \mathrm{SMR}_0)$$

$$= P\left(\frac{\sum_{k=1}^{K} n_{1k}p_{1k}}{\sum_{k=1}^{K} n_{1k}P_k} - \frac{\sum_{k=1}^{K} n_{0k}p_{0k}}{\sum_{k=1}^{K} n_{0k}P_k}\right). \tag{7.4}$$

The expectation of this difference will be something other than Δ regardless of the sample size, except for one very special case.

Take the standard to be the risk-factor-absent group, as in Section 7A.3.1. Although the difference of indirectly standardized rates is still biased, the form of the estimator is informative. Substituting p_{0k} for P_k and p_0^{C} for P in (7.4), we obtain

$$p_1^{\mathrm{I}} - p_0^{\mathrm{I}} = p_0^{\mathrm{C}}\left[\frac{\sum_{k=1}^{K} n_{1k}(p_{1k} - p_{0k})}{\sum_{k=1}^{K} n_{1k}p_{0k}}\right].$$

Now suppose that, for all k, $p_{1k} - p_{0k} = \Delta$, as would be approximately the case in large samples. Then,

$$p_1^{\mathrm{I}} - p_0^{\mathrm{I}} = \Delta\left[\frac{(1/n_{0.})\sum_{k=1}^{K} n_{0k}p_{0k}}{(1/n_{1.})\sum_{k=1}^{K} n_{1k}p_{0k}}\right]. \tag{7.5}$$

This means that, in large samples, $p_1^{\mathrm{I}} - p_0^{\mathrm{I}}$ will be biased unless $\Delta = 0$, and that the greater the confounding, the greater the bias. [The term in brackets in (7.5) can be viewed as a measure of the extent to which the confounding factor distribution in the two risk factor groups differ.]

7A.4 Equivalence of Direct and Indirect Standardization

Whether or not the relative risk is constant within each category of the confounding factor, as was assumed in Section 7A.3, direct and indirect standardization are equivalent if the standard population is taken to be one of the risk factor groups. First, equivalent means that, by choosing the standard appropriately for each type of standardization, the two methods will yield the same estimate of the relative risk.

Suppose that the risk-factor-present group ($r = 1$) is chosen as the standard for direct standardization. Then,

$$N_k = n_{1k} \qquad \text{for all } k,$$

and consequently,

$$p_1^{\mathrm{D}} = p_1^{\mathrm{C}}.$$

For the risk-factor-absent group,

$$p_0^{\mathrm{D}} = \frac{1}{n_{1.}}\sum_{k=1}^{K} n_{1k}p_{0k},$$

and therefore

$$\hat{\theta}^{\mathrm{D}} = \frac{\sum_{k=1}^{K} n_{1k}p_{1k}}{\sum_{k=1}^{K} n_{1k}p_{0k}}.$$

If the other risk factor group, the risk-factor-absent group ($r = 0$), is chosen as the standard for indirect standardization, then

$$P_k = p_{0k} \qquad \text{for all } k$$

and

$$p_0^{\mathrm{I}} = p_0^{\mathrm{C}}.$$

For the risk-factor-present group,

$$p_1^{\mathrm{I}} = p_0^{\mathrm{C}}\frac{\sum_{k=1}^{K} n_{1k}p_{1k}}{\sum_{k=1}^{K} n_{1k}p_{0k}}$$

and therefore

$$\hat{\theta}^{I} = \frac{\sum_{k=1}^{K} n_{1k}p_{1k}}{\sum_{k=1}^{K} n_{1k}p_{0k}} = \hat{\theta}^{\mathrm{D}}.$$

Note, however, that neither the two sets of standardized rates nor their differences are equal; that is, $p_0^{\mathrm{I}} \neq p_0^{\mathrm{D}}$, $p_1^{\mathrm{I}} \neq p_1^{\mathrm{D}}$, and $p_1^{\mathrm{I}} - p_0^{\mathrm{I}} \neq p_1^{\mathrm{D}} - p_0^{\mathrm{D}}$.

APPENDIX 7B STRATIFIED ESTIMATORS OF THE ODDS RATIO

In this appendix various mathematical details regarding the five principal estimators—maximum likelihood, conditional maximum likelihood, Woolf, modified Woolf, and Mantel-Haenszel—of the odds ratio will be given. The estimators due to Birch and Goodman will not be considered, since they are not approximately unbiased in large samples.

The notation for sample quantities is given in Section 7A.1. In addition, the population rate for the kth confounding factor category and rth risk factor category is denoted P_{rk} ($r = 0, 1$ and $k = 1, \ldots, K$). The no-interaction assumption is that the odds ratio is the same in each confounding factor stratum:

$$\psi = \frac{P_{1k}Q_{0k}}{Q_{1k}P_{0k}} \qquad \text{for } k = 1, \cdots, K, \tag{7.6}$$

where $Q_{rk} = 1 - P_{rk}$.

7B.1 Maximum Likelihood and Conditional Likelihood Estimators

The likelihood (ignoring the binomial coefficients) is

$$L = \prod_{k=1}^{K} P_{1k}^{s_{1k}} Q_{1k}^{n_{1k}-s_{1k}} P_{0k}^{s_{0k}} Q_{0k}^{n_{0k}-s_{0k}},$$

where the s_{rk} are the numbers of "successes" ($p_{rk} = s_{rk}/n_{rk}$). In this form there appear to be $2K$ parameters to estimate, the P_{rk}, but actually there are only $K + 1$ independent parameters, owing to the no-interaction assumption (7.6). To reparametrize, let γ denote the natural log of the odds ratio ψ and let

$$\rho_k = \ln\left(\frac{P_{1k}}{Q_{1k}}\right) \qquad \text{for } k = 1, \cdots, K.$$

The natural log of the likelihood is then

$$l = \sum_{k=1}^{K} [s_{1k}\rho_k + n_{1k} \ln Q_{1k} + s_{0k}(\rho_k - \gamma) + n_{0k} \ln Q_{0k}], \tag{7.7}$$

where the Q_{rk} are functions of γ and the ρ_k.

The (usual) maximum likelihood estimator of γ is found by differentiating (7.7) with respect to γ and the ρ_k and then solving for the $K + 1$ unknowns. If the sample sizes within each stratum, the n_{rk}, are all large, the maximum likelihood estimator of γ, $\hat{\gamma}_{\text{ML}}$, will be approximately normally distributed with mean γ and variance W^{-1} (Gart, 1962), where

$$W = \sum_{k=1}^{K} [(n_{0k}P_{0k}Q_{0k})^{-1} + (n_{1k}P_{1k}Q_{1k})^{-1}]^{-1}. \tag{7.8}$$

The maximum likelihood estimator of the odds ratio ψ is

$$\hat{\psi}_{\text{ML}} = \exp(\hat{\gamma}_{\text{ML}}),$$

which will be approximately normally distributed with mean ψ and variance ψ^2/W. The variance of $\hat{\psi}_{\text{ML}}$ can be estimated by replacing each P_{rk} in (7.8) with the corresponding p_{rk}.

This maximum likelihood estimator is identical to that obtained by logit analysis (Chapter 9), using the method of Section 9.8 to handle a confounding factor with more than two categories.

In the alternative asymptotic case, where the number of categories, K, increases, the maximum likelihood estimator given above is not appropriate; as the number of categories increases, the number of parameters also increases, violating one of the assumptions required for the properties of maximum likelihood estimators to hold. In such cases, an alternative maximum likelihood estimator, the conditional maximum likelihood estimator, is appropriate. The term "conditional" comes from the fact that this procedure is based on the likelihood of ψ conditioned on the sufficient statistics for the K nuisance parameters, the ρ_k. This likelihood is (Gart, 1970)

$$L' = \prod_{k=1}^{K} \frac{\binom{n_{1k}}{s_{1k}}\binom{n_{0k}}{t_k - s_{1k}}\psi^{s_{1k}}}{\sum_{j=\max(0,t_k-n_{0k})}^{\min(t_k,n_{1k})} \binom{n_{1k}}{j}\binom{n_{0k}}{t_k - j}\psi^{j}}, \tag{7.9}$$

where $t_k = s_{1k} + s_{0k}$ is the sufficient statistic for ρ_k. The conditional maximum likelihood estimator of ψ, $\hat{\psi}_{\text{CML}}$, is that value of ψ which maximizes L'. Thomas (1975) considers the numerical problem of solving for the maximizing value.

Andersen (1970) considers the properties of conditional maximum likelihood estimators in general. Applying his results to this problem, we obtain that $\hat{\psi}_{\text{CML}}$ will be approximately normally distributed with mean ψ when K is large. The variance formula is not illuminating; the variance can be estimated as the reciprocal of the second derivative of $-L'$.

Birch (1964) showed that $\hat{\psi}_{\text{CML}}$ is the solution of a polynomial equation that involves an expectation taken with respect to the conditional distributions in (7.9). Gart's (1970) approximation to the conditional maximum likelihood estimator is based on approximating this expected value by using a large sample (large n_{rk}) approximation. This estimator, $\hat{\psi}_{\text{AML}}$ (A for asymptotic), is the solution to

$$s_1 = \sum_{k=1}^{K} \hat{s}_k$$

where $s_1 = \sum^K{}_{k=1}\, s_{1k}$ and each $\hat{s}_k$ satisfies

$$\frac{\hat{s}_k(n_{0k} - t_k + \hat{s}_k)}{(t_k - \hat{s}_k)(n_{1k} - \hat{s}_k)} = \hat{\psi}_{\text{AML}}.$$

Again, Thomas (1975) considers the numerical problems of solving for $\hat{\psi}_{\text{AML}}$.

Gart's approximation to the conditional maximum likelihood estimator does require large n_{rk} to be valid, but, unlike the approximations due to Birch and Goodman, is valid for all value of the odds ratio, not just $\psi = 1$.

7B.2 Woolf and Modified Woolf Estimators

The estimator of the log odds ratio proposed by Woolf (1955) is a weighted average of the estimators of the log odds ratio from each of the strata:

$$\hat{\gamma}_w = \frac{\sum_{k=1}^{K} w_k \hat{\gamma}_k}{w}$$

where

$$\hat{\gamma}_k = \ln\left(\frac{p_{1k}q_{0k}}{q_{1k}p_{0k}}\right) = \ln\left[\frac{s_{1k}(n_{0k} - s_{0k})}{(n_{1k} - s_{1k})s_{0k}}\right] \qquad (7.10)$$

$$w_k = [(n_{0k}p_{0k}q_{0k})^{-1} + (n_{1k}p_{1k}q_{1k})^{-1}]^{-1}$$

$$= \left[\frac{n_{0k}}{s_{0k}(n_{0k} - s_{0k})} + \frac{n_{1k}}{s_{1k}(n_{1k} - s_{1k})}\right]^{-1}$$

$$w = \sum_{k=1}^{K} w_k. \qquad (7.11)$$

w_k^{-1} is an estimate of the variance of $\hat{\gamma}_k$, so the Woolf estimator is based on weighting inversely proportional to the variance.

From (7.10) it is clear that the Woolf estimator cannot be calculated if any p_{rk} or q_{rk} is zero. A modification that avoids this problem is based on work of Haldane (1955) and Anscombe (1956). They independently showed that a less biased estimator of the log odds ratio from the kth confounding factor stratum is

$$\hat{\gamma}'_k = \ln\left[\frac{(s_{1k} + 0.5)(n_{0k} - s_{0k} + 0.5)}{(n_{1k} - s_{1k} + 0.5)(s_{0k} + 0.5)}\right]; \qquad (7.12)$$

that is, just add 0.5 to each of the four quantities in the sample odds ratio formula. Gart (1966) suggested a modified Woolf estimator of the form

$$\hat{\gamma}_{\text{MW}} = \frac{\sum_{k=1}^{K} w'_k \hat{\gamma}'_k}{w'},$$

where

$$w'_k = \left[\frac{n_{0k} + 1}{(s_{0k} + 0.5)(n_{0k} - s_{0k} + 0.5)} + \frac{n_{1k} + 1}{(s_{1k} + 0.5)(n_{1k} - s_{1k} + 0.5)}\right]^{-1}$$

$$w' = \sum_{k=1}^{K} w'_k. \qquad (7.13)$$

The results of Gart and Zweifel (1967), who considered various estimators of the log odds and estimators of the variances of the log odds estimators, suggest that $(w'_k)^{-1}$ is generally the least biased estimator of the variance of $\hat{\gamma}'_k$. [An alternative modification to the weights, not considered here, was suggested by Haldane (1955).]

For the asymptotic case of large n_{rk}, the Woolf and modified Woolf estimators have the same large-sample distribution as the maximum likelihood estimator. In particular, the asymptotic variances are equal, so the two Woolf estimators, as well as the maximum likelihood estimator, are asymptotically efficient. Estimated variances for the Woolf and modified Woolf estimators are w^{-1} (7.11) and $(w')^{-1}$ (7.13), respectively.

7B.3 Mantel–Haenszel Estimator

Mantel and Haenszel (1959) proposed the estimator

$$\hat{\psi}_{\text{MH}} = \sum_{k=1}^{K} \frac{s_{1k}(n_{0k} - s_{0k})}{n_{1k} + n_{0k}} \Big/ \sum_{k=1}^{K} \frac{(n_{1k} - s_{1k})s_{0k}}{n_{1k} + n_{0k}}$$

$$= \sum_{k=1}^{K} \frac{m_k \hat{\psi}_k}{m},$$

where

$$m_k = \left(\frac{1}{n_{1k}} + \frac{1}{n_{0k}}\right)^{-1} q_{1k}p_{0k}$$

$$m = \sum_{k=1}^{K} m_k.$$

Hauck (1979) has shown that if the n_{rk} are large, the Mantel–Haenszel estimator is approximately normally distributed with mean ψ and variance

$$V = \frac{\psi^2 \sum_{k=1}^{K} M_k^2 W_k^{-1}}{M^2}, \qquad (7.14)$$

where

$$M_k = \left(\frac{1}{n_{1k}} + \frac{1}{n_{0k}}\right) Q_{1k}P_{0k}$$

$$M = \sum_{k=1}^{K} M_k$$

$$W_k^{-1} = (n_{0k}P_{0k}Q_{0k})^{-1} + (n_{1k}P_{1k}Q_{1k})^{-1}.$$

A sufficient condition for the variance of the Mantel–Haenszel estimator (7.14) to be equal to that of the maximum likelihood estimator (ψ^2/W) is $\psi = 1$.

REFERENCES

Andersen, E. B. (1970), Asymptotic Properties of Conditional Maximum Likelihood Estimators, *Journal of the Royal Statistical Society, Series B,* **32,** 283–301.

Anscombe, F. J. (1956), On Estimating Binomial Response Relations, *Biometrika,* **43,** 461–464.

Birch, M. M. (1964), The Detection of Partial Association I: The 2 × 2 Case, *Journal of the Royal Statistical Society, Series B,* **26,** 313–324.

Bishop, Y. M. M. (1967), Multidimensional Contingency Tables: Cell Estimates, Ph.D. thesis, Harvard University.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1974), *Discrete Multivariate Analysis: Theory and Practice,* Cambridge, MA: MIT Press.

Chiang, C. L. (1961), Standard Error of the Age-Adjusted Death Rate, U.S. Dept. Health, Education and Welfare, Vital Statistics, Special Report 47, pp. 271–285.

Cochran, W. G. (1954), Some Methods for Strengthening the Common χ^2 Tests, *Biometrics,* **10,** 417–451.

Cochran, W. G. (1968), The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies, *Biometrics,* **24,** 295–313.

Fleiss, J. L. (1973), *Statistical Methods for Rates and Proportions,* New York: Wiley.

Gart, J. J. (1962), On the Combination of Relative Risks, *Biometrics,* **18,** 601–610.

Gart, J. J. (1966), Alternative Analyses of Contingency Tables, *Journal of the Royal Statistical Society, Series B,* **28,** 164–179.

Gart, J. J. (1970), Point and Interval Estimation of the Common Odds Ratio in the Combination of 2 × 2 Tables with Fixed Marginals, *Biometrika,* **57,** 471–475.

Gart, J. J. (1971), The Comparison of Proportions: A Review of Significance Tests, Confidence Intervals and Adjustments for Stratification, *Review of the International Statistical Institute,* **39,** 148–169.

Gart, J. J., and Zweifel, R. (1967), On the Bias of Various Estimators of the Logit and Its Variance with Application to Quantal Bioassay, *Biometrika,* **54,** 181–187.

Goldman, A. (1971), The Comparison of Multidimensional Rate Tables—A Simulation Study, Ph.D. thesis, Harvard University.

Goodman, L. A. (1969), On Partitioning χ^2 and Detecting Partial Association in Three-Way Tables, *Journal of the Royal Statistical Society, Series B,* **31,** 486–498.

Haldane, J. B. S. (1955), The Estimation and Significance of the Logarithm of a Ratio of Frequencies, *Annals of Human Genetics,* **20,** 309–311.

Hauck, W. W. (1979), The Large Sample Variance of the Mantel–Haenszel Estimator of a Common Odds Ratio, *Biometrics,* **35,** 817–819.

Herring, R. A. (1936), The Relationship of Martial Status in Females to Mortality from Cancer of the Breast, Female Genital Organs and Other Sites, *The American Society for the Control of Cancer,* **18,** 4–8.

Kalton, G. (1968), Standardization: A Technique to Control for Extraneous Variables, *Applied Statistics,* **17,** 118–136.

Keyfitz, N. (1966), Sampling Variance of Standardized Mortality Rates, *Human Biology,* **3,** 309–317.

Kitagawa, E. M. (1964), Standardized Comparisons in Population Research, *Demography,* **1,** 296–315.

Kitagawa, E. M. (1966), Theoretical Considerations in the Selection of a Mortality Index, and Some Empirical Comparisons, *Human Biology,* **38,** 293–308.

McKinlay, S. M. (1975), The Effect of Bias on Estimators of Relative Risk for Pair Matched and Stratified Samples, *Journal of the American Statistical Association,* **70,** 859–864.

McKinlay, S. M. (1978), The Effect of Nonzero Second-Order Interaction on Combined Estimators of the Odds Ratio, *Biometrika,* **65,** 191–202.

Mantel, N., and Haenszel, W. (1959), Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, *Journal of the National Cancer Institute,* **22,** 719–748.

Miettinen, O. S. (1972), Standardization of Risk Ratios, *American Journal of Epidemiology,* **96.** 383–388.

Miettinen, O. S. (1976), Stratification by a Multivariate Confounder Score, *American Journal of Epidemiology,* **104,** 609–620.

Spiegelman, M., and Marks, H. H. (1966), Empirical Testing of Standards for the Age Adjustment of Death Rates by the Direct Method, *Human Biology,* **38,** 280–292.

Thomas, D. G. (1975), Exact and Asymptotic Methods for the Combination of 2 × 2 Tables, *Computers and Biomedical Research,* **8,** 423–446.

Woolf, B. (1955), On Estimating the Relation between Blood Group and Disease, *Annals of Human Genetics,* **19,** 251–253.

CHAPTER 8

Analysis of Covariance

8.1 Background 140
8.2 Example: Nutrition Study Comparing Urban and Rural Children 141
8.3 The General ANCOVA Model and Method 144
8.4 Assumptions Underlying the Use of ANCOVA 148
8.5 Dealing with Departures from the Assumptions 151
8.5.1 Nonparallel Regressions 152
8.5.2 Nonlinear Regressions 154
8.5.3 Measurement Error 156
8.5.4 Multiple Confounding Variables 156
Appendix 8A Formulas for Analysis-of-Covariance Calculations 157
References 159

In this chapter we consider an adjustment strategy that is appropriate for cohort studies with a numerical outcome factor, a categorical treatment (or risk) factor, and a numerical confounding factor. Under these conditions, the *general linear model* can be applied to the problem of estimating treatment effects. The *analysis of covariance* (ANCOVA) represents the main application of the linear model for this purpose.

8.1 BACKGROUND

The general linear model represents the outcome value as a linear combination (weighted sum) of measured variables. Generally speaking, when these variables are all numerical, the linear model is called a *regression model.* When the variables are all categorical, we refer to the *analysis of variance* (ANOVA).

While both regression and analysis of variance can be formally subsumed under the general linear model, the two techniques have traditionally been treated as distinct. This historical separation occurred for two reasons. First, before high-speed computers were in general use, computational aspects of statistical techniques were of much interest. The most efficient computational procedures for regression and ANOVA were quite different. Second, the two methods tended to be applied to different sorts of problems.

The analysis of variance is usually thought of as a technique for comparing the means of two or more populations on the basis of samples from each. In practice, these populations often correspond to different treatment groups, so that differences in population means may be evidence for corresponding differences in treatment effects.

The ANOVA calculations involve a division of the total sample variance into within-group and between-group components. The within-group component provides an estimate of error variance, while the between-group component estimates error variance plus a function of the differences among treatment means. The ratio of between- to within-group variance provides a test of the null hypothesis that all means are equal. Moreover, the differences among group means provide unbiased estimates of the corresponding population mean differences, and standard errors based on the within-group variance provide confidence intervals for these differences and tests of their significance.

Regression analysis, on the other hand, is primarily used to model relationships between variables. With it, we can estimate the form of a relationship between a response variable and a number of inputs. We can try to find that combination of variables which is most strongly related to the variation in the response.

The analysis of covariance represents a marriage of these two techniques. Its first use in the literature was by R. A. Fisher (1932), who viewed the technique as one that "combines the advantages and reconciles the requirements of the two very widely applicable procedures known as regression and analysis of variance."

Combining regression and ANOVA provides the powerful advantage of making possible comparisons among treatment groups differing prior to treatment. Suppose we can identify a variable X that is related to the outcome, Y, and on which treatment groups have different means. We shall assume for simplicity that X is the only variable on which the groups differ. Then, if we knew the relationship between Y and X, we could appropriately adjust the observed differences on Y to take account of the differences on X.

8.2 EXAMPLE: NUTRITION STUDY COMPARING URBAN AND RURAL CHILDREN

Greenberg (1953) described a nutrition study designed to compare growth of children in an urban environment with that of rural children. Data were ob-