*This investigation was supported in part by an NIH Training Grant Award (No. 5-T01-ES00133) and by an NIH Research Career Development Award (No. 1-K04-ES00003), both from the National Institute of Environmental Health Sciences.

MATCHING ON A CONTINUOUS VARIABLE IN CASE-CONTROL STUDIES*

by

W.J. Raynor
Department of Preventive Medicine
Rush Presbyterian St. Luke's Medical Center, Chicago

L.L. Kupper
Department of Biostatistics
University of North Carolina, Chapel Hill

Institute of Statistics Mimeo Series No. 1127
June 1977
ABSTRACT

This research examines and compares three methods that can be used to match on a continuous variable in case-control studies. Each method is examined to determine the effects of the matching on the distribution of the matched sample, to assess the associated sample size requirements (if any), and to compare a simple mean-difference estimator using the matched samples with an analysis of covariance estimator using random samples.

Chapter I presents a discussion of the case-control study, and a review of the statistical and epidemiologic literature. We study categorical-matching, caliper-matching, and nearest-neighbor matching in Chapters II, III, and IV, respectively. For each method, we first consider the effect of the matching process on the post-matching distribution of the matching variable. We examine the percent reduction in mean-difference between the case and control group matching variable due to the matching method, the post-matching variance of the matching variable in the control group, and, for caliper-matching and nearest-neighbor matching, the post-matching correlation between the paired matching variable values.

The choice of a category structure or a caliper size affects the value of the minimum control pool size needed to match all the cases. The distribution of the least control pool size needed to match all
cases is studied for these two methods. Under the assumption that a simple linear model relates the matching variable and the response variable of interest, a simple estimator of the case-control difference is the mean-difference estimator. For each of the three methods, this estimator is compared with the linear model estimator, based on random samples, using relative efficiency and relative mean square error as criteria for comparison.

Two simple estimators that are commonly cited in discussions on matching are the exactly pair-matched mean-difference estimator and the randomly matched or unmatched mean-difference estimator. We show that these two estimators can be considered to be "limiting" cases of the mean-difference estimators based on the three matching methods.

Finally, at the end of Chapter IV, we compare the three matching methods. Since there are different control pool sample size requirements associated with caliper-matching and category-matching, the two methods are compared at points where both methods were associated with the same expected least control pool sample size. This sample size is also used for nearest-neighbor matching. The comparisons are based on the mean square error of the mean-difference estimator for each method. In Chapter V, we offer some suggestions for future research.
TABLE OF CONTENTS

Page

LIST OF TABLES ........................................ vi
LIST OF FIGURES ........................................ x

CHAPTER

I. INTRODUCTION AND REVIEW OF LITERATURE ............... 1
   1.1. Introduction ........................................ 1
   1.2. Observational Studies ............................... 2
   1.3. Confounding .......................................... 3
   1.4. Effect Modification, Synergism and Antagonism ...... 13
   1.5. Matching ............................................. 14
   1.6. Stratification ........................................ 17
   1.7. Literature Review .................................... 19
   1.8. Summary .............................................. 38

II. CATEGORICAL MATCHING OF CONTINUOUS VARIABLE ....... 40
   2.1. Introduction .......................................... 40
   2.2. The Technique of Categorical Pair-Matching ........ 41
   2.3. Category Matching for Normally Distributed Matching
        Variables ............................................ 47
   2.4. Sample Size Considerations .......................... 58
   2.5. Category Matching and Covariance Analysis ........ 70

III. CALIPER MATCHING ..................................... 108
   3.1. Introduction .......................................... 108
   3.2. The Technique of Caliper Matching and Post-Matching
        Properties of Caliper Matched Variables ............. 110
   3.3. Caliper Matching for Normally Distributed Matching
        Variables ............................................. 116
   3.4. Sample Size Considerations for Caliper Matching .... 139
   3.5. Covariance Analysis and Caliper Matching .......... 145
TABLE OF CONTENTS continued...

IV. NEAREST-NEIGHBOR MATCHING AND COMPARISON OF THE THREE METHODS ........................................ 170
   4.1. Introduction .................................................. 170
   4.2. Nearest-Neighbor Pair-Matching for Normally Distributed Variables ........................................ 173
   4.3. Covariance Analysis and Nearest-Neighbor Pair-Matching ......................................................... 180
   4.4. Mean Square Error Comparisons of the Three Matching Methods ............................................... 195

V. SUMMARY AND PROBLEMS FOR FUTURE RESEARCH ................................................................. 203
   5.1. Summary .......................................................... 203
   5.2. Multiple Matching Variables ........................................... 205
   5.3. Tests for Choosing Matching Variables ........................................................................ 205
   5.4. Other Areas of Interest .................................................. 210

APPENDIX ...................................................................... 212

REFERENCES .................................................................. 228
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7.1</td>
<td>Values of $(1-R^2)$</td>
</tr>
<tr>
<td>1.7.2</td>
<td>Relative Efficiency of Matching and Covariance Analysis</td>
</tr>
<tr>
<td>1.7.3</td>
<td>Results of a Monte Carlo Efficiency Investigation for a Continuous Response</td>
</tr>
<tr>
<td>1.7.4</td>
<td>Probability that the 95% Limits Include $\mu-\mu'$</td>
</tr>
<tr>
<td>1.7.5</td>
<td>Remaining Bias in $\bar{y}-\bar{y}'$ and $\sum_{j} w_j (\bar{y}_j-\bar{y}_j')$</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Minimum $\sigma^2_1$. Intervals, Symmetric about Zero, for $X_1 \sim N(0,1)$</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Results of Matching a $N(\mu_{x_0}, \sigma^2)$ Variable by an Interval Structure that is Equiprobable for a $N(0,1)$ Distribution</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Probability that a Randomly Selected Observation from $X_0 \sim N(\mu_{x_0}, 1)$ will Fall into Category $C_1$, the First of the Intervals that are Equiprobable in a $N(0,1)$ Population</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Expected Values of the Least Sample Size for $r=3$ Categories when $n_2=n_3=m$ and $p_2=p_3=1/3$</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Mean and Standard Deviation of the Least Sample Size of $N(\mu_{x_0},1)$ Controls Necessary to Category Match $M=12$ $N(0,1)$ Cases, Using $r=3,4$ $N(0,1)$ Equiprobable Categories</td>
</tr>
<tr>
<td>2.4.4</td>
<td>90% Bound on the Number of Random Samples from a $N(\mu_{x_0},1)$ Control Distribution Necessary to Category Match $M=60$ $N(0,1)$ Case Observations in a $N(0,1)$ Equiprobable Category Matching Scheme</td>
</tr>
</tbody>
</table>
LIST OF TABLES continued...

TABLE

2.5.1 Monte Carlo Relative Efficiencies, Based on (2.5.19.), at Covariance Analysis and Category Pair-Matching (3 categories) X₁ ~ N(0,1) .................................................. 89

2.5.2 Population Relative Efficiency of Category Pair-Matching to Covariance Analysis for Three Equiprobable N(0,1) Intervals .................................................. 90

2.5.3 Relative Efficiency of Paired-Matching to Unpaired-Matching .................................................. 91

2.5.4 Relative Efficiency of Unpaired Category Matching to Covariance Analysis .................................................. 92

2.5.5 Relative Efficiency of Category Matching to Covariance Analysis Adjusted for the Degrees of Freedom for Error .................................................. 100

2.5.6 MSE(ψ)/MSE(δ) Adjusted for d.f. for Error .................................................. 104

2.5.7 Values of \(\left(\frac{\sigma_{x_1y_1}^2}{(1-\rho_{x_1y_1}^2)}\right)^{1/2}\) .................................................. 106

3.3.1 Percent Reduction in Bias for Caliper Matching .................................................. 131

3.3.2 Proportional Reduction in the Mean Difference Between X₁ and X₀ Due to Caliper Matching .................................................. 133

3.3.3 Post-Caliper Matching Variance of X₀ .................................................. 134

3.3.4 The Post-Caliper Matching Correlation Between X₁ and X₀ .................................................. 137

3.4.1 Mean Ratio of the Least Control Pool Size for the Case Sample Size Required to Caliper Match All Cases .................................................. 142

3.4.2 Proportional Reduction in Mean Difference Due to Caliper Matching ... Divided by the Mean Ratio of the Least Control Pool Size to the Case Sample Size .................................................. 144

3.5.1 Some Values of the Correction Factor \(\frac{n^2}{(n+2)(n-1)}\) .................................................. 153

3.5.2 Relative Efficiency of the Caliper Pair-Matching Estimator to the Covariance Estimator for the Caliper c=0.2\sqrt{\frac{\sigma_{x_0}^2}{2}} .................................................. 155
<table>
<thead>
<tr>
<th>TABLE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.3</td>
<td>Relative Efficiency of the Caliper Pair-Matched Estimator to the Covariance Estimator for the Caliper $c=0.6\sqrt{(1+\sigma_x^2)/2}$</td>
<td>157</td>
</tr>
<tr>
<td>3.5.4</td>
<td>Relative Efficiency of the Caliper Pair-Matched Estimator to the Covariance Estimator for the Caliper $c=1.0\sqrt{(1+\sigma_x^2)/2}$</td>
<td>159</td>
</tr>
<tr>
<td>3.5.5</td>
<td>Relative Mean Square Error of the Covariance Analysis Estimator to the Caliper Pair-Matched Estimator $c=0.2\sqrt{(1+\sigma_x^2)/2}$</td>
<td>163</td>
</tr>
<tr>
<td>3.5.6</td>
<td>Relative Mean Square Error of the Covariance Analysis Estimator to the Caliper Pair-Matched Estimator $c=0.6\sqrt{(1+\sigma_x^2)/2}$</td>
<td>165</td>
</tr>
<tr>
<td>3.5.7</td>
<td>Relative Mean Square Error of the Covariance Analysis Estimator to the Caliper Pair-Matched Estimator $c=1.0\sqrt{(1+\sigma_x^2)/2}$</td>
<td>167</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Percent Reduction in Mean Difference for Random Order Nearest Available Matching</td>
<td>175</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Proportional Reduction in Mean Difference Due to Nearest-Neighbor Matching of n Randomly Ordered Cases from a Control Pool of Size $n_0=\text{rn}$</td>
<td>176</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Variance of n Control Values Chosen from a Pool of Size $n_0=\text{rn}$</td>
<td>178</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Correlation of n Matched Values when Controls are Selected from a Pool of Size $n_0=\text{rn}$ by Random Order Nearest-Neighbor Matching</td>
<td>179</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Relative Efficiency of Random Case Ordering Nearest-Neighbor Pair-Matched Mean Difference Estimator to the Covariance Analysis Estimator, $\sigma^2_{x_0}=1$</td>
<td>184</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Relative Efficiency of Random Case Ordering Nearest-Neighbor Pair-Matched Mean Difference Estimator to the Covariance Analysis Estimator, $\sigma^2_{x_0}=3$</td>
<td>186</td>
</tr>
</tbody>
</table>
LIST OF TABLES continued...

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.3 Relative Mean Square Error of the Covariance Analysis Estimator to the Random Case Ordering Nearest-Neighbor Pair-Matched Mean Difference Estimator $\sigma^2_{x_0} = 1$</td>
<td>191</td>
</tr>
<tr>
<td>4.3.4 Relative Mean Square Error of the Covariance Analysis Estimator to the Random Case Ordering Nearest-Neighbor Pair-Matched Mean Difference Estimator $\sigma^2_{x_0} = 3$</td>
<td>193</td>
</tr>
<tr>
<td>4.4.1 Mean Square Error Function Comparisons of r Category Matching, Caliper Matching, and Nearest-Neighbor Matching, r=3</td>
<td>200</td>
</tr>
<tr>
<td>4.4.2 Mean Square Error Function Comparisons of r Category Matching, Caliper Matching, and Nearest-Neighbor Matching, r=4</td>
<td>201</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3.1</td>
<td>8</td>
</tr>
<tr>
<td>1.3.2</td>
<td>9</td>
</tr>
<tr>
<td>1.3.3</td>
<td>12</td>
</tr>
<tr>
<td>3.3.1</td>
<td>136</td>
</tr>
</tbody>
</table>

3.3.1 Post-Matching Variance as a Function of $c = a \sqrt{1 + \frac{2}{x_0}}$
CHAPTER I
INTRODUCTION AND REVIEW OF LITERATURE

1.1. Introduction

A problem of interest to statisticians and epidemiologists alike concerns how best to choose a control sample in a case-control study, particularly when there are continuous variables to be considered in the choice of the controls. Of particular interest is the gain obtained by matched sampling on these continuous variables, as opposed to simple random sampling of the control subjects. We will consider several aspects of this problem. First, we will examine the effects of three matching methods on the post-matching distribution of a continuous variable, and, compare an estimator of the case-control difference based on a matched sample with another estimator based on a random sample with regard to the efficiency and mean square error of the estimators. Secondly, for two of these methods we examine the distribution of the minimum number of randomly sampled controls necessary to completely match all the cases for a given matching scheme. Finally, we will compare the three matching methods on the basis of the mean square error of their estimators of the case-control difference. These comparisons will be done in such a manner as to equalize the effects of the required control pool size for the three methods.

This chapter presents a general discussion of case-control studies and a review of the literature. The discussion considers the case-control study and the terminology that has grown around it. The
literature review considers the current statistical and epidemiologic literature pertinent to the areas mentioned above.

1.2. **Observational Studies**

The purpose of this chapter is to discuss various aspects of the design and analysis of case-control studies, with special emphasis on clarifying, from a statistical standpoint, the terminology which has developed, and to review the relevant statistical literature which is concerned with matching in case-control studies. The data for such studies generally is "observational" in nature; by observational, we mean, quoting McKinlay (1975), "that type of study which is concerned with investigating relationships among characteristics of human populations, after the manner of an experiment, but comparing groups among which 'treatments' are not randomly assigned."

More specifically, we will be concerned with case-control studies where the dependent variable (which we will alternatively refer to as the response or exposure variable and will denote as Y) can reasonably be treated as a continuous random variable. It will be of primary interest to quantify how Y differs between the case and control groups, after controlling or adjusting for variables $X_1, X_2, \ldots, X_p$, which are variously (although not always appropriately) called "confounders", "concomitants", or "covariates". The necessity for having to worry about the X's (which are generally not of primary interest) stems from the fact that they may be related both to Y and to the case-control group designation (as well as to each other), so that any apparent differences in Y between the case and control groups may be due to either in part or totally to the X's themselves. Accounting for the effects of the X's is especially important in observational studies, since the statistical technique of
randomization cannot be utilized to help negate the influence of such nuisance variables, whether they are measured directly or not.

For example, consider a collection of people with lung cancer to be the set of cases; further, suppose that information is gathered on the response variable of interest (e.g., some quantitative measure of smoking history) and on various covariables (e.g., age, race, sex, and occupation) for these individuals. A group of controls suitable for comparison purposes is then selected. The problem is to study differences in smoking history between these two groups in the presence of possible age, race, sex or occupation differences.

1.3. Confounding

The term "confounding" generally has a different meaning to statisticians than it does to epidemiologists. Statisticians usually deal with the notions of "complete" and "partial" confounding. Two factors are completely confounded when the effects of the two factors are inextricably mixed and cannot be separated at all. An example of such complete statistical confounding in a factorial experiment would be where each distinct level of a factor A is associated with one and only one distinct level of a factor B, in which case the effect of factor A cannot be unentangled at all from that of factor B. Statisticians often deliberately introduce the possibility of complete confounding into an experiment either to satisfy sample size constraints or for logistical reasons. For instance, the use of a Latin Square experimental design automatically completely confounds interactive effects with experimental error, the justification being that the interactions among the particular factors under consideration can reasonably be assumed to be negligible. Partial
confounding enters, for example, into the construction of certain fractional factorial designs.

On the other hand, confounding in an epidemiological sense generally occurs not by design but by chance, because epidemiologic studies are usually observational in nature. MacMahon and Pugh (1970, p. 244) have defined confounding variables as "variables which may introduce differences between cases and controls which do not reflect differences in the variables of primary interest." Miettinen (1974) has offered the following definition: "In practice, a confounding factor for a given hypothesized causal connection between exposure and illness is one which is associated with the exposure and with the illness."

To quantify some of these notions statistically, consider the regression model

$$E(Y/Z,X) = \alpha_0 + (\alpha_1 - \alpha_0)Z + \beta X,$$

where $Y$ is the exposure variable, $X$ is a possible confounding factor, and $Z$ is a dummy variable taking the value 1 for a case and the value 0 for a control. This is a simple two-group analysis of covariance model with a single covariate. The difference between cases and controls is reflected in the parameter $(\alpha_1 - \alpha_0)$, but we can equivalently look at the population partial correlation coefficient between $Y$ and $Z$ controlling or adjusting for the effect of $X$, namely

$$\rho_{YZ/X} = \frac{\rho_{YZ} - \rho_{YX}\rho_{ZX}}{\sqrt{(1-\rho_{YX}^2)(1-\rho_{ZX}^2)}}.$$
The two parameters \((\alpha_1 - \alpha_0)\) and \(\rho_{YZ/X}\) are intimately related in the sense that a statistical test of \(H_0: \alpha_1 = \alpha_0\) versus \(H_a: \alpha_1 \neq \alpha_0\) based on fitting the above regression model is mathematically equivalent to a test of \(H_0: \rho_{YZ/X} = 0\) versus \(H_a: \rho_{YZ/X} \neq 0\), the test being based on the use of a "partial F" statistic. In other words, this test of significance regarding \(\rho_{YZ/X}\) is equivalent to the usual analysis of covariance test of no difference in the population group means after adjusting them for the effect of the covariate.

In this context, then, Miettinen's definition can be translated to mean that \(X\) is a confounding factor if and only if both \(\rho_{YX} \neq 0\) and \(\rho_{ZX} \neq 0\). Note that it is then possible that \(\rho_{YZ/X} = 0\) (or, equivalently, that \(\alpha_1 = \alpha_0\)) when \(\rho_{YZ} \neq 0\), and also that \(\rho_{YZ/X} \neq 0\) when \(\rho_{YZ} = 0\). When \(\rho_{YZ/X} = 0\) and \(\rho_{YZ} \neq 0\), there would be no difference in the population group means after adjusting for \(X\), but there would be a difference before \(X\) is considered. The case \(\rho_{YZ/X} \neq 0\) and \(\rho_{YZ} = 0\) implies a difference in adjusted population means where there is no difference in unadjusted means. Both of these cases describe instances where analysis of covariance can be used to avoid making incorrect inferences due to ignoring \(X\).

On the other hand,

\[
\rho_{YZ/X} = \frac{\rho_{YZ}}{\sqrt{1-\rho_{ZX}^2}} \quad \text{when } \rho_{YX} = 0, \ \rho_{ZX} \neq 0
\]

and

\[
\rho_{YZ/X} = \frac{\rho_{YZ}}{\sqrt{1-\rho_{YX}^2}} \quad \text{when } \rho_{YX} \neq 0, \ \rho_{ZX} = 0.
\]
so that $\rho_{YZ/X}$ is zero or non-zero according as $\rho_{YZ}$ is zero or non-zero.

When $\rho_{YY} = 0$ and $\rho_{ZX} \neq 0$, the groups have different population means for the covariate but there is no regression of $Y$ on $X$, so that there is no need to consider $X$. When $\rho_{YY} \neq 0$ and $\rho_{ZX} = 0$, the covariate population means are equal, so that $X$ can be ignored even though it is related to $Y$ (i.e., $\rho_{YY} \neq 0$); equivalently, if $E(X/Z) = \gamma_0 + \gamma_1 Z$, then

$$E(Y/Z) = E_X E(Y/Z,X) = (\alpha_0 + \beta \gamma_0) + (\alpha_1 - \alpha_0 + \beta \gamma_1)Z,$$

so that the coefficient of $Z$ is exactly $(\alpha_1 - \alpha_0)$ when $\gamma_1 = 0$ (i.e., when $\rho_{ZX} = 0$), regardless of the value of $\beta$.

The remaining case $\rho_{YY} = 0$, $\rho_{ZX} = 0$ means that $X$ has absolutely no effect on $Y$ or $Z$, so that $\rho_{YZ/X} = \rho_{YZ}$ and $\beta = 0$.

In reality, of course, we do not know any of these population parameters. In a typical observational study, for instance, there may be several variables which seem to be operating as confounders to some degree in the observed data set. As is often the situation, these factors will have "somewhat" different distributions in the case and control groups and the data will often suggest "somewhat" of a relationship between these factors and the response. The statistical effort then becomes one of designing the study (e.g., choosing the appropriate control group) and of analyzing the data in order to make valid inferences concerning the true exposure-disease relationship. It is important to emphasize in this regard that the necessity for controlling or adjusting statistically for a possible confounding factor depends on the characteristics of that factor in the observed data; in other words, the
designated confounder is necessarily a sample-based label rather than a population-based label. More specifically, a factor which is strongly suspected to be related to disease and exposure in the population may not manifest any such properties in the particular sample under consideration and so would not be considered to be a confounder. On the other hand, a factor considered a priori not to be related to disease or exposure in the population may turn out to exhibit strong confounding properties in the sample, in which case it is necessary to treat the factor as a confounder in the analysis.

To pursue some of these concepts a bit further, consider the simple situation where the exposure variable \( Y \) is postulated to be linearly related to a single factor \( X \) via the model

\[
Y_{ij} = \alpha_i + \beta_i X_{ij} + e_{ij}, \quad i = 0,1 \text{ and } j = 1,2,\ldots,n_i, \quad (1.3.1)
\]

where \( Y_{ij} \) is the observed value of the exposure variable for the \( j \)-th individual in the \( i \)-th group (\( i = 0 \) for controls and \( i = 1 \) for cases), \( X_{ij} \) is the observed factor value associated with \( Y_{ij} \), and \( e_{ij} \) is the usual error term. Further, for notational convenience, suppose that \( f_i(x) \) denotes the observed distribution of sampled values of \( X \) for the \( i \)-th group, \( i = 0,1 \).

Now, let us examine some possible outcomes that might result from fitting model (1.3.1.) to observed data and then discuss what interpretations can be made. One possible outcome is depicted in Figure 1.3.1 below. Here, we have tried to portray the situation where neither \( \beta_0 \) nor \( \beta_1 \) is statistically significantly different from zero (we have drawn the figure such that \( \hat{\beta}_0 = \hat{\beta}_1 = 0 \)), indicating that there is no evidence that
X and Y are related (for simplicity, let us assume that X and Y are at most linearly related). For these data, then, X would be considered not to be a confounder, regardless of whether $\hat{f}_0(x)$ and $\hat{f}_1(x)$ are the same or not.

![Diagram](image.png)

**FIGURE 1.3.1**

As far as inferences from sample to population are concerned, one would be inclined to take the position that $\beta_0 = \beta_1 = 0$, in which case, the parameter $(\alpha_1 - \alpha_0) = E(Y_{1j} - Y_{0j})$ would be estimated as $(\bar{Y}_1 - \bar{Y}_0)$ with population variance $\sigma^2 \left[ \frac{1}{n_1} + \frac{1}{n_0} \right]$ then estimated as $S_p^2 \left[ \frac{1}{n_1} + \frac{1}{n_0} \right]$, where $S_p^2$ is the usual two-sample pooled estimate based on $(n_1 + n_0 - 2)$ df under the usual homogeneous variance assumption.

Figures 1.3.2(a) and 1.3.2(b) are meant to represent the situation where $\hat{\beta}_0$ and $\hat{\beta}_1$ are both significantly different from zero but are not
significantly different from one another (we have drawn the figure such that \( \hat{\beta}_0 \neq \hat{\beta}_1 > 0 \)), indicating that X and Y are linearly related for these data and that the linear relationship (as reflected in the slope estimates) is, for all practical purposes, the same in the two groups. In Figure 1.3.2(a), \( \hat{f}_0(x) \neq \hat{f}_1(x) \), so that X would be considered not to be a confounder (note that only the condition \( \overline{X}_0 = \overline{X}_1 \) is needed so that \( r_{ZX} = 0 \)). In Figure 2(b), \( \hat{f}_0(x) \neq \hat{f}_1(x) \), so that X would be considered to be a confounder since it is related to both exposure and disease in this data set (although we could still have \( \overline{X}_0 = \overline{X}_1 \)).

\[ \text{FIGURE 1.3.2(a)} \]

\[ \text{FIGURE 1.3.2(b)} \]

With regard to inference-making, the data support the contention that \( \beta_0 = \beta_1 (= \beta, \text{say}) \), so that the model (1.3.1) takes the form

\[ Y_{ij} = \alpha_i + \beta X_{ij} + \epsilon_{ij}, \quad (1.3.2) \]
which (given the usual assumptions) is the ordinary analysis of covariance model. Here,

\[
(\hat{\alpha}_1 - \hat{\alpha}_0) = (\bar{Y}_1 - \bar{Y}_0) - \hat{\beta}(\bar{X}_1 - \bar{X}_0) ,
\]

is the covariance estimator of \((\alpha_1 - \alpha_0) = E(Y_{1j} - Y_{0j} | X_{1j} = X_{0j})\),

where

\[
\hat{\beta} = \frac{\sum_{i=0}^{1} \sum_{j=1}^{n_i} (X_{1j} - \bar{X}_1)(Y_{1j} - \bar{Y}_1)}{\sum_{i=0}^{1} \sum_{j=1}^{n_i} (X_{1j} - \bar{X}_1)^2}
\]

and

\[
\text{Var}(\hat{\alpha}_1 - \hat{\alpha}_0) = \sigma^2 \left[ \left( \frac{1}{n_1} + \frac{1}{n_0} \right) + \frac{(\bar{X}_1 - \bar{X}_0)^2}{\sum_{i=0}^{1} \sum_{j=1}^{n_i} (X_{1j} - \bar{X}_1)^2} \right] ,
\]

which (for fixed sample sizes \(n_1\) and \(n_0\)) is smallest when \(\bar{X}_0 = \bar{X}_1\). When this condition holds, as it does in Figure 2(a) since \(\hat{f}_0(x) \approx \hat{f}_1(x)\), then \((\hat{\alpha}_1 - \hat{\alpha}_0) = (\bar{Y}_1 - \bar{Y}_0)\), with \(\text{Var}(\bar{Y}_1 - \bar{Y}_0) = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_0} \right) \). Note that \(\sigma^2\) is estimated here using the residuals from the estimated regression equation 1.3.2. and not using deviations from the group means as in the situation depicted in Figure 1.3.1. When \(\hat{f}_0(x) \neq \hat{f}_1(x)\), as in Figure 1.3.2(a), then \(\bar{X}_1\) is not necessarily equal to \(\bar{X}_0\). When \(\bar{X}_1 \neq \bar{X}_0\), then the estimator \((\bar{Y}_1 - \bar{Y}_0)\) is biased, but has a smaller population variance than
the unbiased covariance estimator. If one is not opposed to using a
biased estimator, then \((\bar{Y}_1 - \bar{Y}_0)\) will be "better" (in terms of mean square
error) than the covariance estimator in certain situations. But, whether
\(\bar{X}_1\) equals \(\bar{X}_0\) or not, it is still necessary to use analysis of covariance
since, as Greenberg (1953) points out, it provides the correct estimate
of experimental error (one based on \(n_1 + n_0 - 3\) df) under the given model
assumptions. With regard to the above variance considerations, a good
(but not necessarily optimal) procedure under model (1.3.2.) is to use
analysis of covariance and to ensure in some way that \(\bar{X}_1 \neq \bar{X}_0\). One can
often arrange that \(\bar{X}_1 \neq \bar{X}_0\) by selecting the control group appropriately
(e.g., by employing some kind of matching procedure); matching and stra-
tification procedures will be discussed in Sections 1.5 and 1.6.

Figures 1.3.3(a) and 1.3.3(b) depict the complex situation where \(\hat{\beta}_1\)
and \(\hat{\beta}_2\) are significantly different from one another (we have drawn the
figure such that \(\hat{\beta}_1 > \hat{\beta}_0 > 0\)), indicating that the linear relationship
between X and Y is not the same in the two groups. In statistical term-
inology, this is known as "interaction" between X and the case-control
group designation variable Z.

The presence of this "group by covariate" interaction effect greatly
complicates the interpretation of any case-control differences with respec-
to Y, usually meaning that group comparisons can validly be made
only at specific values of the covariate. As far as confounding is con-
cerned, Figure 1.3.3(a) portrays an interaction - no confounding situa-
tion since \(\hat{f}_0(x) = \hat{f}_1(x)\), while Figure 1.3.3(b) illustrates an inter-
action - confounding situation since \(\hat{f}_0(x) \neq \hat{f}_1(x)\).
As far as inferences are concerned, one would tend to believe that \( \beta_1 > \beta_0 > 0 \). If, for example, the parameter of interest is taken to be the true average difference in response for the \( n_1 \) cases, namely

\[
\frac{1}{n_1} \sum_{j=1}^{n_1} E(Y_{1j} - Y_{0j} / X_{1j} = X_{0j}) = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0) \bar{x}_1,
\]

then the unbiased estimator of this quantity is

\[
(\bar{Y}_1 - \bar{Y}_0) - \hat{\beta}_0 (\bar{x}_1 - \bar{x}_0),
\]
which is of the same form as the usual covariance estimator based on model 1.3.2 except that the slope estimate is only based on the control group data and is not the pooled estimate $\hat{\beta}$. These considerations have been discussed by Cochran (1969) and the above estimator has been applied by Belsley (1956).

1.4. Effect Modification, Synergism and Antagonism

The situation depicted in Figures 1.3.3(a) and 1.3.3(b) has been considered in the epidemiologic literature under the names of "effect modification" and "synergism and antagonism". In this section, we shall first consider a possible general etiology leading to a disease and then consider the literature relating to this subject.

Suppose we have an exposure and a risk indicator that can lead to a disease. This could be expressed functionally as:

$$Pr(\text{Disease}) = f(E, W),$$

where $E =$ exposure

$W =$ level of risk indicator.

For simplicity, we will assume that this function can be expressed as:

$$f(E, W) = f(E) + g(W) + h(E, W).$$

This would serve as an example although our remarks apply to a general function. In a case-control situation, we sample according to whether or not the disease is present and attempt to examine the data to determine if the exposures differ between case and control groups.

If we assume that $h(E, W) = 0$, then the post-sampling data should appear as in Figures 1.3.2(a) and 1.3.2(b). By assuming that $h(E, W)$ is zero, we are assuming that $E$ and $W$ do not interact in causing the disease. Rothman (1974, 1976) has considered this situation. He uses the terms "synergy" and "antagonism" to describe the possible forms $h(E, W)$ might
take. When the relationship between the exposure and the risk indicator is such that \( h(E,W) \) is positive, then the probability of the disease is greater than that which would be expected when \( h(E,W) = 0 \). Likewise, when \( h(E,W) < 0 \), the probability is less than would be expected. For the first situation, \( h(E,W) > 0 \), Rothman uses the term, "synergism" and for the second, "antagonism".

Figures 1.3.3(a) and Figure 1.3.3(b) show how the data might appear in a case-control study when \( h(E,W) \neq 0 \). We would expect that the cases would be those subjects for whom there was a high probability of disease, while the controls are those for who there was a lower probability of disease. Thus, when we control for the risk indicator, we would tend to get a different relationship between \( Y \) and \( X \) in the two groups.

Miettinen (1974) has introduced the term, "effect modification", which includes the situation mentioned above. He was writing in response to an article by Fisher and Patil (1974) if which they pointed out that an effect, such as a relative risk or a risk difference, might vary over subcategories of a confounder. Miettinen points out that the effect modifier need not necessarily be a confounder. He also makes the point that all risk indicators modify at least the relative risk or the risk difference. This is because, if the risk indicator does not modify the relative risk, it will modify the risk difference and vice-versa. Thus, the term "effect modification" could be simply defined as non-uniformity of the parameter of interest over subcategories of a confounder.

1.5 Matching

The technique of matching, as the name suggests, involves choosing a group of controls such that members of this group have \( X \) values which "match" those of the \( n_1 \) cases according to some reasonable criterion.
One quite stringent matching criterion is that utilized in the popular technique known as "pair-matching". In a pair-matching scheme, each of the \( n_1 \) cases is matched or "paired up" with a control who has the same set of concomitant variable values as the case does. In some instances, of course, there may be several controls which can be matched to a particular case, and one would then simply choose one of these candidates at random. It is important to note that we are not talking here about the situation where there is a "natural" (and obvious) pairing, as when dealing with "twins" (or their equivalent), in which case there is no randomness at all associated with the forming of pairs and the responses within a pair are inherently correlated to some degree. In other words, we are assuming that the \( Y_{ij} \)'s are statistically independent of one another.

It is easy to see that the use of pair-matching necessitates that the size of the control group sample must equal the size of the case group sample (i.e., \( n_0 = n_1 \)), which could quite possibly represent a sizeable loss in sensitivity if \( n_1 \) is small. In addition, the difficulty in finding exact pair-matched controls increases quite rapidly as the number of \( X \)'s under consideration increases, which necessitates (often at great expense) the establishment of a very large pool of controls in order to be able to find a partner for each of the \( n_1 \) cases.

On the other hand, there are some attractive features associated with exact pair-matching. For example, the use of exact pair-matching assures that the distribution of sample \( X \) values is exactly the same in the two groups. Also, it can be shown that exact pair-matching removes all the bias in \( \overline{Y}_1 - \overline{Y}_0 \) with regard to estimating \( (\alpha_1 - \alpha_0) \), whatever the true model relating \( Y \) and the \( X \)'s happens to be, as long as it is the same model in both groups (i.e., as long as there is no "interaction" in
the sense described earlier). In general, the appropriate statistical procedure to use here (if the assumptions of (1.3.2.) hold) is analysis of covariance, which, when coupled with pair-matching, leads to the unbiased point estimator \((\hat{\alpha}_1 - \alpha_0) = (\overline{Y}_1 - \overline{Y}_0)\) of \((\alpha_1 - \alpha_0)\), with \(\text{Var}(\overline{Y}_1 - \overline{Y}_0) = 2\sigma^2/n_1\).

The analysis of covariance estimate of \(\sigma^2\) would be based on

\((2n_1 - p - 2)\) df, \(p\) being the number of covariates in the model; and, under (1.3.2.), analysis of covariance would provide a more powerful test of \(H_0: \alpha_1 = \alpha_0\) than would the usual paired t-test (i.e., a randomized blocks type of analysis) as long as \((n_1 - p) > 1\).

One way to lower some of the expense of exact pair-matching is to relax the basic requirement that each case-control pair has to match exactly on the \(X\) values. For a given \(X\), for example, we might only require that the \(j\)-th case with value \(X_{1j}\) be matched to a control with value \(X_{0j}\) satisfying

\[|X_{1j} - X_{0j}| \leq c,\]

where \(c\) is some positive constant. Cochran and Rubin (1973) refer to this procedure as paired "caliper-matching". They investigated the behavior of a quantity they called the "percentage reduction in bias" under caliper-matching using \((\overline{Y}_1 - \overline{Y}_0)\) to estimate \((\alpha_1 - \alpha_0)\) in model (1.3.2.) based on an underlying normal distribution for \(X\). As would be expected, they found that "tight" caliper-matching removes practically all the bias, while "loose" caliper-matching removes considerably less (about 75% on the average). As with exact pair-matching, caliper-matching requires that \(n_0 = n_1\), which, in some situations, may be an unnecessarily low control group sample size to be stuck with. For example, in the
situation considered by Cochran and Rubin, one need only equate the sample mean of $X$ in the two groups (i.e., make $\bar{X}_1 = \bar{X}_0$) in order to eliminate the bias in $\bar{Y}_1 - \bar{Y}_0$ due to a possible linear relationship between $X$ and $Y$. And, mean-matching (or "balancing", as it is sometimes called) is certainly a less demanding matching procedure than is exact pair-matching (which is equivalent to matching on all the sample moments of $X$), and so should provide a control group sample size much larger than $n_1$. Similar conclusions should hold regarding the size of a control group chosen to satisfy $|\bar{X}_1 - \bar{X}_0| \leq c$ as opposed to that based on paired caliper-matching.

In general, mean-matching and pair-matching are special cases of a general technique called "moment-matching", in which the control group sample is chosen to match the case group sample with respect to the values of certain sample moments of the $X$'s. As an illustration of an application of moment-matching, given the model $Y_{ij} = \alpha_1 + \beta X_{ij} + \beta_1 X_{ij}^2 + \varepsilon_{ij}$, choosing the control group sample such that the mean and variance of $X$ for this sample is equal to the sample mean and variance of $X$ for the cases ensures that $\bar{Y}_1 - \bar{Y}_0$ will be the analysis of covariance unbiased point estimator of $(\alpha_1 - \alpha_0)$.

1.6. **Stratification**

There are other matching procedures besides the general technique of moment-matching, which includes balancing and pair-matching as special cases. Cochran and Rubin (1973) offer some discussions and evaluations of some of these alternative procedures. One of these techniques is worth some discussion, namely, the technique of "frequency-matching". To "frequency-match", one would first group the $n_1$ cases into strata defined in terms of values of covariates of interest. For example, if
age and sex were concomitants, one might define six strata: males, 30-30; males, 40-49; males, 50-59; females, 30-39; females, 40-49; and, females, 50-59. Then, each case would be allocated to one and only one of the six strata. In general, suppose that \( n_{1\ell} \) of the \( n_{1} \) cases were allocated to the \( \ell \)-th of \( s \) strata, so that \( \sum_{\ell=1}^{S} n_{1\ell} = n_{1} \). The next step would be to come up with \( n_{0\ell} = n_{1\ell} \) (or some multiple thereof) controls which also fall into the \( \ell \)-th stratum based on their covariate values. By doing this for all the strata, one guarantees that the relative frequencies of cases and controls in the various strata are identical, which leads to statistically efficient analyses because of the "balance" introduced. The appropriate case-control comparisons are then made within strata, and the within-strata information can then be combined (as long as group differences do not vary from stratum to stratum) to provide an overall assessment of the exposure-disease relationship.

The motivation behind the use of frequency-matching is analogous to that associated with the use of a randomized block design; the strata are the blocks, within which the "treatment" (i.e., case-control) comparisons are expected to be much more sensitive than they would be without blocking. Note that pair-matching is a very special case of frequency-matching where \( n_{0\ell} = n_{1\ell} = 1, \ell = 1, 2, \ldots, n_{1} \).

The term "stratification", although referring to the process of forming strata as is done in frequency-matching, is to be distinguished from the term "matching" in the following sense. Stratification is generally considered to be an analysis procedure rather than a sampling procedure like matching, or equivalently, it is an "a posteriori" approach (i.e., an approach to be used after the data has been gathered), as opposed to an "a priori" approach like matching (i.e., a sampling
procedure to be decided upon before the data is collected), in which the "balance" achieved by matching is not necessarily guaranteed by stratification. In any case, whether matching or stratification is employed, the important point is that the appropriate analysis is one that takes into account the increase in sensitivity (if any) achieved by forming the strata; in other words, a "stratified" analysis should be performed which is based on utilizing within-strata comparisons. Recently, Miettinen (1976) has suggested stratifying on the basis of a "multi-variable confounder score" calculated using the sample linear discriminant function. The statistical advantages and disadvantages of such an approach have not been thoroughly studied.

1.7. Literature Review

In this section, we will review statistical and epidemiological literature which has considered our research question. This includes discussions of papers which have examined the properties of matched and unmatched estimators of case-control differences, papers which have offered guidelines on the choice of variables to be matched on or otherwise accounted for in the analysis, and some papers which have considered the effect of matching on sample size.

Cochran (1953) examined the variances of estimators of the case-control difference when the data had been randomly sampled, pair-matched, and when covariance adjustment had been used. He assumed that the covariable, x, say, was related to the variable of interest, y, by the following linear regression:

for cases: \[ y = \alpha + \beta x + d \]

for controls: \[ y' = \alpha' + \beta x' + d'. \]
He first considered the situation where the investigator is using an "ideal control" (i.e., \( \mu_x = \mu_x' \), \( \sigma_x^2 = \sigma_x'^2 \), \( \sigma_y^2 = \sigma_y'^2 \)), so that the difference between the two populations is \((\alpha - \alpha')\). The relationship in the population would appear as in Fig. 1.3.2(a) shown previously; and, \(x\) is not a confounder in the sense we described earlier. The variance of \((\bar{y} - \bar{y}')\) is derived for the random sample and pair-matched sample and the two are compared. It should be noted that Cochran assumes that the pair-matches are exact, which is tantamount to assuming that an extremely large pool of samples from which to choose the pairs is available. The variance of \((\bar{y} - \bar{y}')\) is \((1 - R^2)\) that of the random sample variance, assuming equal sample sizes in both groups, where \(R\) is the correlation between \(X\) and \(Y\). Values of \(|R|\) and \((1 - R^2)\) are shown below in Table 1.7.1 Cochran believes that in this case (the "ideal control") the gains in precision are modest for \(|R| \leq 0.5\) and that a variable should not be considered for matching until its

<table>
<thead>
<tr>
<th>TABLE 1.7.1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Values of ((1 - R^2))</strong></td>
</tr>
<tr>
<td>(R)</td>
</tr>
<tr>
<td>((1 - R^2))</td>
</tr>
</tbody>
</table>

correlation with \(y\) is such that the variance will be decreased by some appreciable amount. He suggests that a reasonable level of correlation at which to consider matching is \(|R| \geq 0.3\).

The variance of the pair-matched estimate is \((1 + \frac{1}{2(n-2)})^{-1}\) that of the covariance adjusted estimate. For small \(n\), there is some small increase in precision by choosing a pair-matched sample over a covariance
adjusted sample, but there is no effect gain for large 𝑛 (i.e., 

\(1 + \frac{1}{2(n-2)} \leq 1\)). When \(\mu_x \neq \mu_{x'}\), pair-matching seems to be much more precise than covariance analysis, since the variance of the covariance estimate is 

\[1 + \frac{1}{2(n-2)} + \frac{n(x-x')^2}{4(n-2)\sigma_x^2}\] 

times that of the pair-matched estimate, although Cochran believes that "the increase in precision from pairing relative to covariance is probably not that great in practice."

He makes a judgment in favor of pair-matching in general, though, since it offers "the simpler method of analysis". It could be argued that this no longer applies since this comment was made before the advent of high-speed computers.

Cochran also gives some consideration to the technique of balancing (mean-matching) and notes that it is effective in removing common linear effects, but does not have as great an effect on higher-order terms. He does not consider this technique in combination with covariance analysis where it can be shown that mean-matching removes the variance due to \(\overline{x} - \overline{x}'\).

In a paper published with the one just discussed, Greenberg (1953) considers the use of covariance analysis and balancing (mean-matching). He shows that a covariance estimator is more sensitive that a simple mean difference. He then comments on the ability of balancing to remove the linear effect, and points out that balancing alone does not reduce the error term as much as possible and that covariance analysis will accomplish this. He shows that even if covariance analysis is used, balancing the sample will still increase the sensitivity of the test. The numerator of the F ratio for testing differences between groups is stated to have a value equal to
\[
\delta^2 + n \left\{ 1 - \frac{t_{xx}}{s_{xx}} \right\} \Delta_t^2
\]

where \( \delta^2 \) = true error

\( n \) = number of subjects in each group

\( t_{xx} \) = between group sum of squares for the matching variable

\( s_{xx} \) = within group sum of squares for the matching variable

\( \Delta_t^2 \) = sum of squares for true treatment effects

It can be seen that this is maximized when \( t_{xx} / s_{xx} \) is minimized, and \( t_{xx} \) is equal to 0 when the samples have been mean-matched on the covariate.

If \( t_{xx} / s_{xx} \) is considered as the loss of sensitivity due to failure to balance, some measure can be made of gain due to balancing. Greenberg cites some work showing, when the x's are identically distributed normal variables, that \( \mathbb{E} \left( \frac{t_{xx}}{s_{xx}} \right) = \frac{1}{2n-1} \); so, for large \( n \), there is little gain in balancing, while with small \( n \) it would help to match. He does not cite any results for the case when the means of the x's differ in the population, which would make balancing a more attractive technique.

An extensive Monte Carlo investigation of the efficiency of matched sampling was undertaken by Billewicz (1965). When considering quantitative matching variables, he assumed that the X and Y variates were distributed \( \mathcal{N}(0,1) \) and \( \mathcal{N}(\mu_y,1) \) so that the variance of the difference between the responses is two. Note that he is also assuming that X has the same distribution in both populations.

Billewicz first investigated the number and width of categories which would give the minimum variance for the mean-difference when these categories were used to pair-match the samples. He found three category matching to be only slightly less precise than four
category matching when $R \leq 0.5$ and limited his comparisons to this three category structure for the remainder of the paper. These categories were formed by taking a central interval of $\bar{X} + k$. During this investigation, he found the optimal value of $k$ to be $k = 0.5$. At this value of $k$, the variance increased with $R$, going from 2.06 when $R = 0.2$ to 2.68 when $R = 0.8$.

He then investigated the relative efficiency of matched pairs to covariance analysis. $\mu_{y|x}$ is assumed to be linear in $X$ (as in Fig. 1.3.2). Table 1.7.2 shows some of these efficiencies.

### TABLE 1.7.2
Relative Efficiency of Matching and Covariance Analysis

(3 matching categories)

<table>
<thead>
<tr>
<th>Correlation</th>
<th>% Efficiency for Given Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large Samples</td>
</tr>
<tr>
<td>0.2</td>
<td>98.0</td>
</tr>
<tr>
<td>0.4</td>
<td>97.0</td>
</tr>
<tr>
<td>0.6</td>
<td>89.2</td>
</tr>
<tr>
<td>0.8</td>
<td>74.3</td>
</tr>
</tbody>
</table>

\[ \text{from Billewicz (1965). Relative Efficiency} = \frac{(n_1+1)(n_2+3)S^2_1}{(n_2+1)} \]

\[ \frac{(n_1+3)S^2_2}{S^2_1}, \text{ where } n_1 \text{ and } n_2 \text{ are d.f. of error variances } S^2_1 \text{ and } S^2_2 \text{ for matched and simple random samples respectively (Fisher[1947])}. \] For each value of $r$, 400 matched and simple random samples of each size were generated. Estimates of Relative Efficiency for large samples are based on pooled estimates for variance from all sample sizes.

It can be seen that this efficiency decreases for decreasing $n$ and/or for increasing $R$. It should be noted, however, that for $R = 0.8$, where there are lower efficiencies, a model with more matching categories might show better efficiencies.
He then examined the ability of the matched pairs and of covariance analysis to detect non-parallel linear regressions. To detect non-parallel regression slopes in the matched pairs analysis, coefficients for a regression of the differences on a dummy variable to represent matching categories were tested at 5% level. If this test was significant, he considered it indicative of non-parallel slopes. Using this scheme, matched pairs analysis detected the non-parallel slopes in 19.0% of 200 trials. Covariance analysis detected the difference in 20.5% of the sample runs. No efficiency analysis was attempted. Billewicz points out that "the average user of matched samples would not examine his sample in this respect and so would miss this important feature."

The next topic he considered was parallel non-linear regressions. He used three regressions to investigate the ability of the matched pairs method to detect the non-linearity. The three lines are:

\[
\begin{align*}
    y &= 0.4x - 0.1x^2 \\
    y &= 0.8x - 0.14x^2 \\
    y &= \tanh(x)
\end{align*}
\]

The non-linearity was detected in 12.3%, 20.3% and 19.8% of the 400 trials, respectively, when a sample size of 40 was used. The text does not indicate what treatment difference was added in for the control population. Using a linear regression, he found that matching was no more than 93% efficient compared with the regression in detecting the difference. Using 4 categories improved this ratio slightly.

For a quantitative response with qualitative matching variables, he used equidistant dummy variable values to perform the comparison. The matching variable was bivariate with 3 categories when the case and control covariate was drawn from the same population; pair-matching was
87.4% efficient at detecting the case-control difference relative to simple covariance analysis and covariance analysis on frequency matched data was 101.4% efficient. He comments that "it seems that the crude device of regression on equally spaced dummy variables is at least as effective as matching, while being administratively much simpler." He does not comment on the method of analysis of variance, which would seem to be a more natural approach. To examine the situation where case and control populations have different distributions for the control variates, he made three types of comparisons. He used 100 samples of 80 subjects in each group. In his first comparison, there was no difference in the response other than that due to the difference in the covariate; in the second comparison, the treatment effect was opposite in sign to the difference due to the covariate so that it was obscured by the population difference and in the third comparison, the treatment effect was of the same sign, so as to accentuate the difference. The efficiency of the matched samples was high (98-99%). Covariance analysis of frequency matched samples was no more efficient than that of covariance analysis of random samples. In each of the three situations, the matched pairs estimate of treatment effect was closer to the true value while its variance was higher. He points out that when matching from different populations, the relationship between the two variables need not be the same in both populations. In performing covariance analysis this would be checked as part of the analysis, but this would not necessarily be the case in matched-pairs analysis.

When both the response and matching variate are qualitative, he compares only matched pairs analyses and random samples. The matched samples are shown to be more effective than the random samples. No complex form of categorical analysis is used.
Table 1.7.3 below, summarizes the results of this investigation.

**TABLE 1.7.3**

Results of Monte Carlo Efficiency Investigation for a Continuous Response

<table>
<thead>
<tr>
<th>Type of Covariate</th>
<th>Covariate Relationship</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>X Continuous, same distribution in both populations\textsuperscript{b}.</td>
<td>linear and parallel</td>
<td>Covariance analysis on random samples is always more efficient than matched analysis.</td>
</tr>
<tr>
<td></td>
<td>linear but non-parallel</td>
<td>No efficiency analysis undertaken. Both methods detected non-parallelism at the same rate.</td>
</tr>
<tr>
<td></td>
<td>non-linear and parallel\textsuperscript{c}</td>
<td>Pair-matching is always less efficient.</td>
</tr>
<tr>
<td>X Categorical, with same distribution in both populations.</td>
<td>Equidistant dummy variables used.\textsuperscript{d}</td>
<td>Pair-matching is less efficient than covariance analysis on frequency matched samples.</td>
</tr>
<tr>
<td>With different distributions.</td>
<td>Same</td>
<td>The three methods used above were essentially the same.\textsuperscript{e}</td>
</tr>
</tbody>
</table>

\textsuperscript{b}X~N(0,1)

\textsuperscript{c}Three different functions used; see text above.

\textsuperscript{d}It is not clear whether or not interaction terms were included.

\textsuperscript{e}Matched analysis gave estimates which were closer to the true values, but had larger variances than covariance analysis.

In a paper on the analysis of the matched-pairs design, Meittinen (1968) commented in reference to the case-control design that "... matching in such a design, when not needed for validity, has a general tendency merely to reduce design efficiency" (p. 347).

Bross (1969) took umbrage with this statement and published a paper criticizing "superficial theoretic evaluations" and offered a "simple
hypothetical counter-example". This counter-example used data gathered for a two-strata, pair-matched, case-control study. Bross analyzed the data in two ways: first, using the pair-matched analysis which was appropriate to the sampling plan under which the data was gathered. Second, he used a random reallocation of the same sample as the comparison design. The test statistics of the first design was 1.33 times as large as that of the second. Bross used this to claim that the matched-pairs design improved the efficiency of the study.

Miettinen (1970) responded to this criticism. He immediately pointed out that the supposed "design efficiency" Bross had computed was actually only an analysis efficiency. All that he had shown was that using an analysis that is appropriate to the sampling procedure used to collect the data (as was the pair-matched analysis based on the strata) will be more efficient than a design that is inappropriate to the sampling plan (as was the random repairing disregarding the strata). To compare actual design efficiencies, he reconstructed the source population and computed the expected value of a statistic on that population. Under the assumption that the matching factor was not a risk factor of the disease given exposure (so that both designs are valid), he showed that the matched-pairs design was less efficient. He then suggested some basic requirements in the response-covariable relationship that allow matching. These have already been discussed.

Cochran (1965) considered the possible approaches to "disturbing" variables. He makes the point that the methods of analysis of observational studies should be designed to minimize both bias and variance. The additional constraint of minimizing bias is not considered in the analysis of experimental design, where it is assumed that the experimenter
has randomly allocated the subjects to each of the experimental groups. This non-random allocation is included as the second point in Cochran's definition of the observational study.

Cochran classifies possibly disturbing variables into three classes. The first of these is the set of variables which require some form of matching or adjustment to remove bias. Ideally, the number of variables in this class is to be kept small. The second set contains those variables for which matching or adjustment would be helpful, but, for which we "content ourselves with some verification that their effects produce little or no bias". The third group consists of those variables which are disregarded since their effects are thought to be minor. There is no discussion of this class of variables.

For a categorical variable with k classes in the second group, Cochran suggests the usual $\chi^2$ for a 2 x k contingency table. He does not offer any idea what amount of bias is associated with large $\chi^2$'s in the samples. For continuous $X$, he suggests that if the following regression holds

$$y = \mu_y + \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \epsilon$$

$$y = \mu_y' + \beta_0 + \beta_1 x' + \beta_2 x'^2 + \ldots + \beta_k x^k' + \epsilon',$$

then the bias in ($\bar{y} - \bar{y}'$) is

$$\beta_1 (\mu_1 - \mu_1') + \beta_2 (\mu_2 - \mu_2') + \ldots + \beta_k (\mu_k - \mu_k'),$$

where $\mu_1 = E(X^i_1)$, $\mu_1' = E(X'^i_1)$, so that verification that the sample means are the same guards against bias arising from a linear regression, while
similar sample means and variances guard against a quadratic term and so on. Thus, the common practice of examining the marginal distributions of several covariables could lead to error when the relation between $x$ and $y$ includes cross-products. He suggests that when the relationship is not known it would be best to examine the whole multivariate distribution. This is somewhat at variance with a comment by Miettinen (1974), which we will consider below.

Now the absence of statistical significance in the above tests on sample moments does not necessarily guarantee that there is no bias. Cochran realizes this and says "although these checks on the $x$-distribution are usually made by tests of significance, it is not clear what kind of assurance is given by the finding of a nonsignificant result, nor that a test is the appropriate criterion." He gives an example of this by assuming a linear regression of $y$ on $x$. For repeated samples with $(\bar{x} - \bar{x}')$ fixed

$$ (y - y') \sim N(\mu - \mu', \beta \bar{x} - \bar{x'}), \frac{2\sigma^2}{n} ) . \quad (1) $$

If we assume no bias due to $x$ then

$$ (y - y') \sim N(\mu - \mu', \frac{2}{n} (\sigma_e^2 + \beta^2 \sigma_x^2)) . \quad (2) $$

It can be shown that the 95% confidence interval on $(\mu - \mu')$ for large samples will be, if we assume (2) when (1) is true,

$$ -tv \pm 1.96 \sqrt{1 + v^2} , \quad (3) $$
where
\[ v = \frac{\beta \sigma_t^2}{\sigma_e} \]
\[ t = \frac{(\bar{x} - \bar{x}')}{\sigma_x} \sqrt{n} \]

where \( t \) is the test criterion for \( H_0: \mu_x = \mu_{x'} \). As can be seen from (3), the 95% confidence interval can be shifted from the "correct" central interval to one that is not centered around \( \mu - \mu' \). Table 1.7.4 below contains the actual width, in terms of probability, of the 95% interval generated assuming (2).

<table>
<thead>
<tr>
<th>( v )</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.957</td>
<td>0.961</td>
<td>0.966</td>
<td>0.972</td>
<td>0.988</td>
</tr>
<tr>
<td>1.0</td>
<td>0.950</td>
<td>0.950</td>
<td>0.951</td>
<td>0.953</td>
<td>0.963</td>
</tr>
<tr>
<td>1.5</td>
<td>0.948</td>
<td>0.931</td>
<td>0.924</td>
<td>0.917</td>
<td>0.898</td>
</tr>
<tr>
<td>2.0</td>
<td>0.922</td>
<td>0.903</td>
<td>0.882</td>
<td>0.862</td>
<td>0.785</td>
</tr>
</tbody>
</table>


Thus, if \( t \) gets very large, it is reasonable to suggest that \( x \) be considered for inclusion in the first group of variables.

Cochran considers the four major methods for handling variables in group 1. These are (1) matching, (2) frequency-matching, (3) stratified analysis, and (4) covariance analysis. Cochran believes that pair-matching removes all the bias "provided that the relation between \( y \) and
x is the same in both populations and that x is measured without appreciable error". He finds that methods 2 and 3 are less effective in removing the bias since the distributions of the x's in the two populations are only partially equated. To compare the three methods, he assumes that $x \sim N(0,1)$ in population 1 and $x \sim N(\mu,1)$ in population 2. For Table 1.7.5 below, he considered $\mu = -\frac{1}{2}, \mu = -1$ and constructed the boundaries of his categories so that there is an equal proportion of population 1 in each class. As can be seen, Method 2 (frequency-matching) is superior to Method 3 (stratified analysis) in removing bias.

TABLE 1.7.5c

<table>
<thead>
<tr>
<th>No. of Subclasses</th>
<th>$\mu = -\frac{1}{2}$</th>
<th>$\mu = -1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method 2</td>
<td>Method 3</td>
</tr>
<tr>
<td>2</td>
<td>0.184</td>
<td>0.190</td>
</tr>
<tr>
<td>3</td>
<td>0.105</td>
<td>0.109</td>
</tr>
<tr>
<td>4</td>
<td>0.071</td>
<td>0.076</td>
</tr>
<tr>
<td>5</td>
<td>0.052</td>
<td>0.055</td>
</tr>
</tbody>
</table>

cIbid. Adopted from Table 3.2.

Method 4 (covariance analysis) removes all bias under the same assumptions as above. Using a result from his 1954 paper, he indicates that the relative efficiency of this method to pair-matching is 0.94 and 0.8 when $\mu = -\frac{1}{2}$ and 1, respectively. This is assuming that both methods use the same sample size and that the pair-matching is exact.

Cochran (1968) considers the ability of a categorization scheme on x to remove bias in comparisons due to different distributions in x and
compares the categorization schemes to covariance analysis to compare ability to reduce variance of the estimates. For the first objective, he considers various distributions for the treatment and control group when the distribution of the two groups differs by a shift in location. The results of his analysis showed that the optimum categorization scheme had a large percentage of the observations in the central categories and fewer in the tail categories. This was shown to be slightly better at removing the bias than a categorization scheme with equal proportions in all categories. The increase in percent bias eliminated was greatest for the change from 2 to 3 categories and become smaller with each increase of category number. For practical purposes, the bias removed at 4 categories was not much less than that at 6 categories (88.2% and 94.2% for normal distributions). This is similar to the results of Billewicz who showed 3 categories as having minimum variance. His comparison of percent reductions in variance between stratified matching and covariance analyses for normal distributions showed that covariance analysis was always better than that of a categorization scheme and that this difference increased as the correlation increased.

Cochran (1969) has considered the issue of how to estimate the $\beta$ in the covariance regression. It is assumed that the populations can be described by the model:

\[ y_{1j} = \mu + \alpha_j + \beta x_{1j} + \epsilon_{1j} \quad j = 1, \ldots, n \]

\[ y_{0j} = \mu + \beta x_{0j} + \epsilon_{0j} \quad j = 1, \ldots, n \]

It is assumed that $\sigma^2_{x_1} = \sigma^2_{x_0}$. He then considers three cases for $\alpha_j$, the parameter of interest. In the first case, it is assumed that
\[ \alpha_j = \alpha \nu_j \] and he recommends that \( \beta \) be estimated from the pooled data. As his second case \( \alpha_j = \alpha + \epsilon_j \), so that \( \alpha_j \) has a random component. He shows that for sufficiently large \( n_1, n_0 \), \( \beta \) may be estimated from the pooled data, as above, as long as both \( x \)'s have the same expected value. When they do not, he recommends that a weighted sum of the regression coefficients for each group, the weight being proportional to the inverse of the variances. In his third case, \( \alpha_j = \alpha + \delta x_{ij} + \epsilon_j \), hence, the treatment effect is linearly related to the matching variable. His recommended estimate of the treatment effect in this case is \( \alpha + \delta \bar{x} \), and he feels this can be best estimated by using the data for controls to estimate the common linear trend.

Cochran's idea of studying the percent reduction in bias was used extensively by one of his students, Rubin (1970; 1973a, b; 1976a, b). The 1970 reference is his dissertation and the work done there has been published in the other references, which we will consider here. Rubin (1973a) used Monte Carlo techniques to investigate the percent reduction in bias of \( (\bar{y} - \bar{y}') \) by mean-matching and several pair-matching techniques. Under the assumption of the same linear relationship between \( x \) and \( y \) in both populations and that \( x \) was normal in both populations, he found that mean-matching removed most of the bias in \( (\bar{y} - \bar{y}') \). This matching was done using a control reservoir that was limited to a certain size (2,3,4 \( x \) the case sample). However, Rubin did not spend much time on this technique. After showing the work just mentioned, he makes the comment "since an investigation rarely knows for sure that the response surface is linear, if the choice is between exactly pair-matched samples and exactly mean-matched samples of the same size, obviously, he would choose the exactly pair-matched samples" and then considers only pair-matched samples. It
should be noted that he is assuming that both samples are of the same size (and also the matches are exact) when he makes this comparison. He compares several simple pair-matching methods to find the method which gives the best reduction.

In his second paper (1973b), Rubin examined regression adjusted estimators on pair-matched samples. He assumed either an exponential response surface or quadratic surface and considered estimators that used a linear adjustment. As before, this was done using Monte Carlo simulation techniques. He found that there was usually little gain to be had by using matched pairs, except when the variance of the x's were quite different or the distributions were not symmetric. In this case, there was a definite gain in bias reduction by using the matched samples.

Cochran and Rubin (1973) published a review of their work on bias in observational studies. A good deal of the paper concerned results that are considered elsewhere (Cochran 1965, 1968, 1969; Rubin 1970, 1973, 1976). There are several sections in this paper that we will discuss here. In Section 3.5 of their paper, the authors address the problem of omitted confounding variables. They show that if the response is a linear function of both the confounding variable, x, and the omitted confounder, z, that matching on the x without regard to the z may result in even greater bias in \(\frac{\bar{y}_1 - \bar{y}_2}{\bar{z}_1 - \bar{z}_2}\) than that of the original. A second point that they make is that if z is a linear function of x then the linear regression of y on x will remove all the bias due to z while the pair-matched samples will not. In Section 5, a multivariable generalization of Rubin's work is presented. This section concerns itself only with the theory and not with any Monte Carlo work. An interesting point that comes about from this is that even though each variable may contribute
some bias, the total bias may be zero, since the individual biases are multiplied by their regression coefficients and added to obtain the total bias.

The authors also consider several multivariable matching methods. Two methods, caliper-matching and categorical-matching are straightforward generalizations of univariate techniques. These methods are considered in conjunction with a transformation that produces uncorrelated variables. Other methods which are means of summarizing the variables into a univariate score are also considered. The first of these, discriminant matching, was considered in Rubin (1970) and later in Rubin (1976a). They suggest that a linear discriminant function be formed using the estimated variance covariance matrix from the control population. The performance of this method can then be judged using the tables presented in Rubin (1973a). This approach to using the linear discriminant differs from that recommended by Miettinen (1976) (below). They also suggest the use of metrics, such as the Euclidian distance between the case and control (standardized) variables (see Smith (1976)). Nothing is known about the properties of these methods.

MacKinley (1974) investigated the expected number of matches, and the variance for a categorical model. She used models with 10 and 20 categories and identical and different distributions in the two populations and various sample sizes for the two populations. She found that the expected number of matches increased as the sample size increased in both populations and when the sample size in the control populations was increased relative to the size of the case population. The expected number of matches for dissimilar populations was less than that for identical populations, and decreased as the number of categories increased. She
recommends that the control reservoir should be at least five times as large as the case reservoir when there is a small number of categories; and when the number of categories is large (≥ 100), the ratio of sample sizes should be 10 to obtain near 100% matching. When the two sample sizes are comparable and small, she finds that there will be a 50% loss of potential samples.

Rubin (1976a) considered the properties of matching methods which are "equal percent bias reducing" (EPBR) for multivariate $\tilde{x}$. The author feels methods that are EPBR are "appealing basically because one obtains the same percent reduction in bias for any linear function of the $\tilde{x}$, if and only if, the matching method is EPBR for $\tilde{x}$". Rubin points out that if the response function is known for univariate $y$ or if $\beta_i(\bar{X}_{i1} - \bar{X}_{i2}) \gg \beta_i'(\bar{X}_{i1}' - \bar{X}_{i2}')$, $i \neq i'$, for multivariate $y$, then this method is not as attractive. However, for the situation where the response function is not known, then EPBR methods are quite attractive. He enumerates cases for which the matching is EPBR, then suggests a transformation that will allow a normal sample to fulfill these requirements. This is the linear discriminant function mentioned above. Rubin finds that for ellipsoidal $X$, the estimated linear discriminant function is EPBR. He also shows that the use of certain metrics to define the distance between two observations will be EPBR.

Rubin (1976b) considers the maximum percent reduction in bias attainable for EPBR methods. He proves a theorem showing the maximum percent reduction in bias, $\theta_{\text{max}}^*$, given fixed distributions in $X$, fixed case and control pools from which the matches are to be selected, and fixed final sample sizes. If we define $r_i$ as the ratio of the pool size to the sample size in the $i$th group and $\sigma_i^2$ as the variance of the linear discriminant,
calculated using the estimated variance-covariance matrix from the 2nd group, in the i-th group, he finds the following trends in $\theta_{\text{max}}$:

1. "... everything else being equal, the further apart the means of X in the populations, the more difficult it is to obtain well matched samples."
2. as $r_1$ increases, the better the matches will be.
3. "one gains less and less from successive increases in the ratios $r_1$ and $r_2$." 
4. "the larger the samples, the better the matches."
5. for fixed sample sizes when $r_1 = 1$, $\theta_{\text{max}}$ decreases as $\sigma_1^2 / \sigma_2^2$ increases.

He then makes some suggestions on how to estimate the components of $\theta_{\text{max}}$ and on how to use these estimates to suggest minimal sample sizes. Nothing is said about either the particular methods to be used or the use of regression adjustments on the samples.

Miettinen (1976a, 1976b) has also considered the use of the linear discriminant as a means of matching. His method is different from that expressed in Cochran and Rubin (1973). Miettinen's function uses the pooled variance-covariance matrix for the covariables and the response. Cochran and Rubin (1973) use only the estimated variance-covariance matrix of the control group covariable to estimate the function. Miettinen then computed the score of each observation, arbitrarily giving all observations the same response for scoring purposes. The scores are then used to implement the matching plans and may be used in fitting a model to the response. The use of the response in formulating the function is somewhat questionable on several points. First it assumes that the
investigator has already gathered the response, an assumption that is not made elsewhere. Secondly, the use of the response in the formulation of criteria which will be used to analyze the response is questionable in itself. Thirdly, the presence of the response in the function may diminish the importance of a highly correlated confounder in the analysis. No theoretic or Monte Carlo investigation is made on the properties of this method; rather, it is given with the comment that "it works".

1.8. **Summary and Outline of Work**

For the most part, there has been little analytic study of the distributions of pair-matched variables. Cochran (1953) has considered the case of exact pair-matching, in which case both variables have the same distribution. He has also examined frequency-matching and stratified analysis (Cochran (1968)) to provide some rules for the choice of category structure. Little work has been done on non-standard situations, e.g. non-parallel slopes in the Anova model. The most extensive studies done have been Monte Carlo studies. Billewicz (1965) examined the variance of several category structures when the matched variable was identically distributed in both populations; and, Rubin (1970, 1973a,b) examined the bias reduction abilities of certain mean-matched and pair-matched techniques.

The studies of matching versus linear models are generally in favor of linear models, at least for the case of univariate covariance analysis, which has been extensively examined. Cochran (1954) said that exact pair-matching was always more efficient that covariance analysis of random samples of the same size. Billewicz (1965) found the reverse to be true for categorical pair-matching when the matching variable has the
same distribution in both groups. He did not consider the case where the variables were not identically distributed in the two populations. However, some results for a categorical-matching variable with a different distribution in the two groups suggests that pair-matching would be less efficient than covariance analysis, and that the covariance analysis of matched samples would be more efficient than covariance analysis of random samples. This agrees with the results of Greenberg (1953) who showed that covariance analysis of mean-matched samples to be more efficient than the same analysis of random samples. Cochran (1968) found that a covariance analysis is more efficient than a frequency-matched sample or a stratified analysis. Rubin (1970; 1973a,b) found that a covariance estimator was better at reducing bias than the pair-matched estimator for quadratic or exponential response surfaces. Thus, it seems that when covariance analysis is the correct model and the variable is identically distributed in both populations then analysis of covariance is to be preferred. Little is known about other situations, although it is suggested that a linear model would also be preferred.

Little work has been done on multivariable-matching techniques. Cochran and Rubin (1973) discuss some generalizations of the techniques of caliper-and categorical-matching. Rubin (1976a,b) has presented some results on bias reduction for the class of methods that are equal percent bias reducing (EPBR). He considers the use of a linear discriminant function to give univariate variables for matching. He also considered the use of certain inner product matrices useful in matching to minimize certain distances between the two groups. Little is known about the performance of such matching methods. Miettinen (1976) has also suggested the use of discriminant matching, although his discriminant differs substantially from that of Rubin.
CHAPTER II
CATEGORY MATCHING OF CONTINUOUS VARIABLES

2.1. Introduction

Pair-matching and the analysis of pair-matched data present a particular problem when the matching variables are continuous. In this case, there is no obvious matching criterion as there is when the matching variable is discrete. If an investigator desires to use a pair-matching plan, either a categorization plan must be forced upon the matching variable so that the discrete variable matching methods can be used, or some criterion must be set for determining that two values are "close enough" to be considered as matching values. Categorical pair-matching, caliper-matching, and nearest-neighbor-matching are three widely used techniques that are available to pair-match observations when considering continuous data. Although there has been considerable Monte Carlo study of these techniques, there has been little analytical work regarding these methods and their properties. In this chapter and in the following ones, we shall present some results of an analytical study of the three techniques.

In this chapter, we shall consider categorical matching schemes. Cochran (1968) has investigated some cases of this and has recommended certain procedures. Billewicz (1965) has also done some Monte Carlo study of the method. We will review the results of these papers and extend them in several directions. First, we will examine the post-sampling distribution of a variable that has been sampled according to some categorization scheme. Next, we will consider the sample sizes necessary to fulfill
certain categorical-matching-scheme requirements. Finally, we will compare the categorical methods of analysis with the more usual analysis of covariance when the response variable is also continuous. In this and the following work, our frame of reference shall be the case-control study.

2.2. The Technique of Categorical Pair-Matching

When category pair-matching is to be used in a case-control study, the following procedure is often adopted. One group (which is most often the set of cases) is classified, on the basis of the value of the variable being used for matching, so that each observation falls into one of r mutually exclusive and exhaustive categories. For each case observation in a given category, the data on the second group, the controls, are examined until one observation is found which also belongs to that category. Then, on the basis of common membership in a given category, the two observations are declared to be "matched". This searching can, of course, be done for all categories simultaneously, so that when it is determined to which category a control observation belongs, an unmatched case observation from the same category is paired with it. This matching procedure is continued until all observations in the case group have been matched, or, when there are a limited number of controls, as many cases as possible have been matched.

As can be seen from the method described above, the pairing of observations is arbitrary once a common category has been established, unlike a "natural" pairing (as between twins or siblings) where it might be expected that the response variables would be correlated within the pair. This artificial pairing might be required for certain types of "matched" pairs analyses, but we shall ignore such artificial pairing of
observations within categories in further discussion. This method will then be referred to as category-matching or frequency-matching. This method selects the observation so that each group has the same number of observations (or a common multiple thereof) in each category.

These schemes are used under the assumption that the response variable varies with, or is correlated with, the matching variable. The common aims of these methods are both to reduce the bias in the estimators of the response variable's case-control difference and to increase the precision of these estimators. Since the response variable is assumed to vary with the matching variable, estimators of the case-control difference can be biased if the matching variable has a different distribution in the two groups. Thus, the matching, which tends to make the distribution similar, would reduce this bias. Likewise, the matching reduces the variance of the matching variable, thus reducing the variance of estimators of the response variable and increasing the precision of related tests as compared to random sampling. As we mentioned above, this depends on the assumption that the response variable is related to the matching variable, since, otherwise, a random matching can actually serve to decrease the precision. This was shown by Miettinen (1968) where the estimate from an unnecessary matching procedure had a greater variance than an estimate from an unmatched analysis.

The categorization scheme can arise from two general sources. It can be based on some set of external standards, such as a standard million based on the U.S. population. A second set of standards would be those arising from the distributions of the matching variable in the two groups. Not much beyond description of the results can be said about the external standards. However, the internal standards can be chosen in
such a way as to minimize the bias or some other property of the estimator. Cochran (1965, 1968) considered techniques of categorization based on what we have referred to as internal standards. Under the assumption that the response variable was linearly related to the matching variable, he found an approximation for the reduction in bias due to matching relative to the original, unmatched bias. He also found the optimal categorization structure for a given number of categories when the matching variable was distributed as a standard normal in both populations. He also investigated the effect of category matching on the average within category variance and used these results to verify some findings of Billewicz (1965). Both of these papers were referred to in the literature review and shall be considered in greater detail in the following sections.

Now, to examine the effects of category matching on the post-sampling distribution of the matching variable, we make the following assumptions:
Let the distribution of the matching variable \( X \) in the \( i \)-th population (\( i = 0 \) for controls, \( i = 1 \) for cases) be denoted by \( f_i(x) \) with mean \( \mu_{X_i} \) and variance \( \sigma_{X_i}^2 \). For a given categorization scheme with \( r \) categories, let \( A_1 < A_2 < \ldots < A_{r+1} \) be the bounds of the categories, the \( k \)-th category \( C_k \) having lower boundary point \( A_k \) and upper boundary point \( A_{k+1} \). The probability that an observation from the \( i \)-th group falls into the \( k \)-th category is

\[
\pi_{ik} = \frac{A_{k+1} - A_k}{\int_{A_k}^{A_{k+1}} f_i(x) \, dx}.
\]

If we consider the distribution of \( X_i \) within a category, it is clear that we are considering a truncated version of the original distribution.
Then, the distribution of the truncated random variable $X_{ik}$ associated with $C_k$ is

$$g_{ik}(x) = \begin{cases} 
  f_1(x)/\pi_{ik} & x \in C_k \\
  0 & x \notin C_k
\end{cases} \quad (2.2.1.)$$

$$= I_{[x \in C_k]} f_1(x)/\pi_{ik}$$

The mean and variance of $X_{ik}$ are defined in the usual manner:

$$\mu_{ik} \equiv E(X_{ik}) = \int_{A_k} x \, g_{ik}(x) \, dx, \quad (2.2.2.)$$

$$\sigma^2_{ik} \equiv \text{Var}(X_{ik}) = \int_{A_k} (x-\mu_{ik})^2 g_{ik}(x) \, dx \quad (2.2.3.)$$

When the $X$'s are sampled, they are chosen so that some proportion $p_k$ are in $C_k$. How the $p_k$'s are determined will depend on the type of matching scheme. If the matching scheme is based solely on the distribution $f_1(x)$ of $X_1$, for example, then $p_k = \pi_{1k}$; throughout most of discussion, $p_k = \pi_{1k}$. Now the post-sampling distribution of the $X$'s is composed of the weighted sum of the within category distributions, which by (2.2.1.) is

$$g_1(x|\tilde{A}) = \sum_{k=1}^{r} p_k I_{[x \in C_k]} g_{ik}(x) \quad (2.2.4.)$$

where $\tilde{A} = (A_1, A_2, \ldots, A_{r+1})$. Thus, the post-sampling mean of $X_1$ is
\[
E(X_{\perp} | r, A) = \int_{A_1} x g_{1}(x | r, A) dx = \int_{A_1} x \left( \sum_{k=1}^{r} p_k I_{[x \in C_k]} g_{1k}(x) \right) dx
\]
\[
= \sum_{k=1}^{r} p_k \int_{A_1} x I_{[x \in C_k]} g_{1k}(x) dx
\]
\[
= \sum_{k=1}^{r} p_k \int_{A_k} x g_{1k}(x) dx = \sum_{k=1}^{r} p_k \mu_{1k} ; \quad (2.2.5.)
\]

And, the Post-sampling variance of \( X_{\perp} \) is

\[
V(X_{\perp} | r, A) = E(X_{\perp}^2 | r, A) - (E(X_{\perp} | r, A))^2 ,
\]

where \( E(X_{\perp}^2 | r, A) = \int_{A_1} x^2 g_{\perp}(x | r, A) dx \)

\[
= \sum_{k=1}^{r} p_k \int_{A_k} x^2 g_{1k}(x) dx
\]
\[
= \sum_{k=1}^{r} p_k (\sigma_{1k}^2 + \mu_{1k}^2) = \sum_{k=1}^{r} p_k \sigma_{1k}^2 + \sum_{k=1}^{r} p_k \mu_{1k}^2
\]
so

\[ V(X_1 \mid r, A) = \sum_{k=1}^{r} p_k \sigma_{1k}^2 + \sum_{k=1}^{r} p_k \mu_{1k}^2 - \left( \sum_{k=1}^{r} p_k \sigma_{1k}^2 + \sum_{k=1}^{r} p_k \mu_{1k}^2 \right)^2 \]

(2.2.6.)

The first summation on the R.H.S. of (2.2.6.) is the "average within category variance". We note in passing that if \( p_k = \pi_{ik} \), \( A_1 = -\infty, A_{r+1} = \infty \), the post-sampling distribution is identical to the original distribution since

\[ g_i(x \mid r, A) = \sum_{k=1}^{r} p_k [x \in C_k] g_{ik}(x) \]

\[ = f_i(x) . \]

Hence

\[ E(X_1 \mid r, A) = \int_{A_1}^{|A_{r+1}|} x g_i(x \mid r, A) dx \]

\[ = \mu_{x_1} ; \]

and

\[ V(X_1 \mid r, A) = \int_{A_1}^{|A_{r+1}|} (x - \mu_{x_1})^2 g_i(x \mid r, A) dx \]

\[ = \int_{-\infty}^{\infty} (x - \mu_{x_1})^2 f_i(x) dx \]

\[ = \sigma_{x_1}^2 . \]
This indicates that when the categorization scheme is based on the
distribution of the matching variable in one of the two groups (i.e.,
\( p_k^1 = \pi_k \) or \( p_k^0 = \pi_k \)), the categorization has no effect on that distribution.
Also, if the distributions are identical for both groups, then neither
of them is changed by the sampling.

The formulae developed above can be applied to many types of dis-
tributions, including the normal, Gamma and Beta distributions. In the
next section, we shall consider, in detail, the effects of category-
matching on normal variates.

2.3. Category Matching for Normally Distributed Matching Variables

The results of the last section will be applied to the normal dis-
tribution. Explicit expressions will be developed from the general mo-
ments given in the preceding sections. The work of Cochran (1968) will
also be considered in greater detail.

In what follows, we will have use for the following results from
distributed as a normal variable with mean \( \mu_X \) and variance \( \sigma_X^2 \), is trun-
cated below at \( A \) and above at \( B \), then the mean and variance of the trun-
cated distribution are

\[
\mu_X + \sigma_X \left\{ z \left( \frac{A-\mu_X}{\sigma_X} \right) - z \left( \frac{B-\mu_X}{\sigma_X} \right) \right\}.
\]

\[
\phi \left( \frac{B-\mu_X}{\sigma_X} \right) - \phi \left( \frac{A-\mu_X}{\sigma_X} \right), \tag{2.3.1}
\]
and

\[
\sigma_x^2 \left\{ \frac{1 + \left( \frac{A - \mu_x}{\sigma_x} \right) Z \left( \frac{A - \mu_x}{\sigma_x} \right) - \left( \frac{B - \mu_x}{\sigma_x} \right) Z \left( \frac{B - \mu_x}{\sigma_x} \right)}{\phi \left( \frac{B - \mu_x}{\sigma_x} \right) - \phi \left( \frac{A - \mu_x}{\sigma_x} \right)} \right\}^2,
\]

respectively, where

\[
Z(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}
\]

and

\[
\phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt.
\]

We can now use (2.3.1.) and (2.3.2.) to evaluate (2.2.2.) and (2.2.3.) since each pair of category bounds defines a truncated normal distribution. For notational convenience, let

\[
\xi_{i,k} = \frac{(A_k - u_i) / \sigma_i}{x_i} i=0,1, k=1,2,\ldots,r.
\]
Then

\[ \mu_{1k} = \mu_{x_1} + \sigma_{x_1} \left\{ \frac{Z(\xi_{1,k}) - Z(\xi_{1,k+1})}{\phi(\xi_{1,k+1} - \phi(\xi_{1,k}))} \right\} = \mu_{x_1} + \sigma_{x_1} J_{1,k}, \]

(2.3.3.)

and

\[ \sigma^2_{1k} = \sigma^2_{x_1} \left\{ 1 + \frac{\xi_{1,k} Z(\xi_{1,k}) - \xi_{1,k+1} Z(\xi_{1,k+1})}{\phi(\xi_{1,k+1}) - \phi(\xi_{1,k})} - J^2_{1,k} \right\} \]

\[ = \sigma^2_{x_1} \{ 1 + H_{1,k} - J^2_{1,k} \} \]

(2.3.4.)

Note that \( \pi_{ik} = \Phi(\xi_{1,k+1}) - \Phi(\xi_{1,k}) \). So, the post-sampling distribution of \( X_{1i} \) (i.e., the distribution after categorization) has mean

\[ E(X_{1i}|r,A) = \sum_{k=1}^{r} p_k \mu_{1k} = \mu_{x_1} + \sigma_{x_1} \sum_{k=1}^{r} p_k J_{1,k}, \]

(2.3.5.)

and variance

\[ V(X_{1i}|r,A) = \sum_{k=1}^{r} p_k (\sigma^2_{1k} + \mu^2_{1k}) - \left( \mu_{x_1} + \sigma_{x_1} \sum_{k=1}^{r} p_k J_{1,k} \right)^2 \]

\[ = \sum_{k=1}^{r} p_k \left[ \sigma^2_{x_1} (1 + H_{1,k} - J^2_{1,k}) + \mu^2_{x_1} + 2\mu_{x_1} \sigma_{x_1} J_{1,k} + \sigma^2_{x_1} J^2_{1,k} \right] \]
\[
- \mu_{x_1}^2 - 2 \mu_{x_1} \sigma_{x_1} \sum_{k=1}^{r} p_k J_{i,k} - \sigma_{x_1}^2 \left( \sum_{k=1}^{r} p_k J_{i,k} \right)^2 \\
= \sigma_{x_1}^2 \left[ 1 + \sum_{k=1}^{r} p_k H_{i,k} \right] + \mu_{x_1}^2 + 2 \mu_{x_1} \sigma_{x_1} \sum_{k=1}^{r} p_k J_{i,k} \\
- \left( \mu_{x_1}^2 + 2 \mu_{x_1} \sigma_{x_1} \sum_{k=1}^{r} p_k J_{i,k} + \sigma_{x_1}^2 \sum_{k=1}^{r} p_k J_{i,k} \right)^2 \\
= \sigma_{x_1}^2 \left[ 1 + \sum_{k=1}^{r} p_k H_{i,k} \right] - \sigma_{x_1}^2 \left[ \sum_{k=1}^{r} p_k J_{i,k} \right]^2 \\
= \sigma_{x_1}^2 \left\{ \sum_{k=1}^{r} p_k \left[ 1 + H_{i,k} - J_{i,k}^2 \right] - 2 \sum_{k=1}^{r} p_k J_{i,k} J_{i,k} \right\} \quad (2.3.6.)
\]

The expected value of \((x_1 - x_0)\) based on random samples from \(f_1(x)\) and \(f_0(x)\) is \(\mu_{x_1}^2 - \mu_{x_0}^2\). The expected value of \((x_1 - x_0)\) based on a category-matched sample is, from (2.3.5.), \((\mu_{x_1}^2 - \mu_{x_0}^2) + \sum p_k (J_{1,k} \sigma_{x_1} - J_{0,k} \sigma_{x_0})\).

Thus, the reduction in mean difference is \(- \sum p_k (J_{1,k} \sigma_{x_1} - J_{0,k} \sigma_{x_0})\). If we divide this reduction by the initial mean difference, we arrive at a term for the reduction in mean difference due to category matching expressed as a proportion of the actual difference, namely
This term will be hereafter referred to as the "proportional reduction in mean difference".

If we assume that the response variable is related to x by the linear model \( y = \alpha + \beta x + e \), the bias of \( \bar{y}_1 - \bar{y}_0 \) as an estimator of \( (\alpha_1 - \alpha_0) \) is \( \beta(\mu_{x_1} - \mu_{x_0}) \). By inspection, we can see that the "proportional reduction is bias" due to category matching on the x's is the same as \( (2.3.7.) \). Now, we can assume, without loss of generality, that \( x_1 \sim N(0,1) \).

If we further assume that \( p_k = \pi_{1k} = \Phi(A_{k+1}) - \Phi(A_k) \), then

\[
\sum_{k=1}^{r} p_k J_{1,k} = \sum_{k=1}^{r} \pi_{1,k} (Z(A_k) - Z(A_{k+1}))/\pi_{1,k} = Z(A_1) - Z(A_{r+1}) .
\]

A necessary and sufficient condition for this to be zero is that \( A_1 = -A_{r+1} \) (or, equivalently, for general \( \mu_{x_1} \) and \( \sigma^2_{x_1} \), that \( \xi_{1,1} = -\xi_{1,r+1} \)).

i.e., if the first and last truncation points are symmetric about the mean. A special case of this is when \( A_1 = -\infty \), \( A_{r+1} = \infty \), so that the entire range of \( X_1 \) is included in the categorization, as it is in the categorization schemes we are studying. Under these assumptions that \( X_1 \sim N(0,1) \), \( p_k = \pi_{1k} \), and \( A_1 = -A_{r+1} \), \( (2.3.7.) \) becomes
\[ \frac{\sigma_{x_0}}{-\mu_{x_0}} \sum_{k=1}^{r} \pi_{1,k}^j \pi_{0,k}^j \]  

(2.3.8.)

Now, when \( \mu_{x_0} \) is zero (i.e., \( \mu_{x_0} = \mu_{x_1} \)), this term is undefined, as we would expect, since there is no mean difference to remove via category-matching.

Cochran (1968) assumes a linear model of the same form given above and studied the proportional reduction in bias for category-matching. In general, the proportional reduction in bias would be

\[ 1 - \sum_{k=1}^{r} \frac{p_k (\mu_{1k} - \mu_{0k})}{(\mu_{x_1} - \mu_{x_0})} . \]  

(2.3.9.)

Cochran used a Taylor series expansion to develop an approximation for this expression. Under the assumption that the distributions of \( X_1 \) and \( X_0 \) differ only by the value of a single parameter which takes the value \( \theta \) in \( f_0(x) \) and 0 in \( f_1(x) \), he found an approximation for (2.3.9.). Under the further assumption that \( X_1 \) and \( X_0 \) were i.i.D.N(0,1), he found that the proportional reduction in bias is

\[ M = \sum_{k=1}^{r} \frac{(Z(A_k) - Z(A_{k+1}))^2}{\pi_k} , \]  

(2.3.10.)

where \( \pi_k = \pi_{1,k}^j = \pi_{0,k}^j \) since \( X_1 \sim N(0,1) \).
This derivation can be found on pp. 302-303 of his paper. This approximation does not correspond to the result derived in (2.3.8.). Cochran does, however, note that this expression has been derived elsewhere (for a different purpose). Cox (1957) arrived at a similar expression when he studied the information lost by grouping continuous variables, i.e., by categorizing them. We can also arrive at the same expression by attempting to minimize the average within category variance.

From (2.2.6.) the average within category variance $\overline{\sigma}^2_i$ of $X_i$ is

$$\overline{\sigma}^2_i = \sum_{k=1}^r p_k \sigma^2_{ik}.$$ 

Now, if we specify that $X_i \sim N(\mu_{x_i}, \sigma^2_{x_i})$ and that $p_k = \pi_{i,k}$, then, from (2.3.4.),

$$\overline{\sigma}^2_i = \sum_{k=1}^r \pi_{ik}(1 + H_{i,k} - J_{i,k}) \sigma^2_{x_i}$$

$$= \sigma^2_{x_i} \sum_{k=1}^r \pi_{ik} \left\{ 1 + \left( \frac{\xi_{i,k} Z(\xi_{i,k}) - \xi_{i,k+1} Z(\xi_{i,k+1})}{\pi_{i,k}} \right) - \left( \frac{Z(\xi_{i,k}) - Z(\xi_{i,k+1})}{\pi_{i,k}} \right)^2 \right\}$$

$$= \sigma^2_{x_i} \left\{ 1 + \sum_{k=1}^r \pi_{ik} \left( \frac{Z(\xi_{i,k}) - Z(\xi_{i,k+1})}{\pi_{i,k}} \right)^2 \right\}$$
Now, as $\xi_{i,1} \to -\infty$, $\xi_{i,1}Z(\xi_{i,1}) \to 0$ by L'Hôpital's rule. Likewise, for $\xi_{i,r+1} \to \infty$, $\xi_{i,r+1}Z(\xi_{i,r+1}) \to 0$. Thus, for $A_i \to -\infty$ and $A_{r+1} \to \infty$,

$$
\frac{\sigma^2_i}{\pi_i} = \sigma^2_{xi} \left\{ 1 - \sum_{k=1}^{r} \frac{(Z(\xi_{i,k}) - Z(\xi_{i,k+1}))^2}{\pi_{ik}} \right\} \quad \text{(2.3.11.)}
$$

To minimize this expression, we need to maximize

$$
\sum_{k=1}^{r} \frac{(Z(\xi_{i,k}) - Z(\xi_{i,k+1}))^2}{\pi_{ik}}
$$

This reduces to Cochran's expression (2.3.10.) when we assume $X_i \sim N(0,1)$, since then $\xi_{i,k} = A_k$. Thus, the term Cochran has derived is not the "proportional reduction in bias", but is the proportional reduction in average within category variance.

Cox (1957) found that the average loss of information, $L$, say, due to grouping was equal to $L = (1-M)$ when the variable being grouped, $Z$, say, was $N(0,1)$. From (2.3.11.), we can see that $\sigma^2_i = L$ when $X_i \sim N(0,1)$. Since there is no bias to reduce when $X_i \sim N(0,1)$, it seems reasonable to choose the intervals so as to minimize the average within category variance, or equivalently the loss of information on the matching variable. Cox found the boundary points of the categories giving the maximum values of $M$ for $r = 2, 3, 4, 5, 6$ for intervals which are symmetric around the mean, which in this case is zero. These intervals, the values of
TABLE 2.3.1

Minimum $\hat{\sigma}^2$ Intervals, Symmetric About Zero, for $X_1 \sim N(0,1)$

<table>
<thead>
<tr>
<th>No. of Categories</th>
<th>Boundary Points</th>
<th>$\pi_k$</th>
<th>$M_{max}$</th>
<th>M for Equiprobable Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$-\infty$ 0.0</td>
<td>$\infty$ - - - - .5 .5 - - - -</td>
<td>.637</td>
<td>.637</td>
</tr>
<tr>
<td>3</td>
<td>$-\infty$ -0.61</td>
<td>0.61 $\infty$ - - - - .27 .46 .27 - - - -</td>
<td>.810</td>
<td>.793</td>
</tr>
<tr>
<td>4</td>
<td>$-\infty$ -0.98</td>
<td>0.0 0.98 $\infty$ - - - - .16 .34 .34 .16 - - - -</td>
<td>.882</td>
<td>.861</td>
</tr>
<tr>
<td>5</td>
<td>$-\infty$ -1.23</td>
<td>-0.40 0.40 1.23 $\infty$ - - .11 .24 .30 .24 .11 - -</td>
<td>.920</td>
<td>.897</td>
</tr>
<tr>
<td>6</td>
<td>$-\infty$ -1.45</td>
<td>-0.66 0.0 0.66 1.45 $\infty$ .07 .18 .25 .25 .18 .07</td>
<td>.942</td>
<td>.919</td>
</tr>
</tbody>
</table>

Adapted from Cox, (1957), p. 545.
\( \pi_k \), and the value of \( M_{max} \) are presented in Table 2.3.1. The last column shows the value of \( M \) for intervals that are equiprobable (i.e., \( \pi_k = 1/r \)). A portion of this table was used by Cochran (1968, pg. 304). He entitled the columns of \( M \) as the percent reduction in bias. By comparison of the last two columns, we can see that the choice of boundary points is not very critical. There is a definite improvement in \( M \) by increasing \( r \) from two to three. Further increases in \( r \) are not beneficial; note that the change in \( M \) seems to be halved for each such increase in \( r \). So, either a three or four category-matching scheme would seem to be sufficient to obtain good reductions in the within category variance. The fact that the values for the equiprobable intervals are not that different from the "optimum" intervals suggests that, from the standpoint of convenience, this type of interval be used. Cochran also considered equiprobable categories for several other distributions and found the same patterns occurring.

Neither author has given much attention to the performance of the equiprobable category structure when the two distributions differ in their parameters. Recalling the definitions of confounding given in the last chapter, it would seem reasonable to expect that this situation would occur frequently in practice. When this occurs, we would not expect a categorization scheme based on the case group to perform as well when the distribution of the matching variable is not the same in the two groups. Table 2.3.2 on the following page shows the proportional reduction in the mean of \( \mu_{x_0} \) and one minus the ratio of the average within category variance to \( \sigma^2_{x_0} \) when \( x_0 \) has been matched to a \( N(0,1) \) variable by three and four categories that are equiprobable for the \( N(0,1) \). The weights \( p_k \) are equal to \( 1/r \) \((r = 3, 4)\). The proportional
TABLE 2.3.2

Results of Matching a $N(\mu_{x_0}, \sigma_{x_0}^2)$ Variable by an Interval Structure that is Equiprobable for an $N(0,1)$ Distribution

<table>
<thead>
<tr>
<th>Proportion Reduction in $\mu_{x_0}$</th>
<th>$1-\sigma_{x_0}^2/\sigma_{x_0}^2$</th>
<th>Relative Reduction</th>
<th>$r=3$</th>
<th>$r=4$</th>
<th>$r=3$</th>
<th>$r=4$</th>
<th>Relative Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{x_0}$</td>
<td>$X_{x_0}$</td>
<td>r=3</td>
<td>r=4</td>
<td>r=3</td>
<td>r=4</td>
<td>r=3</td>
<td>r=4</td>
</tr>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.7932</td>
<td>0.8606</td>
<td>0.9218</td>
</tr>
<tr>
<td>3.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.7847</td>
<td>0.8502</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>0.7928</td>
<td>0.8602</td>
<td>0.9216</td>
<td>0.7919</td>
<td>0.8596</td>
<td>0.9211</td>
</tr>
<tr>
<td>3.0</td>
<td>0.7846</td>
<td>0.8500</td>
<td>0.9230</td>
<td>0.7842</td>
<td>0.8498</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>0.7914</td>
<td>0.8593</td>
<td>0.9201</td>
<td>0.7877</td>
<td>0.8567</td>
<td>0.9194</td>
</tr>
<tr>
<td>3.0</td>
<td>0.7841</td>
<td>0.8497</td>
<td>0.9228</td>
<td>0.7828</td>
<td>0.8488</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
<td>0.7891</td>
<td>0.8597</td>
<td>0.9200</td>
<td>0.7810</td>
<td>0.8522</td>
<td>0.9164</td>
</tr>
<tr>
<td>3.0</td>
<td>0.7832</td>
<td>0.8491</td>
<td>0.9224</td>
<td>0.7803</td>
<td>0.8470</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>0.7856</td>
<td>0.8556</td>
<td>0.9186</td>
<td>0.7718</td>
<td>0.8458</td>
<td>0.9125</td>
</tr>
<tr>
<td>3.0</td>
<td>0.7821</td>
<td>0.8483</td>
<td>0.9220</td>
<td>0.7769</td>
<td>0.8445</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.7820</td>
<td>0.8529</td>
<td>0.9170</td>
<td>0.7604</td>
<td>0.8373</td>
<td>0.9077</td>
</tr>
<tr>
<td>3.0</td>
<td>0.7807</td>
<td>0.8472</td>
<td>0.9214</td>
<td>0.7726</td>
<td>0.8414</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.0</td>
<td>0.7692</td>
<td>0.8437</td>
<td>0.9117</td>
<td>0.7252</td>
<td>0.8116</td>
<td>0.8936</td>
</tr>
<tr>
<td>3.0</td>
<td>0.7757</td>
<td>0.8436</td>
<td>0.9195</td>
<td>0.7783</td>
<td>0.8309</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>0.7535</td>
<td>0.8319</td>
<td>0.9058</td>
<td>0.6882</td>
<td>0.7812</td>
<td>0.8809</td>
</tr>
<tr>
<td>3.0</td>
<td>0.7692</td>
<td>0.8388</td>
<td>0.9170</td>
<td>0.7400</td>
<td>0.8171</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
reduction in the mean of $\mu_{x_0}$ is identical to the proportional reduction in mean difference that we derived earlier (2.3.8.). The value of $1 - \frac{\sigma_0^2}{\sigma_{x_0}^2}$ is identical to the term derived by Cochran for his reduction in bias term. As can be seen the two are quite close, although Cochran's term declines faster as $\mu_{x_0}$ increases than does the actual proportional reduction in the mean of $\mu_{x_0}$. Note that there is a small, but steady, decrease in the values of the relative reduction columns comparing $r=3$ and $r=4$. This indicates that there is some small advantage to the investigator in increasing the number of categories in the matching scheme as $(\mu_{x_1} - \mu_{x_0})$ increases, although, as can be seen by examining the relative reduction columns, this advantage is likely to be small. Overall, categorization based on the case population performs quite well when the control population does not have the same normal distribution for its matching variable as does the case population.

2.4. Sample Size Considerations

The results presented in the last section suggest that as the distributions of the matching variable in the case and control groups become more dissimilar, it is advantageous to increase the number of matching categories. This increase in the number of matching categories has the effect of reducing both the post-sampling mean difference of the matching variable and the average within category variance. Increasing the number of matching categories also has the effect of making the cases more difficult to match. As the number of categories increases, the probability of a randomly selected control falling in a particular category would generally decrease, and thus those categories with the smallest probabilities would require examination of a large control group in order to insure that all the cases in that category are matched. As an example, consider
the normal distribution situation previously discussed. Recall that we assumed that \( \mu_{x_0} > \mu_{x_1} = 0 \). When the two normal populations have equal variances, the case observations in the first category, \( C_1 \), would generally be the most difficult ones to match, since the probability that a randomly selected control observation falls into this category will be the smallest of the probabilities that it falls into any one of the \( r \) categories, i.e., \( \pi_{01} = \min_{k} (\pi_{01}, \pi_{02}, \ldots, \pi_{0k}, \ldots, \pi_{0r}) \). Table 2.4.1 on the next page illustrates this point. Each entry is the probability that a randomly selected observation from a normally distributed control population with mean \( \mu_{x_0} \) and unit variance falls into \( C_1 \). Here, \( C_1 \) is the first interval in a categorization scheme that is equiprobable in a \( N(0,1) \) case population. Values are given for \( r = 3, 4, 5, 6 \) categories, and for \( \mu_{x_0} = 0.0, 0.1, 0.2, \ldots, 1.0 \). The value of \( A_2 \) given in the table is the upper boundary point of \( C_1 \); the lower boundary point is, of course, \(-\infty\).

As can be easily seen, the probability decreases quickly as \( \mu_{x_0} \) increases, the decrease becoming more accelerated as the number of categories increases. Although the particular numbers apply only to the normal case, the general principle remains the same: the more categories that are used, the more difficult it becomes to match all of the case observations.

With this in mind, the question facing the investigator becomes the following: how can both the increase in precision and the decrease in bias be made as large as possible while maintaining a reasonable cost? If there are no cost constraints and the control pool from which the samples are to be drawn is quite large, then there is really no practical restriction on the number of categories to be used. However, when there
TABLE 2.4.1

Probability that a Randomly Selected Observation from $X_0 \sim N(\mu_{x_0}, 1)$ will fall into Category $C_1$, the First of the Intervals that are equiprobable in a $N(0,1)$ Population, i.e., $P_r\{X_0 < A_2\}$

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{x_0}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.333</td>
<td>0.250</td>
<td>0.200</td>
<td>0.167</td>
</tr>
<tr>
<td>0.1</td>
<td>0.298</td>
<td>0.218</td>
<td>0.174</td>
<td>0.142</td>
</tr>
<tr>
<td>0.2</td>
<td>0.264</td>
<td>0.189</td>
<td>0.150</td>
<td>0.121</td>
</tr>
<tr>
<td>0.3</td>
<td>0.233</td>
<td>0.164</td>
<td>0.127</td>
<td>0.102</td>
</tr>
<tr>
<td>0.4</td>
<td>0.203</td>
<td>0.140</td>
<td>0.108</td>
<td>0.085</td>
</tr>
<tr>
<td>0.5</td>
<td>0.176</td>
<td>0.119</td>
<td>0.090</td>
<td>0.071</td>
</tr>
<tr>
<td>0.6</td>
<td>0.152</td>
<td>0.100</td>
<td>0.075</td>
<td>0.058</td>
</tr>
<tr>
<td>0.7</td>
<td>0.129</td>
<td>0.084</td>
<td>0.062</td>
<td>0.048</td>
</tr>
<tr>
<td>0.8</td>
<td>0.109</td>
<td>0.070</td>
<td>0.051</td>
<td>0.038</td>
</tr>
<tr>
<td>0.9</td>
<td>0.092</td>
<td>0.057</td>
<td>0.041</td>
<td>0.031</td>
</tr>
<tr>
<td>1.0</td>
<td>0.076</td>
<td>0.047</td>
<td>0.033</td>
<td>0.024</td>
</tr>
<tr>
<td>$(A_2)$</td>
<td>(-0.43)</td>
<td>(-0.67)</td>
<td>(-0.84)</td>
<td>(-0.97)</td>
</tr>
</tbody>
</table>

are a limited number of cases and there are restrictions on the availability of controls, then a point will be reached where it becomes difficult to match all the cases and where the use of a large number of categories becomes inefficient since not all of the cases can be matched. One method for assessing the relative merits of various category structuring schemes is to consider how many controls would have to be examined to obtain $n_k$ observations in each of the $k$ categories, $k=1,2,\ldots,r$.

This problem is similar to one faced in quota sampling. When a quota is being drawn up, it is useful to know the number of people or objects that will have to be examined until the quota is fulfilled. The
distribution of the least sample size (the minimum sample size to fulfill a quota) has been examined by several authors. Johnson (1952, 1957) studied the distribution of the least sample size and its cost when a large sample was taken and any quota that was not met by that sample could be fulfilled by further sampling of the deficient categories. Young (1961) found expressions for the distribution, moments and approximations to the upper percentile points of the least sample size when the restricted subsampling that Johnson studied could not be used. This is exactly the distribution we are considering for a given category matching scheme. We have r categories and each category contains \(n_k\) case observations that we wish to match with \(n_k\) controls. Our "quota" for the k-th category is then \(n_k\). Let \(\sum n_k = M\), the total number of matches we wish to obtain.

The probability that a single randomly selected observation from the distribution of the control group's matching variable will fall into the k-th category is \(p_k (\sum p_k = 1)\). The values of the \(p_k\)'s are determined by the distribution of the matching variable in the control group and the category structure used to categorize the matching variable in the case group. The least sample size random variable will be denoted by \(N\) and any realization of \(N\) will be denoted by \(n\). Young states that the probability that the least sample size \(N\) takes on the value \(n\) is

\[ Pr\{N=n\} = P(n) \]
\[ \sum_{k=1}^{r} \binom{n-1}{n_k-1} p_k^{n_k} (1-p_k)^{(n-n_k)} \sum_{i=1}^{c} \frac{(n-n_k)!}{s_j!} \left( \frac{p_j}{1-p_k} \right)^s_j, \]

\[ j=1 \sum_{j=n-n_k(i+k)}^{s_j} \]

\[ n = M, M+1, \ldots, \infty, \] \hspace{1cm} (2.4.1)

where \( s_j \geq n_j \), \( \sum_{j=1}^{r} s_j = n-n_k \) and \( \sum_{c}^{r} \) indicates the summation over all \( (j+k) \) non-negative integer values of the \( s_j \) fulfilling \( \sum_{j=1}^{r} s_j = (n-n_k) \).

This form can be easily derived by considering that in order for the \( n \)-th control observation sampled to be the last observation sampled, 1) the previous \( n-1 \) control observations must have occurred in such a manner that one and only one category must be one observation short of meeting its quota, while the rest of the categories have met or exceeded their quota; and 2) the \( n \)-th control observation must fall into the deficient category. In order for the first condition to hold, \( M-1 \) observations must have fallen into the correct categories to meet the minimum conditions and \( (n-1) - (M-1) = n-M \) observations must be distributed among all categories excluding the deficient ones. Thus, for \( \sum_{c}^{r} \) indicating \( \sum_{j=1}^{r} a_j = n-M \)
the summation over all $a_j$ such that $0 \leq a_j \leq (n-m)$; $j=1, \ldots, r, j+k$,

and

$$\sum_{j=1}^{r-k} a_j = (n-m),$$

$$\sum_{j=1}^{r-k} (n_j + a_j) = (n_k - 1)$$

\[
P(n) = \sum_{k=1}^{r} p_k \prod_{j=1}^{r} \frac{(n-j+a_j)!}{(n_k-1)! (j+k)}
\]

\[
= \sum_{k=1}^{r} \left( \sum_{j=1}^{r} \frac{n_k}{p_k} \frac{1}{(n_j + a_j)!} \right) \frac{(n-1)! (n-n_k)!}{\pi \prod_{j=1}^{r} (n_j + a_j)! (n_k-1)! (n-n_k)! (j+k)}
\]

\[
= \sum_{k=1}^{r} \left( \sum_{j=1}^{r} \frac{n_k}{p_k} \frac{1}{(n_j + a_j)!} \right) \frac{(n-1)! (n-n_k)!}{\pi \prod_{j=1}^{r} (n_j + a_j)! (n_k-1)! (n-n_k)! (j+k)}
\]

\[
= \sum_{k=1}^{r} \left( \sum_{j=1}^{r} \frac{n_k}{p_k} \frac{1}{(n_j + a_j)!} \right) \frac{(n-1)! (n-n_k)!}{\pi \prod_{j=1}^{r} (n_j + a_j)! (n_k-1)! (n-n_k)! (j+k)}
\]

Then if we substitute $s_j = n_j + a_j$, so that

$$\sum_{j=1}^{r-k} s_j = n - M \equiv (\sum_{j=1}^{r-k} s_j = n-n_k),$$

we will have Young's expression (2.4.1). The $\lambda$-th ascending factorial moment is then defined in the usual manner (for $\lambda \geq 0$):
\[ E \left( \frac{(N+\xi-1)!}{(N-1)!} \right) = \sum_{n=M}^{\infty} \frac{(n+\xi-1)!}{(n-1)!} P(n) \quad . \quad (2.4.2.) \]

Young points out that it is not convenient to use (2.4.1.) to evaluate (2.4.2.) since \( n \) is involved in the limits of the interior summation of (2.4.1.); and, so, he offers an alternate general form. This alternate form is based on transforming the interior sum so that it is not dependent on \( n \) and interchanging summation signs in (2.4.2.) to bring the infinite sum over \( n \) and all terms involving \( n \) into the interior. However, he only evaluates this for the general two-category case and a special case of the three-category situation. In the two-category case, we have a quota of \( n_1 \) in the first category and \( n_2 \) in the second category. The probability of a randomly selected observation falling into category one is \( p_1 \) and of the observation falling into the second category is \( p_2 (p_1 + p_2 = 1) \). Then

\[ E(\frac{N+\xi-1}{N-1}) = \frac{(n_1+\xi-1)!}{(n_1-1)!} \frac{\theta}{p_1} I_{p_2}^{n_2, n_1+\xi} + \frac{(n_2+\xi-1)!}{(n_2-1)!} \frac{\theta}{p_2} I_{p_1}^{n_1, n_2+\xi}, \]

(2.4.3.)

where \( I_p(a, n-a+1) = \sum_{i=a}^{n} \frac{n!}{i!(n-i)!} (1-p)^{n-i} \) is the usual notation for the incomplete beta function ratio. (See Johnson and Kotz, (1969), Chap. 1, for further reference on this ratio.) From this we can derive the mean and variance of \( n \) for two categories; in particular, for \( \xi = 1 \)

\[ E(N) = n_1 p_1^{1-1} I_{p_2}^{n_2, n_1+1} + n_2 p_2^{1-1} I_{p_1}^{n_1, n_2+1} \]
and, for $\lambda = 2$

$$E(N(N+1)) = n_1(n_1+1) p_1^2 I_{p_2}(n_2,n_1+2) + n_2(n_2+1) p_2^2 I_{p_1}(n_1,n_2+2).$$

Then, since $E(N(N+1)) = E(n^2) + E(n) = \text{Var}(n) + E(n) + (E(n))^2$, we have

$$\text{Var}(N) = E(n(n+1)) - E(n) - (E(n))^2.$$

Young also found an explicit form of (2.4.2.) when there are three categories, each having $n_k = m$ and $p_k = 1/3$. This corresponds to the situation where there is an equal number of cases in each category and the probability that a randomly selected control observation falls into any one of the three categories is one-third, i.e., the categories are equiprobable in the controls. If we assume that the matching variable is normally distributed in both groups, this can occur only if the two normal distributions are identical. The expression found by Young is, for $n_1 = n_2 = n_3 = m$,

$$E \left\{ \frac{(N+\lambda-1)!}{(N-1)!} \right\} = 3^{\lambda+1} \left( \frac{(m+\lambda-1)!}{(m-1)!} I_{2/3} \right) (2m, m+\lambda)$$

$$- \sum_{s=0}^{m-1} \frac{(m+s+\lambda-1)!}{(m-1)!s!_2^{m+s+\lambda-1}} I_{1/3} (2m-s, m+s+r)$$

Young used this expression to provide a small table of the expected value of $n$ for several values of $m$. The values he found are reproduced
in Table 2.4.2 below. Note that as the value of \( m \) (and thus \( 3m \), the total number of matches desired) increases, the ratio of \( E(N) \) to \( 3m \)

<table>
<thead>
<tr>
<th>Number of Cases per Category</th>
<th>Total Number of Cases</th>
<th>( E(N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m )</td>
<td>( 3m )</td>
<td>5.5</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9.6</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>13.5</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>17.2</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>20.8</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>38.2</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>71.6</td>
</tr>
<tr>
<td>20</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

*Adapted from Young [1961] Table 3, p. 327.

decreases toward one. He also derived an expression for \( E(N) \) which contains a small number of terms and is thus useful for all \( m \):

\[
E(N) = 3^m + 9m \left( \frac{2^m}{m} \right) _{1/3} (m+2, 2m+1) + (9m+5)(3m)! \cdot 3^{-3m}/[(m-1)!m!(m+1)!] 
\]

Strictly speaking, this expression can be used to find \( E(N) \) whenever the categories are equiprobable in the control distribution and we have the same number of cases in each category, but, in the context of a normally distributed matching variable, this is useful only when the matching variable is identically distributed in both populations in which case matching serves only to increase precision.
In the appendix, we derive an explicit form of (2.4.2.) for the general four category-matching scheme and for the general three category scheme as a special case of the four category-matching scheme. These are expressions (A.1.6) and (A.1.9), respectively.

Expressions (A.1.6) and (A.1.9) can be used to compute the mean and the variance of $N$ for any three or four category-matching scheme on any number of cases. This will give us an estimate of the size of the control group that is required to completely category-match the cases by that category-matching scheme. In the last section, we examined equiprobable three and four category-matching schemes based on the case population's $N(0,1)$ distribution when the control matching variable was distributed as a $N(\mu_{x_0},\sigma_{x_0}^2)$ variable. The entries in Table 2.4.3 are the mean and variance of the number of $N(\mu_{x_0},1)$ control observations required to match $M=12$ $N(0,1)$ cases that are to be category-matched using a $N(0,1)$ equiprobable category-matching scheme. This table was constructed using (A.1.6) and (A.1.9), assuming that the $M=12$ case observations are divided equally among the $r=3,4$ categories, i.e., $n_1=n_2=n_3=4$ when $r=3$, and $n_1=n_2=n_3=n_4=3$ when $r=4$. The values of the $p_k$ in (A.1.6) and (A.1.9) are simply $p_k=\pi_{0k}$. As is expected, the mean number of cases needed to match the controls increases with an increase in $r$, the number of categories, and with the value of $\mu_{x_0}$. Under the condition of a limited control pool size, then, it is more to the advantage of an investigator to use a small number of categories to insure matching all of the case observations.

Explicit forms for $E(N)$ and $V(N)$ can also be derived for $r=5,6,...$ by using the same approach that was used to derive (A.1.6). But, as can be seen by examining that derivation, and by comparing (A.1.6) and
TABLE 2.4.3

Mean and Standard Deviation of the Least Sample Size of $N(x_0^2,1)$ Controls Necessary to Category-Match $M=12$
$N(0,1)$ Cases, Using $r=3, 4$ $N(0,1)$ Equiprobable Categories ($n_i=M/r$)

<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r=3$</td>
<td>$r=4$</td>
</tr>
<tr>
<td></td>
<td>$r=3$</td>
<td>$r=4$</td>
</tr>
<tr>
<td>0.0</td>
<td>17.47</td>
<td>19.90</td>
</tr>
<tr>
<td>0.2</td>
<td>19.72</td>
<td>22.61</td>
</tr>
<tr>
<td>0.4</td>
<td>24.40</td>
<td>28.36</td>
</tr>
<tr>
<td>0.6</td>
<td>32.16</td>
<td>38.16</td>
</tr>
<tr>
<td>0.8</td>
<td>44.49</td>
<td>54.23</td>
</tr>
<tr>
<td>1.0</td>
<td>64.17</td>
<td>80.72</td>
</tr>
</tbody>
</table>

(A.1.9), the expressions for the moments very quickly become quite messy. We can, however, approach the question of the least sample size by examining the probability that the least sample size will exceed a given number of control observations. Although the expected value of the least sample size is the average control pool size necessary to category-match all the cases, there is still an appreciable probability that there will not be a complete match at that sample size. We can then ask: "What is the smallest control sample size that will give a complete match of all case observations with some probability $(1-\alpha)$?"

Young developed a simple approximation for the upper tail probability points of (2.4.2.). He found that

\[ \Pr(N > n^*) \approx \sum_{k=1}^{r} (1-p_k)(n^*-n_k+1,n_k), \]
so that the least integer value \( n_\alpha \), say, satisfying

\[
\sum_{k=1}^{r} I(1-p_k) (n_k - n_k, n_k) \leq \alpha
\]

(2.4.12.)

gives \( \Pr\{N > n_\alpha\} \leq \alpha \). Since this is an approximation, there is a difference between the approximate confidence level of (2.4.12.) and the true confidence level for \( \Pr\{N > n_\alpha\} \). The upper bound on this difference is \( 1/2 \alpha^2 \). Thus, (2.4.12.) gives a reasonably close answer for \( \alpha < 0.1 \) (i.e., \( 1/2 \alpha^2 < 0.005 \)), but cannot be used for larger \( \alpha \). In Table 2.4.3, we have used (2.4.12.) to find the 90% bounds on the number of observations from the control groups \( N(\mu_{x_0}, 1) \) matching variable distribution necessary to category-match \( M = 60 \ N(0, 1) \) case observations using \( r = 3, 4, 5, 6 \) \( N(0, 1) \) equiprobable categories. \( M \) was set at 60 since 60 is the smallest integer that can be divided up evenly for the values of \( r \) i.e., 60 is the smallest common multiple of 3, 4, 5, 6. As is expected, the 90% bounds increase as \( r \), the number of categories, increases, and as \( \mu_{x_0} \) increases. Note that the increase in the 90% bound as \( \mu_{x_0} \) increases, accelerates as \( r \) increases. This indicates that a large number of categories is more difficult to category-match than a smaller number of categories and that this difficulty increases as the mean difference in the distributions of the matching variable increases.

Thus, we conclude that when there are a limited number of controls from which to draw the matching values, an investigator will have the best chance of category-matching all the cases when the number of categories is small (\( r = 3, 4 \)).
TABLE 2.4.

90% Bound on the Number of Random Samples from a \( N(\mu_{x_0}, 1) \)

Control Distribution Necessary to Category Match \( M=60 \) \( N(0,1) \) Case Observations in a \( N(0,1) \) Equiprobable Category Matching Scheme (\( n_1=M/r \))

<table>
<thead>
<tr>
<th>( \mu_{x_0} )</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>83</td>
<td>90</td>
<td>97</td>
<td>104</td>
</tr>
<tr>
<td>0.2</td>
<td>96</td>
<td>105</td>
<td>114</td>
<td>121</td>
</tr>
<tr>
<td>0.4</td>
<td>125</td>
<td>140</td>
<td>154</td>
<td>166</td>
</tr>
<tr>
<td>0.6</td>
<td>169</td>
<td>196</td>
<td>221</td>
<td>242</td>
</tr>
<tr>
<td>0.8</td>
<td>235</td>
<td>283</td>
<td>329</td>
<td>367</td>
</tr>
<tr>
<td>1.0</td>
<td>338</td>
<td>423</td>
<td>506</td>
<td>577</td>
</tr>
</tbody>
</table>

2.5. **Category-Matching and Covariance Analysis**

The material in the previous sections has been presented without explicit reference to the functional relationship between the response variable and the matching variable. A relationship was implicitly assumed to exist to provide motivation for matching. In this section, we will utilize a linear model to describe the relationship between a continuous response variable and a continuous matching variable. In this context, a natural alternative to the use of categorical-matching and a simple mean difference estimate of the case-control difference is an analysis of covariance approach. This approach attempts to model the relationship between the response and matching variate, thus giving an unbiased estimate of the case-control difference (if the assumed model is correct) and providing additional information concerning the nature
of the relationship between the two variates. We note that this approach differs from category-matching in that the latter methods attempts to provide an estimate of the case-control effect with lower bias and increased precision relative to the simple random sample mean difference estimator, but does not attempt to quantify the functional relationship between the two variates. The analysis of covariance model has been considered by numerous authors. Two of them, Billewicz (1965) and Cochran (1968), have compared the technique of category-matching with the technique of analysis of covariance. This work has been done under the assumption that the matching variable is identically distributed in both populations. We shall review their work in this area and extend it to the more realistic case where the two distributions are not identical.

The model which these two authors have used and which we will study is the analysis of covariance model:

$$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}, \quad i=0,1; \quad j=1,\ldots,n_i \quad (2.5.1.)$$

where the $e_{ij}$'s are assumed, as usual, to be independent $N(0,\sigma^2_e)$. Billewicz (1965) performed an extensive Monte Carlo simulation to compare the efficiencies of categorical-matching and covariance analysis. The matching variable was assumed to be identically distributed and the response variate was assumed to be distributed according to a $N(\alpha_i + \beta x_{ij}, 1)$ distribution, conditional on $x$ and the group to which it pertained. He used a matching scheme with three categories and computed a mean difference, $\overline{Y}_i - \overline{Y}_0 = d$, say, to estimate $(\alpha_i - \alpha_0)$, the case-control difference. From the covariance analysis, he estimated $(\alpha_i - \alpha_0)$ as

$$(\alpha_i - \alpha_0) = \overline{Y}_i - \overline{Y}_0 - \hat{\beta} (\overline{X}_i - \overline{X}_0).$$

He computed the sample variance of each
estimate and then computed the sample relative efficiency. He found
that the covariance analysis estimate was, on the average, more effi-
cient than the categorically matched estimate for the range \( \rho_{xy} =
0.2, 0.4, 0.6, 0.8. \)

Cochran (1968) took an analytical approach to the same problem. He
assumed that \( X_i \sim N(\mu_x, \sigma_x^2) \) and \( Y_i \sim N(\mu_y, \sigma_y^2) \). Now, it is well known that
the unconditional variance of \( \overline{Y}_1 - \overline{Y}_0 \) estimated from two independent
random samples of size \( n \) is \( 2\sigma_y^2/n \). Of this variance, \( 2(1-\rho^2)\sigma_y^2/n =
2\sigma_y^2/n \) is due to variations that were unrelated to \( x \), and \( 2\rho^2\sigma_y^2/n =
2\rho^2\sigma_y^2/n \) is due to variations in the \( x \)'s. When the \( x \)'s are category-
matched \( \sigma^2_{1k} = \sigma^2_{k} \), since the two variables are identically distributed;
and, the unconditional variance of \( \beta(\overline{X}_1 - \overline{X}_0) = \beta \sum \frac{n_k}{n} \overline{X}_{1k} - \overline{X}_{0k} \) is
\[ 2\beta^2 \sum \frac{n_k}{n} \sigma^2_k, \]
where \( n_k \) is the number of observations in the \( k \)-th category
\( \sum n_k = n \). Thus, since \( n_k/n \neq \pi_k \), the average value of \( 2\beta^2 \sum \frac{n_k}{n} \sigma^2_k/n \) is
\[ 2\beta^2 \sum \pi_k \sigma^2_k/n. \] Thus, Cochran found that when the category-matched form of
\( \overline{Y}_1 - \overline{Y}_0 \), namely, \( \sum \frac{n_k}{n} (\overline{Y}_{1k} - \overline{Y}_{0k}) \) is considered, the variance in \( \overline{Y}_1 - \overline{Y}_0 \)
due to \( x \) is reduced from its value of \( 2\beta^2\sigma_x^2/n \) based on random samples to
\[ 2\beta^2 \sum \pi_k \sigma^2_k/n \] for category-matched samples. Since the matching is assumed
to affect only the distribution of the \( x \)'s and not that of the \( e \)'s, the
variance of \( \overline{Y}_1 - \overline{Y}_0 \) under category-matching is

\[ 2\sigma_y^2(2\rho^2 + (1-\rho^2))/n = 2\sigma_y^2(1-(1-g)\rho^2)/n, \quad (2.5.2) \]

where \( g = \sum \pi_k \sigma^2_k/\sigma^2_x \). Now, since the matching variable is assumed to be
identically distributed in the two groups, it follows from a previously
derived result that
\[
V(X|\bar{r},A) \approx \sum_{k=1}^{r} p_k \sigma_k^2 + \sum_{k=1}^{r} p_k \mu_k^2 - (E(X|\bar{r},A))^2.
\]

But, in Cochran's development it is assumed that \( p_k = \pi_k \); and since we showed earlier that \( f_1(x) = f_0(x) = f(x) \) and \( p_k = \pi_k \) implies \( V(X|\bar{r},A) = \sigma_x^2 \) and \( E(X|\bar{r},A) = \mu_x \), it follows that

\[
\sigma_x^2 = \sum_{k=1}^{r} \pi_k \sigma_k^2 + \sum_{k=1}^{r} \pi_k \mu_k^2 - \mu_x^2,
\]

so that

\[
(1-g) = \sum_{k=1}^{r} \pi_k \frac{\mu_k^2}{\sigma_k^2} - \frac{\mu_x^2}{\sigma_x^2}.
\]

Cochran defines this quantity as "the proportional reduction in the variance of \((\bar{X}_1 - \bar{X}_0)\)". Cochran makes a further assumption that \( X_i \sim N(0,1) \), so that from (2.3.3.)

\[
(1-g) = \sum_{k=1}^{r} \pi_k \frac{\mu_k^2}{\sigma_k^2}
\]

\[
= \sum_{k=1}^{r} \frac{\pi_k (Z(A_k) - Z(A_k+1))^2}{\pi_k}.
\]
= \sum_{k=1}^{r} \frac{(Z(A_k)-Z(A_k+1))^2}{\pi_k}.

This is the same term he derived as the "proportional reduction in bias", which we discussed in Section 2.3. Now, since the unconditional variance of \(\tilde{Y}_1 - \tilde{Y}_0\) is equal to \(2\sigma^2_y/n\), the percent reduction in the variance of \(\tilde{Y}_1 - \tilde{Y}_0\) due to category-matching on the x's is

\[
100 \cdot \left( \frac{2\sigma^2_y/n - 2\sigma^2_y(1-(1-y)\rho^2)/n}{2\sigma^2_y/n} \right) = 100(1-g)\rho^2.
\]

He investigated the behavior of this quantity for \(\rho = 0.2, 0.1, 0.9\) and \(r=2,3,4,5,6\) equiprobable \(N(0,1)\) intervals. He found for \(\rho<0.7\), that categorical-matching with at least four equiprobable categories attains most of the potential reduction in bias; and further, for a small number of categories (\(\leq 6\)), he concluded that category-matching is always less efficient, in terms of the variance of the estimators of \((\alpha_1 - \alpha_0)\), than covariance analysis (given that the assumed model is correct).

We will use the same covariance model to study the relative efficiency and mean square error of the estimators based on category-matching and covariance analysis when \(X_1 \sim N(0,1)\) and \(X_0 \sim N(\mu_{x_0},1)\). We will also present some approximate results for \(X_0 \sim N(\mu_{x_0},\sigma^2_{x_0})\). For a category-matching scheme with \(r\) categories, a sample of size \(n_k\) in each category (\(\sum n_k = n\)) and the vector of boundary points \(A = (A_1,A_2,\ldots,A_{r+1})\), let \(x_{ikm}\) be the value of the \(m\)-th individual's matching variable in the \(k\)-th category of the \(i\)-th population,
\[ i=0,1; \ k=1,\ldots,r; \ m=1,\ldots,n_k \]

Likewise, define \( y_{ikm} \) for the response associated with \( x_{ikm} \) and \( e_{ikm} \) for the error. It is assumed that \( e_{ikm} \sim N(0, \sigma^2_e) \). So the model which we will consider is

\[
y_{ikm} = \alpha_i + \beta x_{ikm} + e_{ikm}. \tag{2.5.5}
\]

A reasonable estimator of \((\alpha_1 - \alpha_0)\), the case-control difference, based on the category-matched sample is then

\[
\bar{d} = \overline{y}_1 - \overline{y}_0 = 1/n \left\{ \sum_{k} \sum_{m} y_{1km} - \sum_{k} \sum_{m} y_{0km} \right\} = 1/n \sum_{k=1}^{r} n_k (\overline{y}_{1k} - \overline{y}_{0k})
\]

\[
= 1/n \sum_{k=1}^{r} n_k \bar{d}_k. \tag{2.5.6.}\]

Now, from (2.5.5),

\[
\bar{d}_k = (\alpha_1 - \alpha_0) + \beta (\overline{x}_{1k} - \overline{x}_{0k}) + (\overline{e}_{1k} - \overline{e}_{0k}) \tag{2.5.7.}\]

Thus,

\[
E(\bar{d}_k | \overline{x}_k, \overline{x}_{0k}) = (\alpha_1 - \alpha_0) + \beta (\overline{x}_{1k} - \overline{x}_{0k})
\]
\[
\text{and}
\]

\[
\text{var}(\overline{d}_k | \overline{x}_{1k}, \overline{x}_{0k}) = \frac{2\sigma^2_e}{n_k}.
\]

The unconditional moments are then

\[
E(\overline{d}_k) = E_{x_0} E_{x_1} E(\overline{d}_k | \overline{x}_{1k}, \overline{x}_{0k})
\]

\[
= (\alpha_1 - \alpha_0) + \beta (\mu_{1k} - \mu_{0k})
\]

and

\[
\text{var}(\overline{d}_k) = E_{x_0} E_{x_1} \text{var}(\overline{d}_k | \overline{x}_{1k}, \overline{x}_{0k}) + E_{x_0} E_{x_1} E(\overline{d}_k | \overline{x}_{1k}, \overline{x}_{0k})
\]

\[
+ E_{x_1} E_{x_0} \text{var}(\overline{d}_k | \overline{x}_{1k}, \overline{x}_{0k})
\]

\[
= \frac{2\sigma^2_e}{n_k} + \beta^2 \frac{\sigma^2_{1k}}{n_k} + \beta^2 \frac{\sigma^2_{0k}}{n_k}
\]

\[
= \frac{2\sigma^2_e}{n_k} \left[ 1 + \frac{\beta^2}{2\sigma^2_e} (\sigma^2_{0k} + \sigma^2_{1k}) \right].
\]

Since \( \overline{d} = \sum \frac{n_k}{n} \overline{d}_k = \sum p_k \overline{d}_k \), it follows that
\[ E(\overline{d}) = (\alpha_1 - \alpha_0) + \beta \sum_{k=1}^{r} p_k (\mu_{1k} - \mu_{0k}) \quad , \quad (2.5.8. \text{)} \]

and

\[ \text{Var}(\overline{d}) = \sum_{k=1}^{r} \left( \frac{n_k}{n} \right)^2 \frac{2 \sigma_e^2}{n_k} \left( 1 + \frac{\beta^2}{2 \sigma_e^2} \left( \sigma_{0k}^2 + \sigma_{1k}^2 \right) \right) \]

\[ = \frac{2 \sigma_e^2}{n} \left( 1 + \frac{\beta^2}{2 \sigma_e^2} \sum_{k=1}^{r} p_k (\sigma_{0k}^2 + \sigma_{1k}^2) \right) \quad . \quad (2.5.9. \text{)} \]

Also, from (2.5.8.),

\[ \text{BIAS} (\overline{d}) = E(\overline{d}) - (\alpha_1 - \alpha_0) \]

\[ = \beta \sum_{k=1}^{r} p_k (\mu_{1k} - \mu_{0k}) \quad . \quad (2.5.10. \text{)} \]

Now, under the assumption that \( X \sim N(\mu_x, \sigma_x^2) \), we showed earlier that

\[ \mu_{1k} = \mu_{x_i} + \mathbf{J}_{i,k} \sigma_{x_i} \]

and

\[ \sigma_{ik}^2 = \sigma_{x_i}^2 \left\{ 1 + H_{i,k} - \mathbf{J}_{i,k}^2 \right\} \quad . \]

If we apply these results to (2.5.9.) and (2.5.10.), we have
\[ \text{BIAS } (\bar{d}) = \beta (\mu_{x_1} - \mu_{x_0} + \sum_{k=1}^{r} p_k [J_{1k} \sigma_{x_1} - J_{0k} \sigma_{x_0}]) , \quad (2.5.11.) \]

and

\[ \text{Var}(\bar{d}) = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \sum_{k=1}^{r} p_k [(1+H_{1k},-J_{1k})\sigma_{x_1}^2 + (1+H_{0k},-J_{0k})\sigma_{x_0}^2] . \]

(2.5.12.)

These forms can be used to evaluate the effects of any category-matching procedure under our model. Two special cases of this general type of matching are unmatched or randomly matched samples and exact pair-matched samples. In the first special case, there is only one interval, so that (2.5.11.) and (2.5.12.) become

\[ \text{BIAS } (\overline{d}) = \beta (\mu_{x_1} - \mu_{x_0} + J_{11} \sigma_{x_1} - J_{01} \sigma_{x_0}) \]

and

\[ \text{Var}(\overline{d}) = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \{(1+H_{11},-J_{11})\sigma_{x_1}^2 + (1+H_{01},-J_{01})\sigma_{x_0}^2 \} . \]

Now, when \( \xi_{1,1} = -\infty, \xi_{1,2} = \infty, \)

\[ J_{1,1} = \frac{Z(-\infty) - Z(\infty)}{\phi(\infty) - \phi(-\infty)} = \frac{0}{1} = 0 . \]

Then, \( \text{BIAS } (\overline{d}) = \beta (\mu_{x_1} - \mu_{x_0}) \), as we would expect.
Now, $H_{i,1}$ cannot be directly evaluated in the same manner as can $J_{i,1}$ since the numerator of $H_{i,1}$

$$
\xi_{i,1}^2 Z(\xi_{i,1}) - \xi_{i,2}^2 Z(\xi_{i,2})
$$

is an indeterminant form when $\xi_{i,1} = -\infty$ or $\xi_{i,2} = \infty$. However, if we take the limit as $\xi_{i,1} \to -\infty$ and $\xi_{i,2} \to \infty$, we can apply L'Hospital's rule to find the limit. Now, since $D_{\xi_{i,j}} Z(\xi_{i,j}) = -\xi_{i,j} Z(\xi_{i,j})$ exists for all values of $\xi_{i,j}$,

$$
\lim_{\xi_{i,1} \to -\infty} \frac{\xi_{i,1}^2}{\sqrt{2\pi}} e^{1/2 \xi_{i,1}^2} = \lim_{\xi_{i,1} \to -\infty} \frac{D\xi_{i,1} Z(\xi_{i,1})}{\sqrt{2\pi}} e^{1/2 \xi_{i,1}^2}
$$

$$
= \lim_{\xi_{i,1} \to -\infty} \left( \frac{\xi_{i,1}^2}{\sqrt{2\pi}} e^{1/2 \xi_{i,1}^2} \right)^{-1} = 0
$$
Likewise,

$$\lim_{\xi_{1,2} \to \infty} \xi_{1,2} Z(\xi_{1,2}) = 0,$$

so

$$\lim_{\xi_{1,1} \to -\infty} \xi_{1,1} = 0 \quad \text{and} \quad \lim_{\xi_{1,2} \to \infty} \xi_{1,2} = 0.$$ 

Thus,

$$\lim_{\xi_{1,1} \to -\infty} \lim_{\xi_{1,2} \to \infty} \text{Var}(\overline{d}) = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \left( \frac{\sigma^2}{x_1} + \frac{\sigma^2}{x_0} \right).$$

which agrees with the results from the usual derivation of the bias and variance of the mean difference in $Y_1$ and $Y_0$ from randomly matched samples based on Model (2.5.3.).

When the cases and controls are exactly pair-matched, so that $x_{0km} = x_{1km}$, the situation can be conceptually viewed as one involving category pair-matched data with $r=2n+1$ categories. Here, $n$ of these intervals would have zero "width" with $p_k = 1/n$, so that there is one pair per interval and $n+1$ of these intervals are empty ($p_k = 0$) with an unspecified width. There is one of these empty intervals between every one of the ordered pairs, giving $n-1$ intervals and an interval covering the range of the variable for each of the upper and lower ends. Thus, we have $2n+1$ mutually exclusive categories that cover the entire range of $x$. Since $p_k = 0$ for the empty intervals, (2.5.11.) and (2.5.12.) become, using the notation that $p_k = 0$ for $k$ odd and $p_k = 1/n$ for $k$ even,
\[ \text{BIAS } (\overline{d}) = \beta (\mu - \mu_0) + \sum_{k=1}^{n} p_{2k}[J_{1,2k} \sigma_{x_1} - J_{0,2k} \sigma_{x_0}] \]

\[ \text{Var}(\overline{d}) = \frac{2}{n} \sigma^2 + \frac{\beta^2}{n} \sum_{k=1}^{n} p_{2k} \left[ (1 + H_{1,2k} - J_{2,1,2k}) \sigma^2_{x_1} + (1 + H_{0,2k} - J_{2,0,2k}) \sigma^2_{x_0} \right] \]

For exact pair-matching, \( \xi_{1,2k} = \xi_{1,2k+1} \); under this condition, however, the expressions given above are indeterminate, since both the numerators and denominators of all \( H_{1,2k} \) and all \( J_{1,2k} \) are zero. So, we will find the limit of these terms as \( \xi_{1,2k} \rightarrow \xi_{1,2k+1} \). Any \( J_{1,2k} \) is the form

\[ J_{1,2k} = \frac{Z(\xi_{1,2k}) - c_1}{c_2 - \phi(\xi_{1,2k})} \]

where \( c_1 = Z(\xi_{1,2k+1}) \), \( c_2 = \phi(\xi_{1,2k+1}) \).

Thus,

\[ \lim_{\xi_{1,2k} \rightarrow \xi_{1,2k+1}} \frac{Z(\xi_{1,2k}) - c_1}{c_2 - \phi(\xi_{1,2k})} = \lim_{\xi_{1,2k} \rightarrow \xi_{1,2k+1}} \frac{D_{\xi_{1,2k}}(Z(\xi_{1,2k}) - c_1)}{D_{\xi_{1,2k}}(c_2 - \phi(\xi_{1,2k}))} = \lim_{\xi_{1,2k} \rightarrow \xi_{1,2k+1}} \frac{Z(\xi_{1,2k})}{\xi_{1,2k}} \]

\[ = \lim_{\xi_{1,2k} \rightarrow \xi_{1,2k+1}} \frac{Z(\xi_{1,2k})}{\xi_{1,2k}} \frac{z(\xi_{1,2k})}{z(\xi_{1,2k})} \]
\[
= \xi_{1,2k+1} = \frac{A_{2k+1} - \mu x_1}{\sigma x_1}
\]

If we apply this result to the bias term, we get

\[
\lim \quad \text{BIAS} (\bar{d}) = \beta (\mu x_1 - \mu x_0) \\
\xi_{1,2k} + \xi_{1,2k+1} \\
\forall k, k=1, \ldots, n
\]

\[
+ \sum_{k=1}^{r} p_{2k} \left[ \left( \frac{A_{2k+1} - \mu x_1}{\sigma x_1} \right) \sigma x_1 - \left( \frac{A_{2k+1} - \mu x_0}{\sigma x_0} \right) \sigma x_0 \right]
\]

\[
= \beta (\mu x_1 - \mu x_0) + \sum_{k=1}^{r} p_{2k} \left[ A_{2k+1} - \mu x_1 - A_{2k+1} - \mu x_0 \right]
\]

\[
= \beta (\mu x_1 - \mu x_0) + \sum_{k=1}^{r} \frac{1}{r} (\mu x_0 - \mu x_1)
\]

\[
= 0
\]

Likewise, for \( c_3 = \xi_{1,2k+1} Z(\xi_{1,2k+1}) \)

\[
\lim \quad H_{1,2k} = \lim \quad \frac{\xi_{1,2k} Z(\xi_{1,2k}) - c_3}{c_2 - \phi(\xi_{1,2k})}
\]

\[
\xi_{1,2k} + \xi_{1,2k+1}
\]

\[
\xi_{1,2k} + \xi_{1,2k+1}
\]
\[
\begin{align*}
&= \lim_{\xi_{i,2k} \to \xi_{i,2k+1}} \frac{D_{\xi_{i,2k}} (\xi_{i,2k}^2 (\xi_{i,2k}^2) - c_2)}{D_{\xi_{i,2k}} (c_2 - \phi(\xi_{i,2k}))} \\
&= \lim_{\xi_{i,2k} \to \xi_{i,2k+1}} \frac{\xi_{i,2k} D_{\xi_{i,2k}} Z(\xi_{i,2k}) + Z(\xi_{i,2k}) D_{\xi_{i,2k}} \xi_{i,2k} - \phi(\xi_{i,2k})}{D_{\xi_{i,2k}}} \\
&= \lim_{\xi_{i,2k} \to \xi_{i,2k+1}} -\frac{(1-\xi_{i,2k}^2)Z(\xi_{i,2k})}{Z(\xi_{i,2k})} \\
&= -(1 - \xi_{i,2k+1}^2).
\end{align*}
\]

If we apply this to the variance term, we get

\[
\begin{align*}
\lim_{\xi_{i,2k} \to \xi_{i,2k+1}} \frac{2\sigma^2}{n} & \sum_{k=1}^{n} p_{2k}(1-(1-\xi_{i,2k+1}^2) - \xi_{i,2k+1}^2)\sigma^2_{x_i} \\
& + \frac{\bar{g}^2}{n} \left( \sum_{k=1}^{n} p_{2k}(1-(1-\xi_{i,2k+1}^2) - \xi_{i,2k+1}^2)\sigma^2_{x_i} \right) \\
& + \left[1-(1-\xi_{0,2k+1}^2) - \xi_{0,2k+1}^2 \right]\sigma^2_{x_0}) \\
& = \frac{2\sigma^2}{n}.
\end{align*}
\]
This is the expected result which is ordinarily derived via more standard procedures.

Billewicz and Cochran considered the case where the matching variable is distributed as N(0,1) in both groups. In this case, since 
\[ \mu_{x_1} = 0, \quad \sigma_{x_1}^2 = 1, \]
we have
\[ H_{i,k} = \frac{A_k Z(A_k) - A_{k+1} Z(A_{k+1})}{\phi(A_{k+1}) - \phi(A_k)} \equiv H_k \]
and
\[ J_{i,k} = \frac{Z(A_k) - Z(A_{k+1})}{\phi(A_{k+1}) - \phi(A_k)} \equiv J_k. \]

Thus, the estimator \( \overline{d} \) is unbiased, since
\[ \text{BIAS} (\overline{d}) = \beta \{ \mu_{x_1} - \mu_{x_0} + \sum p_k (J_k - J_k) \} = 0. \quad (2.5.13.) \]
The variance of \( \overline{d} \) becomes
\[ \text{Var}(\overline{d}) = \frac{2\sigma^2}{n} + \frac{2\beta^2}{n} \sum_{k=1}^{r} p_k [1 + H_k - J_k^2]. \quad (2.5.14.) \]

Using Billewicz's method of pair-matching, the entire range of \( x \) is divided into a small number of categories, so that all \( p_k > 0 \). For a random sample of the cases, \( p_k = n_k / n \) is an estimate of the \( \pi_k = \phi(A_{k+1}) - \phi(A_k) \). For the true average value of \( p_k \), then,
\[ \sum_{k=1}^{r} p_k H_k = \sum_{k=1}^{r} \pi_k \left( \frac{A_k Z(A_k) - A_{k+1} Z(A_{k+1})}{\pi_k} \right) \]

\[ = A_1 Z(A_1) - A_{r+1} Z(A_{r+1}) \]

From L'Hôpital's rule, \( \lim A_1 Z(A_1) = \lim A_{r+1} Z(A_{r+1}) = 0 \), so that
\[ A_1 \to -\infty \quad A_{r+1} \to +\infty \]

\[ \sum_{k=1}^{r} p_k H_k = 0 \], under the \( N(0,1) \) assumption. The unconditional variance of \( \bar{d} \) for the \( N(0,1) \) distribution is then approximately

\[ \text{Var}(\bar{d}) = \frac{2\sigma_e^2}{n} + \frac{2\beta^2}{n} \left\{ 1 - \sum_{k=1}^{r} \pi_k J_k \right\} \]

\[ = \frac{2\sigma_e^2}{n} \left\{ 1 + \frac{\rho_{xy}^2}{(1-\rho_{xy}^2)} \right\} \left\{ 1 - \sum_{k=1}^{r} \pi_k J^2_k \right\} \]  
\[ (2.5.15.) \]

since \( \sigma_e^2 = \sigma_y^2 (1-\rho_{xy}^2) \) and \( \beta^2 = \rho_{xy}^2 \sigma_x^2 \sigma_y^2 / \sigma_x^2 = \rho_{xy}^2 \sigma_y^2 \).

Under the usual covariance model \( y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}, i=0,1 \) and \( j=1,2,...,n \), the estimator of \( (\alpha_1 - \alpha_0) \), the case-control difference, is

\[ \hat{\psi} = \bar{y}_1 - \bar{y}_0 - \hat{\beta}(\bar{x}_1 - \bar{x}_0) \]
This estimator has an advantage over \( \bar{d} \), the category-matched estimator, since it is always unbiased if the given model holds, while \( \bar{d} \) is only unbiased when \( \mu_{x_1} = \mu_{x_0} \). In particular,

\[
E(\hat{\psi}|\bar{x}_1, \bar{x}_0) = (\alpha_1 - \alpha_0), \text{ so } E(\hat{\psi}) = (\alpha_1 - \alpha_0).
\]

The conditional variance of \( \hat{\psi} \) can be shown to be

\[
\text{Var}(\hat{\psi}|\bar{x}_1, \bar{x}_0) = \frac{2\sigma_x^2}{n} \left\{ 1 + \frac{n}{2} \cdot \frac{(\bar{x}_1 - \bar{x}_0)^2}{(n-1)(S_{x_1}^2 + S_{x_0}^2)} \right\},
\]

where \( S_{x_i}^2 = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)/(n-1), \) the sample estimate of the variance of \( x_i \). Since the expected value of \( \hat{\psi} \) given \( x \) does not involve \( x \), the unconditional variance of \( \hat{\psi} \) is

\[
\text{Var}(\hat{\psi}) = \frac{2\sigma_x^2}{n} \left\{ 1 + \frac{n}{2} \cdot E \left[ \frac{(\bar{x}_1 - \bar{x}_0)^2}{(n-1)(S_{x_1}^2 + S_{x_0}^2)} \right] \right\}.
\]

Evaluating the expectation in (2.5.16.) is difficult when \( \sigma_{x_1}^2 \neq \sigma_{x_0}^2 \). Under random sampling, the numerator, \( (\bar{x}_1 - \bar{x}_0)^2 \), is always distributed as a multiple of a non-central \( \chi^2 \), but the denominator, \( S_{x_1}^2 + S_{x_0}^2 \), is distributed as a multiple of a \( \chi^2 \) only when \( \sigma_{x_1}^2 = \sigma_{x_0}^2 \). Otherwise, it is distributed as a weighted sum of two \( \chi^2 \) variables. When \( \sigma_{x_1}^2 = \sigma_{x_0}^2 \), say, \( n(x_1 - x_0)^2/(S_{x_1}^2 + S_{x_0}^2) \) is simply the square of a t statistic. As was pointed out by Cochran (1954), the unconditional variance is then...
\[
\text{Var}(\psi) = \frac{2\sigma^2}{n} \left\{ 1 + \frac{1}{2(n-2)} + \frac{n(\mu_x - \mu_{x_0})^2}{4(n-2)\sigma_x^2} \right\} \quad (2.5.17.)
\]

and, for the \( N(0,1) \) case Billewicz and Cochran have studied, this simplifies to
\[
\text{Var}(\hat{\psi}) = \frac{2\sigma^2}{n} \left\{ 1 + \frac{1}{2(n-2)} \right\}.
\]

Using this simplified form of (2.5.17.), we can then analytically verify the results of Billewicz's Monte Carlo investigation of the relative efficiency of \( \overline{d} \) to \( \psi \). The standard definition of the relative efficiency of \( \overline{d} \) to \( \psi \) is
\[
\text{RE}(\overline{d}, \hat{\psi}) = \frac{\text{Var}(\hat{\psi})}{\text{Var}(\overline{d})}
\]

\[
= \frac{1 + \frac{1}{2(n-2)}}{1 + \frac{\rho_{xy}^2}{1 - \rho_{xy}^2} \left\{ 1 - \sum_{k=1}^{r} \tau_k J_k^2 \right\}}
\]

when \( x_i \sim N(0,1) \) \quad (2.5.18.)

Billewicz used a different expression for the relative efficiency than that given by (2.5.18.). Fisher (1947) showed that the efficiency of the estimate of the mean of a normal population was \( 1/\sigma^2 \) when \( \sigma^2 \) was known, but that this efficiency fell to \( (\nu+1)/(\nu+3)S^2 \) when \( \sigma^2 \) had to be estimated using \( \nu \) degrees of freedom. Billewicz used the expression
\[ \frac{(\nu_d + 1)(\nu^*_\psi + 3)}{(\nu_d + 3)(\nu^*_\psi + 1)} \frac{S^2}{\nu} \]

as the relative efficiency, where \( S^2_{\nu} \) is the estimate of \( \text{Var} \left( \hat{\nu} \right) \) with \( \nu^*_\psi = 2n-3 \) degrees of freedom for error, and \( S^2_d = \sum_k \sum_m \{(y_{1km} - y_{0km}) - (\bar{y}_{1k} - \bar{y}_{0k})\}^2 / (n-r) \) is the estimate of \( \text{Var}(\overline{d}) \) with \( \nu^*_d = (n-r) \) degrees of freedom. We can see immediately that if \( \nu^*_\psi = \nu^*_d \), the form given above is the sample estimator of (2.5.18.). In general, however, \( \nu^*_\psi = \nu^*_d \). In particular, when \( \hat{\nu} \) is estimated from two independent random samples each of size \( n \), \( \nu^*_\psi \) is equal to \( 2n-3 \). The degrees of freedom for the estimator of \( \text{Var}(\overline{d}) \) depend on how it is calculated.

If the (artificial) pairing is retained within each category and the paired differences are treated as in the usual paired analysis, the degrees of freedom for estimating the within category error is then \( n_k - 1 \) for the \( k \)-th category. The estimate of variance for the average difference is then calculated from the pooled within category error and would have \( (n-r) \) degrees of freedom. (We note at this point that this pooling ignores the known heterogeneity of the within category variances and does not consider the distribution of \( \hat{\nu}^2_{1k} \). Thus, when \( \nu^*_\psi = 2n-3 \) and \( \nu^*_d = (n-r) \), the relative efficiency Billewicz calculated was

\[ \frac{(n-r+1)(2n)}{(n-r+3)(2n-2)} \frac{S^2_{\psi}}{S^2_d} \]  

(2.5.19.)

Now, if the observations had not been considered to be pair-matched within category, which would be valid since such pairing is purely
artificial, the degrees of freedom for error would have been \( \sum (2n_k - 2) = 2(n-r) \). Thus, the unnecessary pair-matching has "cost" \( n-r \) degrees of freedom.

Table 2.5.1 shows the results from Billewicz's Monte Carlo investigation. As reported earlier, the relative efficiency of category pair-matching to covariance analysis is less than one for the values he studied. The efficiency is increasing in \( n \), the sample size, and decreasing in \( \rho_{xy} \), the correlation between the matching variable \( x \) and the response variable \( y \). As is indicated, Billewicz

<table>
<thead>
<tr>
<th>( \rho_{xy} )</th>
<th>Large Sample</th>
<th>40 Pairs</th>
<th>20 Pairs</th>
<th>10 Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.98</td>
<td>0.96</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>0.4</td>
<td>0.97</td>
<td>0.94</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>0.6</td>
<td>0.89</td>
<td>0.88</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>0.8</td>
<td>0.74</td>
<td>0.73</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

\(^b\) from Billewicz (1965), Table 2.

used a 3 category structure to category-match his data. We did not indicate exactly what the category bounds were, since the table includes a series of bounds, ranging from \( A_2 = A_3 = -0.2 \) to \( A_2 = -A_3 = -1 \). Compare his results to Table 2.5.2, below. The values in this table are the population relative efficiencies based on (2.5.18.), adjusted, as were Billewicz's Monte Carlo values, for the different degrees of
freedom of the estimators respective error estimates. The values in this table are slightly higher than in Billewicz's table.

**TABLE 2.5.2**

Population Relative Efficiencies of Category Pair-Matching to Covariance Analysis for Three Equiprobable N(0,1) Intervals
(These efficiencies are corrected for different error d.f.s)

<table>
<thead>
<tr>
<th>$\rho_{xy}$</th>
<th>Large Sample</th>
<th>40 Pairs</th>
<th>20 Pairs</th>
<th>10 Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>0.4</td>
<td>0.96</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>0.6</td>
<td>0.90</td>
<td>0.89</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>0.8</td>
<td>0.73</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
</tbody>
</table>

but the same trends for $\rho_{xy}$ and n hold. It is expected that these would be slightly higher than Billewicz's values, since as was mentioned above, Billewicz combined the results for a number of different boundary points, while our table is based on only one set which is almost optimum (See Table 2.3.3).

Earlier we mentioned that there is a loss of (n-r) degrees of freedom when the data are paired, as opposed to the unpaired analysis. Now, since the expected value of the estimators of $\text{Var}(\bar{d})$ are the same when the variance is calculated from paired or unpaired data, the relative efficiency of the two methods would be $(n-r+1)(2n-2r+3)/(n-r+3)(2n-2r+1)$ when the efficiency is calculated using (2.5.18.) with Billewicz's adjustment for different degrees of freedom. Table 2.5.3 gives values of this relative efficiency of analysis using pair-matching within
category to analysis using unpaired category-matching for any three category structure. Thus, for small samples the pairing represents an appreciable loss in efficiency. The values of the relative efficiency of unpaired category matching to covariance analysis can be computed by dividing the relative efficiencies in Table 2.5.2 by the factor given above. The relative efficiency of unpaired-category matching to covariance analysis is then

\[
\frac{2n(2n-2r+1)(1 + 1/(2(n-2)))}{(2n-2)(2n-2r+3)\{1+[(\rho^2_{xy}/(1-\rho^2_{xy}))\{1 - \frac{\sum_{k} J_{k}^2}{k}\}]\}}
\]

Values of this relative efficiency are displayed in Table 2.5.4. As expected, for small values of n, the relative efficiencies are larger than the corresponding values in Table 2.5.2. The relative efficiencies are now decreasing as n increases, whereas in Table 2.5.2 they were decreasing. Note that, for small sample sizes, unpaired-category matching is as efficient as covariance analysis for low \(\rho_{xy}\).

Up to this point, we have assumed that the matching variable is identically distributed in the two groups. No attention has been given to the more realistic case where the distribution is not identical in
TABLE 2.5.4
Relative Efficiency of Unpaired Category Matching to Covariance Analysis (3 Equiprobable N(0,1) Categories)

<table>
<thead>
<tr>
<th>$\rho_{xy}$</th>
<th>Large Sample</th>
<th>40 Cases</th>
<th>20 Cases</th>
<th>10 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.991</td>
<td>1.006</td>
<td>1.026</td>
<td>1.075</td>
</tr>
<tr>
<td>0.4</td>
<td>0.962</td>
<td>0.976</td>
<td>0.996</td>
<td>1.044</td>
</tr>
<tr>
<td>0.6</td>
<td>0.896</td>
<td>0.910</td>
<td>0.927</td>
<td>0.972</td>
</tr>
<tr>
<td>0.8</td>
<td>0.731</td>
<td>0.743</td>
<td>0.757</td>
<td>0.793</td>
</tr>
</tbody>
</table>

the two groups. Recall that Miettinen (1975) defined confounding variables as those variables that were related to the response and to the disease (i.e., the case-control designation). When choosing among those variables that are related to the response, the investigator will first choose to control those that are related to the disease. Since the disease has already occurred, this essentially means that those variables related to the disease are those variables which have different distributions in the case and control groups. Likewise, the variables which Cochran (1965) classified into the first class (requiring adjustment or matching) would be those that have a different distribution in the case and control groups (assuming homogeneous slopes). The studies of Billewicz and Cochran have concentrated on the situation where the matching variable was N(0,1) in both groups. Cochran did indicate that the percent reductions in bias did tend to remain constant as $\mu_{x_0}$ and $\sigma^2_{x_0}$ differed from $\mu_{x_1}$ and $\sigma^2_{x_1}$, as was verified by our investigation of the reduction in mean difference in Section 2.3. We will consider the relative efficiency of the two estimators, their mean square error,
and some cost considerations for situations when the variable is not identically distributed in the two groups. The mean square error is of interest since the categorically matched estimator of the mean difference is biased. The mean square error offers a means of simultaneously considering the bias and the variance of an estimator. Since the cost of obtaining the samples for the two methods differs, it will be of interest to compare the relative cost of the two methods when comparing the relative efficiencies.

Earlier, we found that, for \( x_1 \sim N(\mu_{x_1}, \sigma^2_{x_1}) \),

\[
\text{BIAS} (\overline{d}) = \beta (\mu_{x_1} - \mu_{x_0} + \sum_{k=1}^{r} p_k (J_{1,k} \sigma_{x_1} - J_{0,k} \sigma_{x_0})) ,
\]

and

\[
\text{Var}(\overline{d}) = \frac{2 \sigma^2}{n} + \frac{\beta^2}{n} \sum_{k=1}^{r} p_k \left( [1 + H_{1,k} - J_{1,k}^2] \sigma^2_{x_1} + [1 + H_{0,k} - J_{0,k}^2] \sigma^2_{x_0} \right) ,
\]

and, for the covariance estimator \( \hat{\psi} \),

\[
\text{BIAS} (\hat{\psi}) = 0 ,
\]

\[
\text{Var} (\hat{\psi}) = \frac{2 \sigma^2}{n} \left\{ 1 + E \left( \frac{n(\overline{x_1} - \overline{x_0})^2}{2(n-1)(S_{x_1}^2 + S_{x_0}^2)} \right) \right\} .
\]

As we indicated above, the expectation term in the variance of \( \hat{\psi} \) can be exactly evaluated when \( \sigma^2_{x_1} = \sigma^2_{x_0} \). However, when \( \sigma^2_{x_1} \neq \sigma^2_{x_0} \), the exact
distribution of the term within the expectation is not known and must be approximated. Since under random sampling, \((n-1)S^2_{x_1}\) is distributed as \(\frac{\sigma^2_{x_1} \chi^2_{n-1}}{\chi^2_{n-1}}\),

\[
(n-1)(S^2_{x_1} + S^2_{x_0}) - \sigma^2_{x_1} \chi^2_{n-1} + \sigma^2_{x_0} \chi^2_{n-1},
\]

so that

\[
E((n-1)(S^2_{x_1} + S^2_{x_0})) = (n-1)(\sigma^2_{x_1} + \sigma^2_{x_0}),
\]

and

\[
\text{Var}((n-1)(S^2_{x_1} + S^2_{x_0})) = 2(n-1)(\sigma^4_{x_1} + \sigma^4_{x_0}).
\]

Using a result from Box (1954), we can approximate the distribution of a weighted sum of chi-squares by the distribution of a multiple of a single chi-square. This is done by choosing the multiplier and the degrees of freedom in the approximation so that the first two moments of the distributions are equal. In this way, the approximate distribution of \((n-1)(S^2_{x_1} + S^2_{x_0})\) is then taken to be:

\[
(n-1)(S^2_{x_1} + S^2_{x_0}) - \frac{(\sigma^4_{x_1} + \sigma^4_{x_0})}{(\sigma^2_{x_1} + \sigma^2_{x_0})} \chi^2 \left( \frac{(n-1)(\sigma^2_{x_1} + \sigma^2_{x_0})^2}{(\sigma^4_{x_1} + \sigma^4_{x_0})} \right).
\]

Under random sampling, the distribution of the numerator can be exactly evaluated, namely
\[
(x_1 - x_0)^2 \sim N \left( \mu_{x_1 - x_0}, \frac{(\sigma^2_{x_0} + \sigma^2_{x_1})}{n} \right),
\]

and

\[
\frac{\sqrt{n} (x_1 - x_0)}{\sqrt{\sigma^2_{x_1} + \sigma^2_{x_0}}} \sim N \left( \frac{\sqrt{n} (\mu_{x_1} - \mu_{x_0})}{\sqrt{\sigma^2_{x_1} + \sigma^2_{x_0}}}, 1 \right),
\]

so,

\[
\frac{n(x_1 - x_0)^2}{(\sigma^2_{x_1} + \sigma^2_{x_0})} \sim \chi^2_{1, \lambda},
\]

where

\[
\lambda = \frac{n(\mu_{x_1} - \mu_{x_0})^2}{2(\sigma^2_{x_1} + \sigma^2_{x_0})}.
\]

The numerator is then distributed as a multiple of a non-central chi-square variable. Under random sampling from normal distributions, the numerator and denominator are independently distributed chi-squares, so the ratio of the two is distributed as a multiple of a non-central F-distribution. Now,
\[
\frac{\frac{\mathbf{E}\left\{ \frac{n(x_1-x_0)^2}{2(n-1)(x_1^2+x_0^2)} \right\}}{2(n-1)(x_1^2+x_0^2)}}{\frac{(\sigma_{x_1}^2+\sigma_{x_0}^2)^2}{2(\sigma_{x_1}^2+\sigma_{x_0}^4)}} = \frac{\frac{n(x_1-x_0)^2}{(\sigma_{x_1}^2+\sigma_{x_0}^2)}}{\frac{2(n-1)(x_1^2+x_0^2)(\sigma_{x_1}^2+\sigma_{x_0}^2)}{(\sigma_{x_1}^4+\sigma_{x_0}^4)}} \cdot \mathbf{E}\left\{ \frac{\mathbf{F}_{1,\lambda}}{\mathbf{F}_{f,\lambda}} \right\}
\]

The numerator within the expectation is distributed as a \( X_{1,\lambda}^2 \) and the denominator is distributed as a \( X_f^2 \) where \( f = (n-1)(\sigma_{x_1}^2+\sigma_{x_0}^2)^2/(\sigma_{x_1}^4+\sigma_{x_0}^4) \).

Thus, if we divide the numerator and the denominator by their respective degrees of freedom, the ratio is distributed as a non-central \( F \) with 1 and \( f \) degrees of freedom and non-centrality parameter \( \lambda \). So,

\[
\frac{n(x_1-x_0)^2}{2(n-1)(x_1^2+x_0^2)} = \frac{2(\sigma_{x_1}^2+\sigma_{x_0}^2)^2}{2(\sigma_{x_1}^4+\sigma_{x_0}^4)} \cdot \frac{4\sigma_{x_1}^4}{(n-1)(\sigma_{x_1}^2+\sigma_{x_0}^2)^2} \cdot \mathbf{E}\left\{ \frac{\mathbf{F}_{1,\lambda}}{\mathbf{F}_{f,\lambda}} \right\}
\]

\[
= \frac{1}{2(n-1)} \mathbf{E}\left\{ \frac{\mathbf{F}_{1,\lambda}}{\mathbf{F}_{f,\lambda}} \right\}
\]

\[
= \frac{1}{2(n-1)} \left\{ \frac{f}{f-2} (1+2\lambda) \right\}
\]
\[
\begin{align*}
\frac{1}{2(n-1)} \cdot \frac{(n-1)(\sigma^2_{x_1} + \sigma^2_{x_0})^2}{(n-1)(\sigma^2_{x_1} + \sigma^2_{x_0})^2 - 2(\sigma^4_{x_1} + \sigma^4_{x_0})} \left\{ \frac{n(\mu_{x_1} - \mu_{x_0})^2}{(\sigma^2_{x_1} + \sigma^2_{x_0})} \right\}
\end{align*}
\]

\[
\begin{align*}
\frac{(\sigma^2_{x_1} + \sigma^2_{x_0}) + (\sigma^2_{x_1} + \sigma^2_{x_0})n(\mu_{x_1} - \mu_{x_0})^2}{2(n-1)(\sigma^2_{x_1} + \sigma^2_{x_0})^2 - 4(\sigma^4_{x_1} + \sigma^4_{x_0})}.
\end{align*}
\] (2.5.20.)

As a special case, when \( \sigma^2_{x_1} = \sigma^2_{x_0} = \sigma^2_x \), (2.5.20.) reduces to

\[
\frac{1}{2(n-2)} \cdot \frac{n(\mu_{x_1} - \mu_{x_0})^2}{4(n-2)\sigma^2_x}
\]

which is the same expression as that obtained by Cochran (1954).

We will use these results to compare the relative efficiency of the category-matched estimator, \( \bar{d} \), to the covariance estimator, \( \hat{\psi} \). Assume as before that \( X_1 \sim N(0,1) \) and \( X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0}) \), \( \mu_{x_0} > 0 \). Then

\[
\text{Var}(\hat{\psi}) \approx \frac{2\sigma^2_e}{n} \left\{ \frac{(1+\sigma^2_{x_0})^2 + (1+\sigma^2_{x_0})n(\mu_{x_0})^2}{2(n-1)(1+\sigma^2_{x_0})^2 - 4(1+\sigma^4_{x_0})} \right\}
\]
\[
\text{Var}(\bar{d}) = \frac{2\sigma_e^2}{n} \left\{ 1 + \frac{\beta^2}{2\sigma_e^2} \sum_{k=1}^{r} p_k \left[ 1 + H_{1,k} - J_{1,k}^2 \right]
+ \sigma_x^2 \left[ 1 + H_{0,k} - J_{0,k}^2 \right] \right\}.
\]

Then, since \(\sigma_e^2 = \sigma_{y_1}^2 (1 - \rho_{x_1y_1}^2)\),

\[
\text{Var}(\bar{d}) = \frac{2\sigma_e^2}{n} \left\{ 1 + \frac{\rho_{x_1y_1}^2}{2(1 - \rho_{x_1y_1}^2)} \sum_{k=1}^{r} p_k \left[ 1 + H_{1,k} - J_{1,k}^2 \right]
+ \sigma_x^2 \left[ 1 + H_{0,k} - J_{0,k}^2 \right] \right\},
\]

where \(\rho_{x_1y_1}^2\) is the correlation between \(x_1\) and \(y_1\).

As was considered earlier, there is a difference in the degrees of freedom for error available to estimate the variance of each estimator. The covariance estimator \(\hat{\psi}\) has \((2n-3)\) degrees of freedom available for error estimation, while the category-matched estimator \(\bar{d}\) has \(\sum_{k} 2(n_k-1) = (2n-2k)\) degrees of freedom for estimation when the observations are not paired and the variance of \(\bar{d}\) is estimated using the pooled within category sums of squares for error. If we compute the relative efficiency in the manner suggested by Billewicz, then the relative efficiency is

\[
\frac{2n(2n-2r+1)}{2(n-1)(2n-2r+3)} \frac{\text{Var}(\hat{\psi})}{\text{Var}(\bar{d})} \quad (2.5.21.)
\]
Table (2.5.5.) shows computed values of (2.5.21.) for three categories that are equiprobable based on the case \( N(0,1) \) population. As before, it was assumed that \( p_k = \pi_k \). The tables cover the values of \( \mu_{x_0} = 0.0, (0.2), 1.2, \sigma_{x_0}^2 = 1, 3; \rho_{x_1 y_1} = 0.2, 0.4, 0.6; n = 10, 20, 40, \) large samples. The relative efficiency decreases with increasing sample size, \( \rho_{x_1 y_1} \), and \( \sigma_{x_0}^2 \). It increases with increasing \( \mu_{x_0} \). Thus, for large \( \mu_{x_0} \) relative to \( \sigma_{x_0}^2 \), the category-matched estimator has a smaller variance than the covariance estimator for small sample sizes and small correlations and is, therefore, a more efficient estimator. However, for large samples and/or large correlations, the covariance variance estimator is the most efficient estimator. In practice, when there are small \( \mu_{x_0} \) and small correlations, either estimator may be used since their variances are almost equal, although the covariance estimator is simpler to use, since it requires only random samples.

These comparisons of relative efficiency do not take into account the known bias of \( \bar{a} \), and, thus, do not accurately describe the relative merits of the two estimators under our assumed model. One measure that simultaneously takes into account both variance and the squared bias is the mean square error of the estimator. The mean square error of an estimator, \( \hat{\theta} \), say, is defined to be

\[
\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{BIAS}^2(\hat{\theta}) .
\]  

(2.5.22.)

The mean square error contains both the bias and the variance on the same scale and gives equal weights to both quantities. When \( X_1 \sim N(0,1) \) and \( X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2) \), (2.5.22.) becomes
TABLE 2.5.5

Relative Efficiency of Category Matching to Covariance Analysis Adjusted for the Degrees of Freedom for Error

\[ x_0 \sim N(\mu_{x_0}, \sigma^2_{x_0}), x_1 \sim N(0,1) \]

(3 Equiprobable N(0,1) Categories)

\[ \rho_{x_1y_1} = 0.2 \]

\[ \rho_{x_1y_1} = 0.4 \]

<table>
<thead>
<tr>
<th>( \mu_{x_0} )</th>
<th>( \sigma^2_{x_0} )</th>
<th>Large Sample</th>
<th>n=40</th>
<th>n=20</th>
<th>n=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>0.9915</td>
<td>1.0035</td>
<td>1.0147</td>
<td>1.0328</td>
</tr>
<tr>
<td>0.0</td>
<td>3.0</td>
<td>0.9825</td>
<td>0.9946</td>
<td>1.0059</td>
<td>1.0254</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>1.0013</td>
<td>1.0139</td>
<td>1.0256</td>
<td>1.0449</td>
</tr>
<tr>
<td>0.2</td>
<td>3.0</td>
<td>0.9874</td>
<td>0.9997</td>
<td>1.0114</td>
<td>1.0316</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>1.0310</td>
<td>1.0451</td>
<td>1.0584</td>
<td>1.0812</td>
</tr>
<tr>
<td>0.4</td>
<td>3.0</td>
<td>1.0021</td>
<td>1.0152</td>
<td>1.0278</td>
<td>1.0502</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
<td>1.0804</td>
<td>1.0971</td>
<td>1.1131</td>
<td>1.1418</td>
</tr>
<tr>
<td>0.6</td>
<td>3.0</td>
<td>1.0265</td>
<td>1.0411</td>
<td>1.0552</td>
<td>1.0811</td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>1.1496</td>
<td>1.1698</td>
<td>1.1896</td>
<td>1.2266</td>
</tr>
<tr>
<td>0.8</td>
<td>3.0</td>
<td>1.0606</td>
<td>1.0773</td>
<td>1.0936</td>
<td>1.1243</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.2385</td>
<td>1.2633</td>
<td>1.2880</td>
<td>1.3356</td>
</tr>
<tr>
<td>1.0</td>
<td>3.0</td>
<td>1.1045</td>
<td>1.237</td>
<td>1.1492</td>
<td>1.1719</td>
</tr>
<tr>
<td>1.2</td>
<td>1.0</td>
<td>1.3471</td>
<td>1.3775</td>
<td>1.4082</td>
<td>1.4688</td>
</tr>
<tr>
<td>1.2</td>
<td>3.0</td>
<td>1.1582</td>
<td>1.1805</td>
<td>1.2031</td>
<td>1.2478</td>
</tr>
</tbody>
</table>
TABLE 2.5.5

\[ \rho_{x_1y_1} = 0.4 \] continued...

<table>
<thead>
<tr>
<th>( \mu_{x_0} )</th>
<th>( \sigma^2_{x_0} )</th>
<th>Large Sample</th>
<th>n=40</th>
<th>n=20</th>
<th>n=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>1.1139</td>
<td>1.1334</td>
<td>1.1522</td>
<td>1.1885</td>
</tr>
<tr>
<td>0.8</td>
<td>3.0</td>
<td>0.9968</td>
<td>1.0124</td>
<td>1.0278</td>
<td>1.0567</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.1990</td>
<td>1.2230</td>
<td>1.2470</td>
<td>1.2931</td>
</tr>
<tr>
<td>1.0</td>
<td>3.0</td>
<td>1.0372</td>
<td>1.0552</td>
<td>1.0732</td>
<td>1.1080</td>
</tr>
<tr>
<td>1.2</td>
<td>1.0</td>
<td>1.3030</td>
<td>1.3324</td>
<td>1.3621</td>
<td>1.4207</td>
</tr>
<tr>
<td>1.2</td>
<td>3.0</td>
<td>1.0864</td>
<td>1.1074</td>
<td>1.1285</td>
<td>1.1705</td>
</tr>
</tbody>
</table>

\[ \rho_{x_1y_1} = 0.6 \]

<table>
<thead>
<tr>
<th>( \mu_{x_0} )</th>
<th>( \sigma^2_{x_0} )</th>
<th>Large Sample</th>
<th>n=40</th>
<th>n=20</th>
<th>n=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>0.8958</td>
<td>0.9067</td>
<td>0.9168</td>
<td>0.9331</td>
</tr>
<tr>
<td>0.0</td>
<td>3.0</td>
<td>0.8066</td>
<td>0.8165</td>
<td>0.8258</td>
<td>0.8418</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>0.9045</td>
<td>0.9158</td>
<td>0.9264</td>
<td>0.9438</td>
</tr>
<tr>
<td>0.2</td>
<td>3.0</td>
<td>0.8103</td>
<td>0.8204</td>
<td>0.8300</td>
<td>0.8466</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>0.9304</td>
<td>0.9431</td>
<td>0.9551</td>
<td>0.9757</td>
</tr>
<tr>
<td>0.4</td>
<td>3.0</td>
<td>0.8216</td>
<td>0.8324</td>
<td>0.8427</td>
<td>0.8610</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
<td>0.9734</td>
<td>0.9884</td>
<td>1.0029</td>
<td>1.0288</td>
</tr>
<tr>
<td>0.6</td>
<td>3.0</td>
<td>0.8404</td>
<td>0.8523</td>
<td>0.8639</td>
<td>0.8850</td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>1.0336</td>
<td>1.0517</td>
<td>1.0696</td>
<td>1.1028</td>
</tr>
<tr>
<td>0.8</td>
<td>3.0</td>
<td>0.8665</td>
<td>0.8801</td>
<td>0.8934</td>
<td>0.9185</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.1106</td>
<td>1.1328</td>
<td>1.1550</td>
<td>1.1977</td>
</tr>
<tr>
<td>1.0</td>
<td>3.0</td>
<td>0.9000</td>
<td>0.9156</td>
<td>0.9312</td>
<td>0.9614</td>
</tr>
<tr>
<td>1.2</td>
<td>1.0</td>
<td>1.2043</td>
<td>1.2315</td>
<td>1.2590</td>
<td>1.3131</td>
</tr>
<tr>
<td>1.2</td>
<td>3.0</td>
<td>0.9407</td>
<td>0.9589</td>
<td>0.9772</td>
<td>1.0136</td>
</tr>
</tbody>
</table>
\[ \text{MSE}(\hat{\psi}) = \text{Var}(\hat{\psi}) + \text{BIAS}^2(\hat{\psi}) = \text{Var}(\hat{\psi}) \]

\[
= \frac{2\sigma^2}{n} \left\{ 1 + \frac{(1+\sigma^2_{x0})^2 + (1+\sigma^2_{x0})n\mu^2_{x0}}{2(n-1)(1+\sigma^2_{x0})^2 - 4(1+\sigma^4_{x0})} \right\},
\]

(2.5.23.)

and,

\[ \text{MSE}(\bar{d}) = \text{Var}(\bar{d}) + \text{BIAS}(\bar{d}) \]

\[
= \text{var}(\bar{d}) + \beta^2 [-\mu_{x0} + \sum_{k=1}^{r} p_k \{J_{1,k} - \sigma_{x0} J_{0,k}\}]^2
\]

\[
= \text{var}(\bar{d}) + \frac{2\sigma^2}{n} \cdot \frac{n\delta^2}{2\sigma^2} [-\mu_{x0} + \sum_{k=1}^{r} p_k \{J_{1,k} - \sigma_{x0} J_{0,k}\}]^2
\]

\[
= \frac{2\sigma^2}{n} \left\{ 1 + \frac{\rho^2_{x1y1}}{2(1-\rho^2_{x1y1})} \left[ \sum_{k=1}^{r} p_k [(1+H_{1,k} - J_{0,k})^2 + \sigma^2_{x0}(1+H_{0,k} - J^2_{0,k})] \right. \\
+ n[-\mu_{x0} + \sum_{k=1}^{r} p_k \{J_{1,k} - \sigma_{x0} J_{0,k}\}]^2 \right\}. \quad (2.5.24.)
\]
We will use (2.5.23.) and (2.5.24.) to compare the mean square error of
the estimators. The adjusted relative mean square error of category-
matching relative to covariance analysis is

\[
\frac{2n(2n-2r+1) \text{MSE}(\hat{\psi})}{2(n-1)(2n-2r+3) \text{MSE}(\hat{d})}
\]  

(2.5.25.)

Table 2.5.6 shows values of (2.5.25.) presented in the same format as
was Table 2.5.5. The asymptotic value of the ratio is zero when
\(\mu_{x_0} \neq 0\) and \(\rho_{x_1y_1} \neq 0\) since the bias of \(\tilde{d}\) does not depend on \(n\). Thus,
as \(n\) increases, the MSE(\(\tilde{d}\)) approaches BIAS(\(\tilde{d}\)), while MSE(\(\hat{\psi}\)) approaches
zero and their ratio approaches zero. The adjustment in (2.5.25.) is
the same as that used in (2.5.22.) and adjusts the ratio for the different
degrees of freedom in the two estimators. Table 2.5.6. shows the
same trends as does Table 2.5.5, although the increase in the adjusted
relative mean square error as \(\mu_{x_0} \neq 0\) increases is not as dramatic as that
of the adjusted relative efficiency. This is to be expected, since the
denominator of the adjusted relative mean square error includes a
squared bias term which makes it larger than the denominator of the
adjusted relative efficiency. For the most part, the mean square error
is less than that of \(\hat{\psi}\) for small samples and when the linear relation-
ship \((y_{1km} = \alpha_i + \beta x_{1km} + e_{1km})\) is weak \((\rho_{x_1y_1} < 0.4)\). Under these
conditions, the bias of \(\tilde{d}\) is very small, since, when \(X_i \sim N(0,1),\)

\[
\text{BIAS (}\tilde{d}\) = -\beta (\mu_{x_0} + \sum_{k=1}^{r} \pi_{1,k} x_{0,k} \sigma_{x_0})
\]
TABLE 2.5.6

MSE(\(\hat{\psi}\))/MSE(\(\bar{d}\)) Adjusted for d.f. for Error

\(X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0}), X_1 \sim N(0,1)\)

(3 Equi probable N(0,1) Categories)

\(\rho_{x_1y_1} = 0.2\)

<table>
<thead>
<tr>
<th>(\mu_{x_0})</th>
<th>(\sigma^2_{x_0})</th>
<th>Large Sample</th>
<th>n=40</th>
<th>n=20</th>
<th>n=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>0.9915</td>
<td>1.0035</td>
<td>1.0147</td>
<td>1.0328</td>
</tr>
<tr>
<td>0.0</td>
<td>3.0</td>
<td>0.9825</td>
<td>0.9946</td>
<td>1.0059</td>
<td>1.0254</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>0.8505</td>
<td>1.0132</td>
<td>1.0252</td>
<td>1.0447</td>
</tr>
<tr>
<td>0.2</td>
<td>3.0</td>
<td>0.8298</td>
<td>0.9990</td>
<td>1.0110</td>
<td>1.0314</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>0.5998</td>
<td>1.0421</td>
<td>1.0569</td>
<td>1.0805</td>
</tr>
<tr>
<td>0.4</td>
<td>3.0</td>
<td>0.5682</td>
<td>1.0122</td>
<td>1.0263</td>
<td>1.0494</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
<td>0.4073</td>
<td>1.0899</td>
<td>1.1094</td>
<td>1.1399</td>
</tr>
<tr>
<td>0.6</td>
<td>3.0</td>
<td>0.3759</td>
<td>1.0339</td>
<td>1.0516</td>
<td>1.0792</td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>0.2855</td>
<td>1.1552</td>
<td>1.1825</td>
<td>1.2229</td>
</tr>
<tr>
<td>0.8</td>
<td>3.0</td>
<td>0.2582</td>
<td>1.0640</td>
<td>1.0868</td>
<td>1.1208</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.2098</td>
<td>1.2390</td>
<td>1.2755</td>
<td>1.3291</td>
</tr>
<tr>
<td>1.0</td>
<td>3.0</td>
<td>1.1866</td>
<td>1.1020</td>
<td>1.1317</td>
<td>1.1741</td>
</tr>
<tr>
<td>1.2</td>
<td>1.0</td>
<td>0.1610</td>
<td>1.2381</td>
<td>1.3878</td>
<td>1.4580</td>
</tr>
<tr>
<td>1.2</td>
<td>3.0</td>
<td>0.1413</td>
<td>1.1475</td>
<td>1.1860</td>
<td>1.2389</td>
</tr>
</tbody>
</table>

\(\rho_{x_1y_1} = 0.4\)

<table>
<thead>
<tr>
<th>(\mu_{x_0})</th>
<th>(\sigma^2_{x_0})</th>
<th>Large Sample</th>
<th>n=40</th>
<th>n=20</th>
<th>n=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>0.9621</td>
<td>0.9738</td>
<td>0.9846</td>
<td>1.0022</td>
</tr>
<tr>
<td>0.0</td>
<td>3.0</td>
<td>0.9249</td>
<td>0.9362</td>
<td>0.9479</td>
<td>0.9653</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>0.5437</td>
<td>0.9807</td>
<td>0.9936</td>
<td>1.0131</td>
</tr>
<tr>
<td>0.2</td>
<td>3.0</td>
<td>0.5113</td>
<td>0.9379</td>
<td>0.9504</td>
<td>0.9702</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>0.2388</td>
<td>1.0010</td>
<td>1.0202</td>
<td>1.0455</td>
</tr>
<tr>
<td>0.4</td>
<td>3.0</td>
<td>0.2201</td>
<td>0.9429</td>
<td>0.9608</td>
<td>0.9849</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
<td>0.1258</td>
<td>1.0334</td>
<td>1.0636</td>
<td>1.0990</td>
</tr>
<tr>
<td>0.6</td>
<td>3.0</td>
<td>0.1144</td>
<td>0.9508</td>
<td>0.9779</td>
<td>1.0092</td>
</tr>
</tbody>
</table>
TABLE 2.5.6

\[ \rho_{y_1} = 0.4 \text{ continued...} \]

\[ \rho_{y_1} = 0.6 \]

<table>
<thead>
<tr>
<th>( \mu_{x_0} )</th>
<th>( \sigma^2_{x_0} )</th>
<th>Large Sample</th>
<th>( n=40 )</th>
<th>( n=20 )</th>
<th>( n=10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>0.0773</td>
<td>1.0758</td>
<td>1.1226</td>
<td>1.1728</td>
</tr>
<tr>
<td>0.8</td>
<td>3.0</td>
<td>0.0694</td>
<td>0.9611</td>
<td>1.0010</td>
<td>1.0428</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.0528</td>
<td>1.1254</td>
<td>1.1951</td>
<td>1.2656</td>
</tr>
<tr>
<td>1.0</td>
<td>3.0</td>
<td>0.0469</td>
<td>0.9730</td>
<td>1.0297</td>
<td>1.0851</td>
</tr>
<tr>
<td>1.2</td>
<td>1.0</td>
<td>0.0388</td>
<td>1.1788</td>
<td>1.2788</td>
<td>1.3759</td>
</tr>
<tr>
<td>1.2</td>
<td>3.0</td>
<td>0.0341</td>
<td>0.9857</td>
<td>1.0629</td>
<td>1.1355</td>
</tr>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>0.8958</td>
<td>0.9067</td>
<td>0.9168</td>
<td>0.9331</td>
</tr>
<tr>
<td>0.0</td>
<td>3.0</td>
<td>0.8066</td>
<td>0.8165</td>
<td>0.8258</td>
<td>0.8418</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>0.2859</td>
<td>0.9079</td>
<td>0.9224</td>
<td>0.9418</td>
</tr>
<tr>
<td>0.2</td>
<td>3.0</td>
<td>0.2610</td>
<td>0.8136</td>
<td>0.8265</td>
<td>0.8448</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>0.0953</td>
<td>0.9111</td>
<td>0.9386</td>
<td>0.9672</td>
</tr>
<tr>
<td>0.4</td>
<td>3.0</td>
<td>0.0869</td>
<td>0.8052</td>
<td>0.8287</td>
<td>0.8538</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
<td>0.0461</td>
<td>0.9149</td>
<td>0.9641</td>
<td>1.0085</td>
</tr>
<tr>
<td>0.6</td>
<td>3.0</td>
<td>0.0418</td>
<td>0.7917</td>
<td>0.8321</td>
<td>0.8648</td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>0.0274</td>
<td>0.9170</td>
<td>0.9964</td>
<td>1.0638</td>
</tr>
<tr>
<td>0.8</td>
<td>3.0</td>
<td>0.0246</td>
<td>0.7739</td>
<td>0.8361</td>
<td>0.8881</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.0184</td>
<td>0.9155</td>
<td>1.0324</td>
<td>1.1306</td>
</tr>
<tr>
<td>1.0</td>
<td>3.0</td>
<td>0.0163</td>
<td>0.7527</td>
<td>0.8403</td>
<td>0.9120</td>
</tr>
<tr>
<td>1.2</td>
<td>1.0</td>
<td>0.0134</td>
<td>0.9085</td>
<td>1.0689</td>
<td>1.2059</td>
</tr>
<tr>
<td>1.2</td>
<td>3.0</td>
<td>0.0118</td>
<td>0.7288</td>
<td>0.8440</td>
<td>0.9394</td>
</tr>
</tbody>
</table>
\[
\begin{align*}
&= -\sigma e^{\beta^2 - \frac{\beta^2}{\sigma^2} (\mu_x + \sigma_x \sum_{k=1}^{I} \pi_{1,k} + J_{0,k})} \\
&= -\sigma e^{\left(\frac{\rho_{x_1y_1}^2}{(1-\rho_{x_1y_1}^2)}\right)^{1/2}} \left(\mu_x + \sigma_x \sum_{k=1}^{I} \pi_{1,k} + J_{0,k}\right).
\end{align*}
\]

The factor \((\rho_{x_1y_1}^2 / (1-\rho_{x_1y_1}^2))^{1/2}\) will be quite small when \(\rho_{x_1y_1}\) is close to zero. Values of \(|\rho_{x_1y_1}|\) and \((\rho_{x_1y_1}^2 / (1-\rho_{x_1y_1}^2))^{1/2}\) are shown in Table 2.5.7. However, for large \(\rho_{x_1y_1}\) the bias can be large, and it is in these situations that the covariance estimator is better.

<table>
<thead>
<tr>
<th>(\rho_{x_1y_1})</th>
<th>(0.1)</th>
<th>(0.2)</th>
<th>(0.3)</th>
<th>(0.4)</th>
<th>(0.5)</th>
<th>(0.6)</th>
<th>(0.7)</th>
<th>(0.8)</th>
<th>(0.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\rho_{x_1y_1}^2 / (1-\rho_{x_1y_1}^2))</td>
<td>(0.10)</td>
<td>(0.20)</td>
<td>(0.31)</td>
<td>(0.42)</td>
<td>(0.58)</td>
<td>(0.75)</td>
<td>(0.98)</td>
<td>(1.33)</td>
<td>(2.07)</td>
</tr>
</tbody>
</table>

In summary we suggest that, for small sample sizes and small correlations, the category-matched estimator would be a better estimator than the covariance estimator under the given assumptions, which are: we have a normally distributed matching variable, and the analysis of covariance model is correct. There is one word of caution regarding the use of \(\bar{d}\). The distribution of its estimator of error is not known.
It seems reasonable to assume that for small \( \rho_{x_1y_1} \) it would be similar to a multiple of a \( \chi^2_{2(n-r)} \), since the contribution of \( \sigma^2_{x_1} \) and \( \sigma^2_{x_0} \) to \( \text{Var}(d) \) is quite small (See (2.5.15.).) However, as \( \rho_{x_1y_1} \) increases the contributions of \( \sigma^2_{x_1} \) and \( \sigma^2_{x_0} \) to \( \text{Var}(d) \) would become large and their effects on the distribution of \( \text{Var}(d) \) is not known.

There is one further point to consider in these comparisons. In Section 2.4, we discussed the distribution of the least sample size required to match a given case sample. Now, if we compare the cost of obtaining the estimate it can be seen that the covariance estimate can be much less costly to obtain than the category-matched estimate. Table 2.4.3 shows the 90% bound on the least sample size. This is the number of controls that will have to be examined to be 90% sure of matching all the controls. For the no bias case (\( \mu_{x_0} = 0.0 \)), the bound is \( \sim 1.4 \) times the case group sample size; and when the mean of the control group is large (\( \mu_{x_0} = 1.0 \)), the ratio has increased to \( \sim 5.6 \). So, a considerable number of observations must be examined to find the required number of matches. Covariance analysis requires only a random sample of the desired size. So, if we assume that the cost of analysis is linear in number of controls sampled, then covariance estimator will be uniformly more cost efficient than the category-matched estimator.
CHAPTER III
CALIPER MATCHING

3.1. Introduction

In Section 2.1., we mentioned that there are basically two approaches to pair-matching observations on a continuous matching variable. The first approach, which we considered in the last chapter, is to force a discrete structure upon the continuous variable by dividing its range into a series of mutually exclusive intervals. The observations are then classified into categories based on the values of the matching variable. The second approach is to form a criterion for determining when two observations are close enough to be considered "identical". There are two types of pair-matching that use this approach. The first of these, which we will consider in this chapter, is known as "caliper matching" or, equivalently, as "matching with a tolerance". Using this method, a control observation is eligible to be paired with a case observation when the control observation is within the caliper or tolerance around the case observation. The second method, which we will consider in the next chapter, is known as "nearest neighbor pair-matching". As its name suggests, it pairs a case observation with the control that is "closest" to it.

The technique of caliper matching does have an intuitive appeal to it. For a small caliper, it seems reasonable to believe that the corresponding expected response values will differ only by the amount due to the case-control difference, and, thus, the within-pair difference
should give a reasonable estimate of the case-control difference. However, there is no strict definition as to what constitutes a "small" caliper. In the statistical literature, we have found only one reference to the properties of this technique. Cochran and Rubin (1974, p. 421) present a small table (p. 421) showing the percent reduction in bias when \( x \) is normal. In passing, they note that nothing is known concerning sample size requirements implied by various caliper values and by the relative values of the mean and variance of the case and control matching variable distributions.

The results of our study of this technique are presented in a similar fashion to the results on categorical matching. We first consider the general post-sampling distribution of the caliper matched variable and these results are applied to the normal distribution. We will consider the values of the percent reduction in the mean difference of the case and control matching variable due to caliper matching and the variance of the post-caliper matching distribution of the control-matching variable. Recall from the last chapter that the percent reduction in mean difference due to caliper matching is equal to the percent reduction in the bias of the mean difference estimator of the case-control difference when it is assumed that the response and matching variables are related by the usual linear analysis of covariance model. We will then present an approximate distribution for the minimum number of randomly sampled controls necessary to caliper-match a case sample using a given caliper. This will enable us to judge the cost of using a given caliper. Finally, we will compare the mean difference estimator based on caliper matching to the analysis of covariance estimator based on random samples.
3.2. The Technique of Caliper Matching and Post-Sampling Properties of Caliper Matched Variables

For a given caliper, the technique of caliper matching is quite simple to apply. It is assumed that a case sample of size \( n_1 \) has already been collected and is randomly ordered. A control observation is then randomly selected. This observation is compared with each of the case samples' matching variable values until it first falls within the caliper of some observation. The case and control observations are then considered to be a match or a pair and are not considered further in the matching process. If a control value is not within the caliper of any case, it is discarded. This process is repeated until all the cases have been matched or all the controls have been examined. The observations which have been matched are then used in the analysis. A variation of this technique would be to order the case observations so that those that will be the most difficult to match are the first to be examined. For example, for normally distributed variables, if the mean of the control group is greater than that of the case group, the observations from the lower tail of the case distribution would be the most difficult to match, since the probability of a control group observation being that low would generally be quite small. Thus, if the case group observations are ordered from low to high, the lower observations are examined first. When there are overlapping calipers, this would match the lower tail observations first, thus, possibly lowering the number of case observations left unmatched after all matching has been done. Of course, when the control pool is quite large relative to the case group and the matching variable distribution is similar in the two groups, all observations would probably be matched anyway; but,
otherwise, the variation given should produce more matches. However, we will not consider this alternative approach in the following development, as the statistical theory becomes quite involved.

We would like to emphasize several points about caliper matching. First, the matches produced by this technique are much less arbitrary than those produced by category matching. Using category matching, the so-called matches have nothing more in common than their common membership in the same category. Any case in a given category could be legitimately matched with any control in the same category. This is not so with caliper matching. Two cases that have similar matching variable values will generally have some overlap in the region for which they will each accept a control observation as a match. In this region of overlap, the assignment of a control to one particular case is arbitrary. However, outside this region of overlap, a control can only be assigned to one or the other of the two cases. In actual practice, of course, there will be regions where several intervals overlap. However, the arbitrariness of the matching is still much less pronounced than it is for category matching. Secondly, note that the distribution of the case matching variate is not affected by this procedure. In category matching, the post-matching distribution of the matching variable depended on the weights used to sample each category. The situation we considered in the last chapter mostly considered the situation \( p_k = \pi_{1k} \), so that the case distribution was unchanged; but, as we pointed out in Section 2.2, other weights could, of course, be chosen. The next two points were made by Cochran and Rubin (1974, p. 420) in their discussion of caliper matching. Tight caliper matching should remove almost all of the bias in estimating the true mean difference between the response variables.
for any smooth regression that is common in the two populations, and, the mean response difference for a tight caliper match can be used to investigate trends with respect to the matching variate since its value "is close to a constant for any pair".

We will now investigate the post caliper matching properties of the control group for general distributions. Let \( X_1 \) be the matching variable in the case group and \( X_0 \) be the matching variable in the control group. \( X_1 \) is distributed according to the density function \( f_1(x) \) over the region of interest, with mean \( \mu_1 \) and variance \( \sigma_1^2 \). The response variable is \( Y_1 \) and it is assumed that its value is a function of \( x_1 \) and an error component. For a single observation from the case group, \((x_{1j}, y)\), we wish to find a control observation that differs from it by less than a given caliper (or tolerance), which we denote by \( c \). Suppose we randomly sample the controls until an observation \( x_{0j} \) falls within that caliper. Then

\[
|x_{1j} - x_{0j}| \leq c
\]

(3.2.1.)

\( x_{1j} \) and \( x_{0j} \) are then considered to be matched. Now it is clear that \( x_{0j} \) is a randomly selected observation from the distribution of \( x_0 \) over the restricted range \((x_{1j} - c)\) to \((x_{1j} + c)\), which is a truncated form of the original distribution for \( x_0 \). So the density function of the matched observation \( x_0 \) is

\[
g(x_0, x_1, c) = \begin{cases} 
  f(x_0) / \pi_0(x_1, c) & (x_{1j} - c) < x_0 < (x_{1j} + c) \\
  0 & \text{otherwise,}
\end{cases}
\]

(3.2.2.)
\[
\pi_0(x_1, c) = \int_{x_1-c}^{x_1+c} f_0(t) \, dt.
\]

From this definition, we can see that \( g(x_0|x_1, c) \) is a conditional distribution for \( X_0 \), given \( x_1 \). The conditional moments are defined in the usual manner, so that

\[
E(X_0|x_1, c) = \mu(x_1, c) = \int_{x_1-c}^{x_1+c} t \, g(t|x_1, c) \, dt , \tag{3.2.3}
\]

and

\[
\text{Var}(X_0|x_1, c) = \sigma^2(x_1, c) = \int_{x_1-c}^{x_1+c} [t - \mu(x_1, c)]^2 g(t|x_1, c) \, dt . \tag{3.2.4}
\]

Using standard relationships, we can also define the unconditional post-caliper-matching distribution of \( X_0 \) and the associated unconditional mean and variance. The unconditional distribution is

\[
h(x_0|c) = \int_{-\infty}^{\infty} g(x_0|x_1, c) \, f_1(x_1) \, dx_1 ; \tag{3.2.5}
\]

and, the unconditional mean and variance are

\[
E(X_0|c) \equiv \mu(c) = E_{x_1}[\mu(x_1, c)]
\]

\[
= \int_{-\infty}^{\infty} \mu(x_1, c) \, f_1(x_1) \, dx_1 , \tag{3.2.6}
\]
\begin{align*}
\text{Var}(X_0|c) \equiv \sigma^2(c) &= E_{x_1}[\sigma^2(x_1,c)] + \text{Var}_{x_1}[\mu(x_1,c)] \\
&= \int_{-\infty}^{\infty} \sigma^2(x_1,c) f_1(x_1) dx_1 + \int_{-\infty}^{\infty} [u(x_1,c)-\mu(c)]^2 f_1(x_1) dx_1.
\end{align*}

(3.2.7.)

In our discussion of this method at the beginning of this section, we indicated that the pairing of the case and control observations is less arbitrary for caliper-matching than it is for category-matching. For the latter method, we showed that the within-category pairing could be dispensed with, since there are numerous observations per group within each category and the variance of the estimator ($\overline{Y}_1 - \overline{Y}_0$) of the case-control difference could be estimated using the pooled within-category sums of squares for error. However, for caliper-matching, there is only observation per "category" per group, where we use the term "category" to indicate the interval $(x_1-c)$ to $(x_1+c)$. Thus, we cannot estimate the within-category error for each group as we did for category-matching, and we will be required to use the differences $(y_{1j} - y_{0j}) = d_j$ to estimate the variance of $(\overline{Y}_1 - \overline{Y}_0)$. This variance will depend on the post-caliper-matching correlation between $X_1$ and $X_0$. For a given caliper $c$, the post-caliper-matching covariance between $X_1$ and $X_0$ is

\begin{align*}
\text{Cov}(X_1, X_0|c) &= E[(X_1-\mu_{x_1})(X_0-\mu(c))] \\
&= \int_{-\infty}^{\infty} \int_{x_1-c}^{x_1+c} x_1 x_0 f_1(x_1) g(x_0|x_1,c) dx_0 dx_1 - \mu_{x_1} \mu(c)
\end{align*}
\[
\int_{-\infty}^{\infty} x_1 \left\{ \int_{x_1 - c}^{x_1 + c} x_0 g(x_0 | x_1, c) dx_0 \right\} f_1(x_1) dx_1 - \mu_{x_1} \mu(c)
\]

\[
= \int_{-\infty}^{\infty} x_1 \mu(x_1, c) f_1(x_1) dx_1 - \mu_{x_1} \mu(c)
\]  

(3.2.8.)

If we denote the post-caliper-matching correlation between \(X_1\) and \(X_0\) given a caliper of \(c\) as \(\rho(c)\), then

\[
\rho(c) = \frac{\text{Cov}(X_1, X_0 | c)}{\sigma_{x_1} \sigma(c)}
\]  

(3.2.9.)

Since \(X_0\) increases with \(X_1\), this will always be positive; and, we would expect that \(\rho(c)\) ranges in value from zero, when \(c \to \infty\), to one, when \(c \to 0\).

As in the last chapter, we are particularly interested in the proportional reduction in the mean difference between \(X_1\) and \(X_0\) due to caliper-matching, as this gives an indication of the effectiveness of the matching method in lowering the bias in the estimator of the case-control difference parameter. As a matter of fact, this is precisely equal to the proportional reduction in bias when the linear model \(y = \alpha + \beta x + e\) holds. The initial (i.e., pre-matching) difference in the means of \(X_1\) and \(X_0\) is \((\mu_{x_1} - \mu_{x_0})\) and the difference in means after caliper-matching is \((\mu_{x_1} - \mu(c))\). So, the proportional reduction in the mean difference between \(X_1\) and \(X_0\) due to caliper-matching with a caliper of size \(c\) is
\[
\frac{(\mu_{x_1} - \mu_{x_0}) - (\mu - \mu(c))}{(\mu_{x_1} - \mu_{x_0})} = \frac{(\mu(c) - \mu_{x_0})}{(\mu_{x_1} - \mu_{x_0})}
\]

\[
= 1 - \frac{(\mu_{x_1} - \mu(c))}{(\mu_{x_1} - \mu_{x_0})}
\]

(3.2.10.)

The form of the distributions \(f_1(x)\) and \(f_0(x)\) will dictate whether or not \(\mu(c)\) can be explicitly determined. In the next section, we will examine (3.2.10.) under the assumption that \(X_1 \sim N(\mu_{x_1}, \sigma^2_{x_1})\).

3.3. **Caliper-Matching for Normally Distributed Matching Variables**

In this section we shall apply the results of the last section when \(f_0(x)\) and \(f_1(x)\) are normal distributions. The mean and the variance of \(X_0\) conditional on \(x_1\) and \(c\) are simply those of a truncated normal distribution. The mean and the variance of the truncated normal were given in the last chapter [expressions (2.3.1.) and (2.3.2.)]. We will use the following notation to indicate the upper and lower boundary points of the caliper, standardized with respect to the controls \(N(\mu_{x_0}, \sigma^2_{x_0})\)

matching variate distribution, namely

\[
L(x_1, c) = \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} , \quad U(x_1, c) = \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}}
\]

Then, applying (2.3.1.) and (2.3.2.), we have:

\[
\mu(x_1, c) = \mu_{x_0} + \frac{Z(L(x_1, c)) - Z(U(x_1, c))}{\phi(U(x_1, c)) - \phi(L(x_1, c))} \frac{\sigma_{x_0}}{\phi(U(x_1, c)) - \phi(L(x_1, c))}
\]
\[ \sigma^2(x_1, c) = \left\{ \frac{L(x_1, c)Z(L(x_1, c)) - U(x_1, c)Z(U(x_1, c))}{\phi(U(x_1, c)) - \phi(L(x_1, c))} - J^2(x_1, c) \right\} \sigma^2_{x_0} \]

\[ = \left\{ 1 + H(x_1, c) - J^2(x_1, c) \right\} \sigma^2_{x_0} \quad (3.3.2.) \]

Using these forms, the unconditional mean and variance are

\[ \mu(c) = \int_{-\infty}^{\infty} \frac{1}{\sigma_{x_0}} x_1 \left( \frac{1}{\sigma_{x_1}} \right) \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 \]

\[ = \mu_{x_0} + \sigma_{x_0} \int_{-\infty}^{\infty} J(x_1, c)Z\left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 \quad (3.3.3.) \]

and

\[ \sigma^2(c) = \int_{-\infty}^{\infty} \sigma^2_{x_0} \left( 1 + H(x_1, c) - J^2(x_1, c) \right) Z\left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 \]

\[ + \int_{-\infty}^{\infty} (\mu(x_1, c) - \mu(c))^2 Z\left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 \]
\[
\begin{align*}
= \sigma_{x_0}^2 + \sigma_{x_0}^2 \int_{-\infty}^{\infty} H(x_1, c) Z\left(\frac{x_1 - \mu_{x_0}}{\sigma_{x_1}}\right) dx_1 - \sigma_{x_0}^2 \int_{-\infty}^{\infty} J^2(x_1, c) Z\left(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}}\right) dx_1 \\
+ \int_{-\infty}^{\infty} \mu^2(x_1, c) Z\left(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}}\right) dx_1 - \mu^2(c),
\end{align*}
\]

\[
= \sigma_{x_0}^2 + \sigma_{x_0}^2 \mu_{x_0}^2 - \sigma_{x_0}^2 \int_{-\infty}^{\infty} J(x_1, c) Z\left(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}}\right) dx_1 \\
+ \sigma_{x_0}^2 \int_{-\infty}^{\infty} H(x_1, c) Z\left(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}}\right) dx_1 \\
+ \sigma_{x_0}^2 \mu_{x_0}^2 - \mu^2(c).
\]  

(3.3.4.)

Now, note that

\[
J(-x_1 + \mu_{x_0}, c) = \frac{Z\left(-\frac{(x_1 - \mu_{x_0}) - c - \mu_{x_0}}{\sigma_{x_0}}\right) - Z\left(-\frac{(x_1 - \mu_{x_0}) + c - \mu_{x_0}}{\sigma_{x_0}}\right)}{\phi\left(-\frac{(x_1 - \mu_{x_0}) + c - \mu_{x_0}}{\sigma_{x_0}}\right) - \phi\left(-\frac{(x_1 - \mu_{x_0}) - c - \mu_{x_0}}{\sigma_{x_0}}\right)}
\]

\[
= \frac{Z\left(-\frac{x_1 + c}{\sigma_{x_0}}\right) - Z\left(-\frac{x_1 - c}{\sigma_{x_0}}\right)}{\phi\left(-\frac{x_1 + c}{\sigma_{x_0}}\right) - \phi\left(-\frac{x_1 + c}{\sigma_{x_0}}\right)}
\]
\[ z \left( \frac{(x_1+c)-\mu_{x_0}+\mu_{x_0}}{\sigma_{x_0}} \right) - z \left( \frac{(x_1-c)-\mu_{x_0}+\mu_{x_0}}{\sigma_{x_0}} \right) \\ \frac{1 - \phi \left( \frac{(x_1-c)-\mu_{x_0}+\mu_{x_0}}{\sigma_{x_0}} \right)}{1 - \phi \left( \frac{(x_1+c)-\mu_{x_0}+\mu_{x_0}}{\sigma_{x_0}} \right)} \\
\]

\[ = -J(x_1+c, \mu_{x_0}) . \]

Thus, \( J(x_1,c) \) is an odd function of \( x_1 \) around \( \mu_{x_0} \). It is well known that \( Z\left(\frac{x_1-\mu_{x_1}}{\sigma_{x_1}}\right) \) is an even function of \( x_1 \) around \( \mu_{x_1} \); i.e. \( Z\left(\frac{x_1-\mu_{x_1}}{\sigma_{x_1}}\right) = Z\left(-\frac{x_1-\mu_{x_1}}{\sigma_{x_1}}\right) \) . So, if \( \mu_{x_1} = \mu_{x_0} \), the product of the two is an odd function around \( \mu_{x_0} = \mu_{x_1} \) and so

\[ \int_{-\infty}^{\infty} J(x_1,c)Z\left(\frac{x_1-\mu_{x_1}}{\sigma_{x_1}}\right) dx_1 = 0 \quad \text{when} \quad \mu_{x_1} = \mu_{x_0} . \]
Thus, when \( \mu_{x_1} = \mu_{x_0}, \mu(c) = \mu_{x_0} \) from (3.3.3), and \( \sigma^2(c) = \sigma_{x_0}^2 \)

\[ + \sigma_{x_0}^2 \int_{-\infty}^{\infty} H(x_1, c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 \text{ from (3.3.4).} \]

Now it is fairly simple to show that \( H(x_1, c) \) is an even function of \( x_1 \) around \( \mu_{x_0} \), so that \( \sigma^2(c) + \sigma_{x_0}^2 \) since \( \int_{-\infty}^{\infty} H(x_1, c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 \neq 0 \) in general. Thus, when the two distributions have the same prematching means, the post-caliper-matching means are also the same. However, the post-matching-variance of \( X_0 \) will in general be different from the initial variance. We will consider this in greater detail when we consider specific values of \( \sigma^2(c) \).

In the last section, we developed (3.2.8.) and (3.2.9.) to express the post-category-matching covariance and correlation between \( X_1 \) and \( X_0 \). If we apply (3.3.1.), (3.3.3.) and (3.3.4.) to these expressions, we find

\[
\text{Cov}(X_1, X_0 | c) = \int_{-\infty}^{\infty} x_1 \mu(x_1, c) f_1(x_1) dx_1 - \mu_{x_1} \mu(c)
\]

\[
= \int_{-\infty}^{\infty} x_1 \left( \mu_{x_0} + J(x_1, c) \sigma_{x_0} \right) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1
\]

\[- \mu_{x_1} (\mu_{x_0} + \sigma_{x_0}) \int_{-\infty}^{\infty} J(x_1, c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1
\]
\[= \mu_{x_1} \mu_{x_0} + \sigma_{x_0} \int_{-\infty}^{\infty} x_1 J(x_1, c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \, dx_1 \]

\[= \mu_{x_1} \mu_{x_0} - \sigma_{x_0} \mu_{x_1} \int_{-\infty}^{\infty} x_1 J(x_1, c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \, dx_1 \]

\[= \sigma_{x_0} \int_{-\infty}^{\infty} (x_1 - \mu_{x_1}) J(x_1 - c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \, dx_1 \quad , \quad (3.3.5.) \]

and, so

\[\rho(c) = \frac{\text{Cov}(X_1, X_0 | c)}{\sigma_{x_1} \sigma(c)} \]

\[= \frac{\sigma_{x_0}}{\sigma_{x_1} \sigma(c)} \int_{-\infty}^{\infty} (x_1 - \mu_{x_1}) J(x_1, c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \, dx_1 \quad . \quad (3.3.6.) \]

In the last section, we mentioned that this correlation should go from zero as \(c\) approached infinity and should go to one as \(c\) approached zero. We will now show that this does hold for expression (3.3.6.). In
passing, we shall also find the corresponding limiting values of \( \mu(c) \) and \( \sigma^2(c) \). Consider \( J(x_1, c) \):

\[
\lim_{c \to \infty} J(x_1, c) = \lim_{c \to \infty} \frac{Z\left(\frac{x_1^{-c} - \mu}{\sigma x_0}\right) - Z\left(\frac{x_1^{c} - \mu}{\sigma x_0}\right)}{\phi\left(\frac{x_1^{c} - \mu}{\sigma x_0}\right) - \phi\left(\frac{x_1^{-c} - \mu}{\sigma x_0}\right)}
\]

\[
= \frac{Z(-\infty) - Z(\infty)}{\phi(\infty) - \phi(-\infty)} = 0 .
\]

Thus

\[
\lim_{c \to \infty} \mu(x_1, c) = \mu x_0 \quad \text{and} \quad \lim_{c \to \infty} J(x_1, c) = \mu x_0 .
\]

And,

\[
\lim_{c \to \infty} \rho(c) = \frac{\sigma x_0}{\sigma x_1} \cdot \frac{1}{\lim_{c \to \infty} \sigma(c)} \int_{-\infty}^{\infty} (x_1^{-\mu} x_1) (\lim_{c \to \infty} J(x_1, c)) \left(\frac{x_1^{-\mu} x_1}{\sigma x_1}\right) dx_1
\]
= 0 if \( \lim_{c \to \infty} \sigma(c) > 0 \).

Now, as \( c \to \infty \), \( \mu(c) \to \mu \), and the correlation between \( X_1 \) and \( X_0 \) approaches zero if \( \lim_{c \to \infty} \sigma(c) > 0 \), which will later be shown to be true. Now,

\[
\lim_{c \to \infty} J(x_1, c) = \lim_{c \to \infty} \frac{\left( \frac{x_1 - c - \mu}{\sigma x_0} \right) - \left( \frac{x_1 + c - \mu}{\sigma x_0} \right)}{\phi \left( \frac{x_1 + c - \mu}{\sigma x_0} \right) - \phi \left( \frac{x_1 + c - \mu}{\sigma x_0} \right)}.
\]

Since both the numerator and denominator of \( J(x_1, c) \) go to zero as \( c \) goes to zero, we will use L'Hospital's rule to find this limit. In particular,

\[
\lim_{c \to \infty} J(x_1, c) = \lim_{c \to \infty} \frac{D_c \left( \frac{x_1 - c - \mu}{\sigma x_0} \right) - \frac{x_1 + c - \mu}{\sigma x_0} \right) \right)}{D_c \left( \frac{x_1 + c - \mu}{\sigma x_0} \right) - \frac{x_1 - c - \mu}{\sigma x_0} \right)} \right)}
\]

\[
= \lim_{c \to \infty} \frac{\left( \frac{x_1 - c - \mu}{\sigma x_0} \right) - \left( \frac{x_1 + c - \mu}{\sigma x_0} \right)}{\left( \frac{x_1 + c - \mu}{\sigma x_0} \right) + \left( \frac{x_1 - c - \mu}{\sigma x_0} \right) \left( \frac{1}{\sigma x_0} \right)}
\]
Thus,

$$\lim_{c \to \infty} \mu(x_1, c) = \mu_{x_0} + \sigma_{x_0} \left( \frac{x_1 - \mu_{x_0}}{\sigma_{x_0}} \right) = x_1.$$ 

so that

$$\lim_{c \to \infty} \mu(c) = \int_{-\infty}^{\infty} x_1 Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 = \mu_{x_1}.$$ 

And,

$$\lim_{c \to \infty} \rho(c) = \frac{\sigma_{x_0}}{\sigma_{x_1}} \cdot \frac{1}{\lim_{c \to \infty} \sigma(c)} \int_{-\infty}^{\infty} (x_1 - \mu_{x_1}) \left( \frac{x_1 - \mu_{x_0}}{\sigma_{x_0}} \right) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 =$$

$$= \frac{\sigma_{x_1}}{\lim_{c \to \infty} \sigma(c)}.$$ 

Now, as $c \to 0$, $\mu(c) \approx \mu_{x_1}$, and the correlation becomes $\sigma_{x_1} / \lim_{c \to \infty} \sigma(c)$, if $\lim_{c \to \infty} \sigma(c)$ is not zero. Now for $H(x_1, c)$,
\[
\lim_{c \to \infty} H(x_1, c) = \lim_{c \to \infty} \left( \frac{x_1 - \mu x_0}{\sigma x_0} \right) z \left( \frac{x_1 - \mu x_0}{\sigma x_0} \right) - \left( \frac{x_1 + c - \mu x_0}{\sigma x_0} \right) z \left( \frac{x_1 + c - \mu x_0}{\sigma x_0} \right) - \left( \frac{x_1 - \mu x_0}{\sigma x_0} \right) z \left( \frac{x_1 - \mu x_0}{\sigma x_0} \right)
\]

\[
= \lim_{c \to \infty} \left( \frac{x_1 - \mu x_0}{\sigma x_0} \right) z \left( \frac{x_1 - \mu x_0}{\sigma x_0} \right) - \lim_{c \to \infty} \left( \frac{x_1 + c - \mu x_0}{\sigma x_0} \right) z \left( \frac{x_1 + c - \mu x_0}{\sigma x_0} \right).
\]

In the last chapter, we showed that \( \lim_{w \to \infty} wZ(w) = 0 \). By the same line of reasoning, both the limits given above are also zero, and so

\[
\lim_{c \to \infty} H(x_1, c) = 0.
\]

Therefore,

\[
\lim_{c \to \infty} \sigma^2(x_1, c) = \sigma^2_{x_0} \{1 + \lim_{c \to \infty} H(x_1, c) = (\lim_{c \to \infty} J(x_1, c))^2\}
\]

\[
= \sigma^2_{x_0},
\]

and, from (3.3.4.),

\[
\lim_{c \to \infty} \sigma^2(c) = \sigma^2_{x_0} + \mu^2_{x_0} - (\lim_{c \to \infty} \mu(c))^2 + 2\mu_{x_0} \int_{-\infty}^{\infty} (\lim_{c \to \infty} J(x_1, c))
\]

\[
= \int_{-\infty}^{\infty} \left( \frac{x_1 - \mu x_1}{\sigma x_1} \right) dx_1 + \sigma^2_{x_0} \int_{-\infty}^{\infty} (\lim_{c \to \infty} H(x_1, c)) z \left( \frac{x_1 - \mu x_1}{\sigma x_1} \right) dx_1.
\]
\[ = \sigma_{x_0}^2 \]

Since \( \lim_{c \to \infty} \sigma^2(c) = \sigma_{x_0}^2 > 0, \lim_{c \to \infty} \rho(c) = 0. \) So, in the limit as \( c \to \infty, \)
\( \mu(c) \to \mu_{x_0}, \sigma^2(c) \to \sigma_{x_0}^2, \) and \( \rho(c) \to 0, \) the original, presampling values for \( X_0. \) Now,
\[
\lim_{c \to \infty} H(x_1, c) = \lim_{c \to \infty} \frac{\left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) Z \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) - \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right) Z \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right)}{\Phi \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right) - \Phi \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right)}.
\]

This is an indeterminate form at \( c = 0, \) so that we will have to use L'Hopital's rule to evaluate this limit. Thus,
\[
\lim_{c \to \infty} H(x_1, c) = \lim_{c \to \infty} \frac{D_c \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) Z \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) - \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right) Z \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right)}{D_c \left\{ \Phi \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right) - \Phi \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) \right\}}
\]
\[
= \lim_{c \to \infty} \left\{ -\frac{1}{\sigma_{x_0}} \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) Z \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) + \frac{1}{\sigma_{x_0}} \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right)^2 Z \left( \frac{x_1 - c - \mu_{x_0}}{\sigma_{x_0}} \right) \right. \\
\left. - \frac{1}{\sigma_{x_0}} Z \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right) + \frac{1}{\sigma_{x_0}} Z \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right) \left( \frac{x_1 + c - \mu_{x_0}}{\sigma_{x_0}} \right)^2 \right\}.
\]
\[
\left\{ \frac{1}{\sigma_x} \left( \frac{x_1 + \mu - \mu}{\sigma} \right) x_0 \right\} + \frac{1}{\sigma_x} \left( \frac{x_1 - \mu - \mu}{\sigma} \right) x_0 \right\}
\]

\[
= \frac{2 \left\{ -1 + \left( \frac{x_1 - \mu}{\sigma x_0} \right)^2 \right\} \left( \frac{x_1 - \mu}{\sigma x_0} \right)}{2 \left( \frac{x_1 - \mu}{\sigma x_0} \right)}
\]

\[
= \left( \frac{x_1 - \mu}{\sigma x_0} \right)^2 - 1 = \left\{ \lim_{c \to \infty} J(x_1, c) \right\}^2 - 1.
\]

Thus,

\[
\lim_{c \to \infty} \sigma^2(x_1, c) = \sigma^2_{x_0} \left\{ 1 + \lim_{c \to \infty} H(x_1, c) - (\lim_{c \to \infty} J(x_1, c))^2 \right\}
\]

\[
= \sigma^2_{x_0} \left\{ 1 - 1 + (\lim_{c \to \infty} J(x_1, c))^2 \right\} - (\lim_{c \to \infty} J(x_1, c))^2 = 0.
\]
And
\[
\lim_{c \to \infty} \sigma^2(c) = \sigma_{x_0}^2 + \mu_{x_0}^2 - (\lim_{c \to \infty} \mu(c))^2 + 2\mu_{x_0} \sigma_{x_0} \int_{-\infty}^{\infty} \lim_{c \to \infty} J(x_1, c) dx_1
\]

\[
z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 + \sigma_{x_0}^2 \int_{-\infty}^{\infty} \lim_{c \to \infty} H(x_1, c) z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1
\]

\[
= \sigma_{x_0}^2 + \mu_{x_0}^2 - \mu_{x_1}^2 + 2\mu_{x_0} \sigma_{x_0} \int_{-\infty}^{\infty} \left( \frac{x_1 - \mu_{x_0}}{\sigma_{x_0}} \right) z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1
\]

\[
+ \sigma_{x_0}^2 \int_{-\infty}^{\infty} \left\{ -1 + \left( \frac{x_1 - \mu_{x_0}}{\sigma_{x_0}} \right)^2 \right\} z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1
\]

\[
= \sigma_{x_0}^2 + \mu_{x_0}^2 - \mu_{x_1}^2 + 2\mu_{x_0} \mu_{x_1} - 2\mu_{x_0}^2 - \sigma_{x_0}^2 + \sigma_{x_1}^2
\]

\[
+ \mu_{x_1}^2 - 2\mu_{x_1} \mu_{x_0} + \mu_{x_0}^2
\]

\[
= \sigma_{x_1}^2.
\]
Thus,
\[ \lim_{c \to 0} \rho(c) = \frac{\sigma_{x_1}}{\lim_{c \to 0} \sigma(c)} = 1. \]

Thus, we have shown that as \( c \) approaches infinity, \( \rho(c) \), the post-caliper-matching correlation between \( X_1 \) and \( X_0 \), approaches zero; and, as \( c \) approaches zero, \( \rho(c) \) approaches one. We have also shown that, as \( c \) approaches infinity, the post-caliper-matching mean and variance of \( X_0 \) approach the initial mean and variance of \( X_0 \); and, as \( c \) approaches zero, they approach the mean and variance of \( X_1 \).

From (3.2.10.) and (3.3.3.), the proportional reduction in mean difference (or, equivalently, mean bias under the straight line \( y_{ij} = \alpha_i + \beta x_{ij} + e_{ij} \) model assumption) is

\[ 1 - (\mu_{x_1} - \mu(c))/(\mu_{x_1} - \mu_{x_0}) = (\mu(c) - \mu_{x_0})/(\mu_{x_1} - \mu_{x_0}) \]

\[ = -\frac{\sigma_{x_0}}{(\mu_{x_1} - \mu_{x_0})} \int_{-\infty}^{\infty} J(x_1, c) \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1. \quad (3.3.7.) \]

Unfortunately, the integrals in expressions (3.3.3.) thru (3.3.7.) cannot be explicitly evaluated. Thus, to make any further progress, we will have to resort to using numerical integration techniques. Note that the integrals of \( J(x_1, c) \) and \( H(x_1, c) \) that we have to consider are both of the general form
\[ \frac{1}{\sigma_{x_1} \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right)^2} h(x_1, c) \, dx_1, \]

where \( h(x_1, c) \) is some function of \( x_1 \) and \( c \). We will assume, as we did in the last chapter, that \( X_1 \sim N(0,1) \). The above integral is then expressible in the general form

\[ c^* \int_{-\infty}^{\infty} e^{-x_1^2} g(x_1, c) \, dx_1, \]

where \( g(x_1, c) \) is some function of \( x_1 \) and \( c \). This type of integral can be evaluated using Gaussian-Hermitian Quadrature formulas.

Cochran and Rubin (1974, pp. 420-421) offer a brief discussion of caliper-matching. They assumed that the usual straight line analysis of covariance model held. As a measure of the initial mean difference in \( X_1 \) and \( X_0 \), they used the quantity

\[ B = (\mu_{x_1} - \mu_{x_0}) \left( \frac{\sigma_{x_1}^2 + \sigma_{x_0}^2}{2} \right)^{1/2}. \]

This caliper was of the form \( c = a \sqrt{\frac{\sigma_{x_1}^2 + \sigma_{x_0}^2}{2}} \). They then present the following table of the proportional reduction in bias for \( X_1 \), \( X_0 \) normal. They also indicate that for \( 0.5 \leq B \leq 1 \) the results were similar, being slightly lower by about 1% to 1 1/2%. This indicates that a tight matching will remove almost all the bias, while a loose
TABLE 3.3.1

Percent Reduction in Bias for Caliper Matching
to Within $\pm \sqrt{\sigma^2 + \sigma^2} / 2$ With X Normal. $B < 0.5$

<table>
<thead>
<tr>
<th>$a$</th>
<th>$\sigma^2_{X_1} / \sigma^2_{X_0} = 1/2$</th>
<th>$\sigma^2_{X_1} / \sigma^2_{X_0} = 1$</th>
<th>$\sigma^2_{X_1} / \sigma^2_{X_0} = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>0.4</td>
<td>0.96</td>
<td>0.95</td>
<td>0.93</td>
</tr>
<tr>
<td>0.6</td>
<td>0.91</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>0.8</td>
<td>0.86</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>1.0</td>
<td>0.79</td>
<td>0.74</td>
<td>0.69</td>
</tr>
</tbody>
</table>


Matching will remove only about 80% of the bias. Rubin has stated, in a private conversation, that these results were obtained by numerical integration of

$$\int_{-\infty}^{\infty} \int_{x_1-c}^{x_1+c} x f_0(x) dx dx_1.$$  

The proportional reduction in $B$ that they used can be seen to be equivalent to the proportional reduction that we have been using.

As a comparison, we used (3.3.7.) to generate Table 3.3.2 below. We have assumed that $X_1 \sim N(0,1)$ and $X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$ and that the caliper was of the form $c = \sqrt{\sigma^2 + \sigma^2} / 2$. In order to be able to compare our results with those of Cochran and Rubin, we have presented results for $\mu_{x_0}$ chosen so that $B = 0.5$. As indicated in the table, this leads to values of $\mu_{x_0} = 0.5, 0.6, 0.7$ when $\sigma^2_{x_0} = 1, 2, 3$, respectively. The column for $\sigma^2_{x_0} = 1$ in this table corresponds to the column labeled...
\( \sigma^2_{x_1} / \sigma^2_{x_0} = 1 \) in Table 3.3.1, and the column labeled \( \sigma^2_{x_0} = 2 \) corresponds to
the one labeled \( \sigma^2_{x_1} / \sigma^2_{x_0} = 1/2 \). For these columns our results are essen-
tially identical with those of Cochran and Rubin. We have also in-
cluded a column for \( \sigma^2_{x_0} = 3 \). We also found that amount of bias reduction
increases with increasing variance, and also that the proportional re-
duction in mean difference increases as \( \mu_{x_0} \) increases for a given \( \sigma^2_{x_0} \)
and \( a \), indicating that the performance of this method improves under
more difficult situations. This is opposite to the effect found for
categorical-matching. Recall that the reduction in mean difference
under category-matching decreased as \( \mu_{x_0} \) increased. Although the
patterns are not strong, this still gives some indication that caliper-
matching might be preferable to category-matching; this will be con-
sidered further in the next chapter. As with Cochran and Rubin's
results, a "tight" caliper (\( a = 0.2 \)) removes essentially all the mean
difference, while a "loose" caliper (\( a = 1.0 \)) removes from 75\% to 80\%.
Note that for very small values of \( a \), there is very little change in
the proportional reduction. This indicates that from a practical
standpoint, it would be advisable to pick a fairly large caliper so
that there would be a better chance of matching all the case obser-
vations. This point is considered further in the next section, where
we discuss sample size requirements for various caliper sizes.

Table 3.3.3 gives the post-caliper-matching variance for \( X_0 \), based
on integrals (3.3.3.) and (3.3.4.), for various values of \( \sigma^2_{x_0} \), \( \mu_{x_0} \), and
\( c = a \sqrt{1 + \sigma^2_{x_0}} / 2 \). As is obvious from the table, caliper pair-matching
does quite well in reducing \( \sigma^2_{x_0} \). There are several patterns that are
worth mentioning. The first is that the value of \( \sigma^2(c) \) increases as \( c \)
<table>
<thead>
<tr>
<th>a</th>
<th>$\sigma_{x_0}^2 = 1$</th>
<th>$\sigma_{x_0}^2 = 2$</th>
<th>$\sigma_{x_0}^2 = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>0.2</td>
<td>0.987</td>
<td>0.990</td>
<td>0.991</td>
</tr>
<tr>
<td>0.3</td>
<td>0.971</td>
<td>0.978</td>
<td>0.980</td>
</tr>
<tr>
<td>0.4</td>
<td>0.949</td>
<td>0.961</td>
<td>0.965</td>
</tr>
<tr>
<td>0.5</td>
<td>0.923</td>
<td>0.940</td>
<td>0.946</td>
</tr>
<tr>
<td>0.6</td>
<td>0.893</td>
<td>0.915</td>
<td>0.924</td>
</tr>
<tr>
<td>0.7</td>
<td>0.859</td>
<td>0.887</td>
<td>0.898</td>
</tr>
<tr>
<td>0.8</td>
<td>0.823</td>
<td>0.856</td>
<td>0.869</td>
</tr>
<tr>
<td>0.9</td>
<td>0.784</td>
<td>0.823</td>
<td>0.838</td>
</tr>
<tr>
<td>1.0</td>
<td>0.745</td>
<td>0.788</td>
<td>0.805</td>
</tr>
</tbody>
</table>
TABLE 3.3.3

Post-Caliper Matching Variance of $x_0$, Using a Caliper of the Form

$$c = a \sqrt{1 + \frac{\sigma^2}{x_0}} / 2$$

for $x_1 \sim N(0,1)$, $x_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$

<table>
<thead>
<tr>
<th>$a$</th>
<th>$\frac{\sigma^2}{x_0}$</th>
<th>$\mu_{x_0}$ =</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>3</td>
<td>1.002</td>
<td>1.002</td>
<td>1.002</td>
<td>1.002</td>
<td>1.002</td>
<td>1.002</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>1.009</td>
<td>1.009</td>
<td>1.009</td>
<td>1.009</td>
<td>1.009</td>
<td>1.009</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0.952</td>
<td>0.952</td>
<td>0.952</td>
<td>0.952</td>
<td>0.953</td>
<td>0.953</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>1.002</td>
<td>1.002</td>
<td>1.002</td>
<td>1.002</td>
<td>1.001</td>
<td>1.001</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>1.036</td>
<td>1.036</td>
<td>1.036</td>
<td>1.036</td>
<td>1.036</td>
<td>1.036</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>0.904</td>
<td>0.904</td>
<td>0.905</td>
<td>0.905</td>
<td>0.907</td>
<td>0.908</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>2</td>
<td>1.007</td>
<td>1.007</td>
<td>1.007</td>
<td>1.007</td>
<td>1.007</td>
<td>1.007</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>1.082</td>
<td>1.082</td>
<td>1.082</td>
<td>1.082</td>
<td>1.081</td>
<td>1.081</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.853</td>
<td>0.854</td>
<td>0.855</td>
<td>0.857</td>
<td>0.359</td>
<td>0.962</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>2</td>
<td>1.021</td>
<td>1.021</td>
<td>1.021</td>
<td>1.020</td>
<td>1.020</td>
<td>1.019</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.148</td>
<td>1.148</td>
<td>1.148</td>
<td>1.147</td>
<td>1.146</td>
<td>1.144</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>0.809</td>
<td>0.810</td>
<td>0.811</td>
<td>0.814</td>
<td>0.817</td>
<td>0.822</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>2</td>
<td>1.046</td>
<td>1.046</td>
<td>1.045</td>
<td>1.044</td>
<td>1.043</td>
<td>1.042</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>1.235</td>
<td>1.234</td>
<td>1.233</td>
<td>1.231</td>
<td>1.229</td>
<td>1.226</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>1</td>
<td>0.777</td>
<td>0.777</td>
<td>0.779</td>
<td>0.782</td>
<td>0.786</td>
<td>0.792</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>2</td>
<td>1.084</td>
<td>1.084</td>
<td>1.083</td>
<td>1.081</td>
<td>1.078</td>
<td>1.075</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>3</td>
<td>1.341</td>
<td>1.340</td>
<td>1.338</td>
<td>1.335</td>
<td>1.330</td>
<td>1.324</td>
<td></td>
</tr>
</tbody>
</table>
increases when \( \sigma^2_{x_0} = 2 \) or 3. However, for \( \sigma^2_{x_0} = 1 \), the values of \( \sigma^2(c) \) decrease with increasing \( c \), indicating, in this case, that the post-caliper-matching variance decreases as the caliper width increases.

This pattern would seem to contradict the earlier results showing that as \( c \) goes to infinity, \( \sigma^2(c) \) approaches \( \sigma^2_{x_0} \). This apparent conflict can be explained by considering the function defining \( \sigma^2(c) \). The pattern observed for \( \sigma^2_{x_0} = 1 \) holds for all \( \mu_{x_1} \), but becomes clearest when we assume \( \mu_{x_0} = \mu_{x_1} = 0 \). Let \( \mu_{x_0} = \mu_{x_1} = 0 \) and \( \sigma^2 = 1 \). In an earlier discussion of the function \( J(x_1,c) \), we noted that this was an odd function of \( x_1 \) around \( \mu_{x_0} \), so that, when \( \mu_{x_0} = \mu_{x_1} \), 

\[
\int_{-\infty}^{\infty} J(x_1,c) \times \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) dx_1 \text{ is zero, and } \mu(c) = \mu_{x_0} = \mu_{x_1}, \text{ which in this case is also zero.}
\]

If we apply these two facts to equation (3.3.4.) for \( \sigma^2(c) \), we find that 

\[
\sigma^2(c) = \sigma^2_{x_0} \left[ 1 + \int_{-\infty}^{\infty} H(x_1,c)Z(x_1)dx_1 \right].
\]

Now, as \( c \) approaches zero, we know that \( \sigma^2(c) \) approaches \( \sigma^2_{x_0} \); and, as \( c \) approaches infinity, \( \sigma^2(c) \) approaches \( \sigma^2_{x_1} \). So, it would seem logical that \( \sigma^2(c) \) would vary between \( \sigma^2_{x_0} \) and \( \sigma^2_{x_1} \). However, this is not actually the case, since the function \( \int_{-\infty}^{\infty} H(x_1,c)Z(x_1)dx_1 \) is not a monotonic function of \( c \). Figure (3.3.1.) shows a detailed graph of \( \sigma^2(c) \) over the range \( a = 0.0 \) to \( a = 6.0 \) for \( \sigma^2_{x_0} = 0.5, 1.0, 1.5, 2.0 \). At the two extremes of \( a = 0.0 \) and \( a = 6.0 \), \( \sigma^2(c) \) reaches the limiting values of \( \sigma^2(c) = \sigma^2_{x_0} \) and \( \sigma^2(c) = \sigma^2_{x_1} \), respectively. However, as can be seen, \( \sigma^2(c) \) is not necessarily bounded by these two values. For \( \sigma^2_{x_0} = 0.5, 1.0, 1.5 \), the minimum value of \( \sigma^2(c) \) occurs at some value of \( a \) greater than 0.0, and the values of \( a \) corresponding to the minimum of \( \sigma^2(c) \) decreases as \( \sigma^2_{x_0} \) increases, till, at \( \sigma^2_{x_0} = 2 \), the minimum occurs at a
FIGURE 3.3.1
Post-Matching Variance as a Function of $c = \sqrt{\frac{\sigma^2}{(1+c^2x_0)^2}}$
<table>
<thead>
<tr>
<th>$a$</th>
<th>$\sigma^2_{x_0}$</th>
<th>$\mu_{x_0} = 0.0$</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>0.1</td>
<td>2</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>0.1</td>
<td>3</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td>0.2</td>
<td>2</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0.973</td>
<td>0.973</td>
<td>0.973</td>
<td>0.973</td>
<td>0.973</td>
<td>0.973</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>0.960</td>
<td>0.960</td>
<td>0.960</td>
<td>0.961</td>
<td>0.961</td>
<td>0.961</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>0.948</td>
<td>0.948</td>
<td>0.948</td>
<td>0.948</td>
<td>0.948</td>
<td>0.949</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>0.938</td>
<td>0.939</td>
<td>0.939</td>
<td>0.940</td>
<td>0.941</td>
<td>0.942</td>
</tr>
<tr>
<td>0.6</td>
<td>2</td>
<td>0.912</td>
<td>0.912</td>
<td>0.912</td>
<td>0.913</td>
<td>0.913</td>
<td>0.914</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>0.888</td>
<td>0.888</td>
<td>0.888</td>
<td>0.888</td>
<td>0.889</td>
<td>0.889</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.889</td>
<td>0.890</td>
<td>0.891</td>
<td>0.893</td>
<td>0.996</td>
<td>0.899</td>
</tr>
<tr>
<td>0.8</td>
<td>2</td>
<td>0.847</td>
<td>0.847</td>
<td>0.848</td>
<td>0.849</td>
<td>0.851</td>
<td>0.853</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>0.811</td>
<td>0.811</td>
<td>0.812</td>
<td>0.812</td>
<td>0.814</td>
<td>0.815</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>0.825</td>
<td>0.826</td>
<td>0.829</td>
<td>0.833</td>
<td>0.839</td>
<td>0.845</td>
</tr>
<tr>
<td>1.0</td>
<td>2</td>
<td>0.879</td>
<td>0.769</td>
<td>0.771</td>
<td>0.773</td>
<td>0.777</td>
<td>0.781</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>0.723</td>
<td>0.724</td>
<td>0.726</td>
<td>0.726</td>
<td>0.729</td>
<td>0.732</td>
</tr>
<tr>
<td>1.2</td>
<td>1</td>
<td>0.749</td>
<td>0.750</td>
<td>0.755</td>
<td>0.762</td>
<td>0.772</td>
<td>0.783</td>
</tr>
<tr>
<td>1.2</td>
<td>2</td>
<td>0.682</td>
<td>0.683</td>
<td>0.686</td>
<td>0.690</td>
<td>0.695</td>
<td>0.702</td>
</tr>
<tr>
<td>1.2</td>
<td>3</td>
<td>0.632</td>
<td>0.632</td>
<td>0.634</td>
<td>0.637</td>
<td>0.641</td>
<td>0.645</td>
</tr>
</tbody>
</table>

TABLE 3.3.4

The Post-Caliper Matching Correlation Between $X_1$ and $X_0$, Using a Caliper of the Form $\sqrt{1+\sigma^2_{x_0}}/2$, When $X_1 \sim N(0,1)$, $X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$
value of a which is quite close to zero. This same situation holds when \( \mu_{X_0} \neq \mu_{X_1} \). This suggests that when the matching variable distribution has the same or similar variance in the two populations that a wider caliper will give a smaller variance of \( X_0 \) after caliper-matching than a narrow caliper. This will be balanced by the fact that the correlation between \( X_1 \) and \( X_0 \) will not be as strong for a wider caliper as it is for a narrow caliper.

Table 3.3.4 shows the post caliper matching correlation between \( X_0 \) and \( X_1 \). As can be seen, it is almost perfect for \( a \leq 0.2 \) and decreases slowly as \( a \) increases. However, even when \( a = 1.2 \), the correlation is still quite strong and will have to be considered in any subsequent statistical analysis. We will consider this further in Section 3.5., when we discuss the analysis of caliper matched data.
3.4. **Sample Size Considerations for Caliper Matching**

Cochran and Rubin (1974) point out that, unless the ratio of the control pool size to the case sample size is quite large, "there is a non-negligible probability that some of the desired matches are not found". They indicate that nothing is known about the distribution of the number of matches that can be obtained from a given sample of \( n_0 \) controls and \( n_1 \) cases and, thus, do not offer any advice concerning the size of the control pool needed. It is possible, however, to develop an approximation to the distribution of the least control pool size needed to match a given case sample of size \( n_1 \). This will be of use in selecting an appropriate caliper width and in studying the amount of bias removed by various caliper widths.

The exact distribution of the least control pool size would be quite difficult to derive. However, we can develop an approximation. Consider the following argument: A given value of \( x_1 \) will be caliper matched if a randomly selected observation from the control pool has an \( X_0 \) value in the interval \((x_1 - c, x_1 + c)\). The probability of this event is

\[
\phi \left( \frac{x_1 + c - \mu}{\sigma} x_0 \right) - \phi \left( \frac{x_1 - c - \mu}{\sigma} x_0 \right) \quad \text{under the assumption that } X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0}).
\]

Now, since \( x_1 \) is a randomly selected value from a \( N(\mu_{x_1}, \sigma^2_{x_1}) \) distribution, the average probability \((p_c)\), say, of matching \( x_1 \) is

\[
p_c = \int_{-\infty}^{\infty} \left\{ \phi \left( \frac{x_1 + c - \mu}{\sigma} x_0 \right) - \phi \left( \frac{x_1 - c - \mu}{\sigma} x_0 \right) \right\} \Phi \left( \frac{x_1 - \mu}{\sigma} x_1 \right) \, dx_1. \quad (3.4.1)
\]
If we have a case sample of size \( n_1 \) that we wish to match, we can approximate the distribution of the least control pool size required to match these \( n_1 \) cases, that case with value \( x_1 \) having probability
\[
\Phi \left( \frac{x_1 + c - u_{X_0}}{\sigma_{X_0}} \right) - \Phi \left( \frac{x_1 - c - u_{X_0}}{\sigma_{X_0}} \right)
\]
of being matched by a randomly selected control, by the distribution of the least control pool size \( (N_0, \text{say}) \) required to match \( n_1 \) case observations, each of which has a probability \( p_c \) of being matched. This is simply the negative binomial distribution with parameters \( n_1 \) and \( p_c \) where \( n_1 \) is the number of matches desired and \( p(c) \) is the probability of a match. The probability distribution of \( N_0 \) is
\[
P_{N_0}(n_0) = \binom{n_0 - 1}{n_1 - 1} \binom{n_0 - n_1}{n_1} p_c (1-p_c)^{n_0 - n_1} \quad n_0 = n_1, n_1+1, n_1+2, \ldots.
\]
The expected value of \( N_0 \) is \( E(N_0) = n_1 / p_c \) and its variance is \( V(N_0) = n_1(1-p_c)/p_c^2 \). Any upper tail probability would be defined as
\[
Pr\{N_0 > w\} = 1 - \sum_{i=n_1}^{w-1} \binom{i-1}{n_1-1} p_c (1-p_c)^{i-n_1}.
\]
Again, since \( p_c \) cannot be explicitly evaluated we will use a numerical integration technique to compute actual values of \( p_c \). As before, with integrals involving \( J(x_1, c) \) and \( H(x_1, c) \), \( p_c \) can be expressed in a form suitable for integration using Gaussian-Hermite numerical quadrature formulas.
We will use these results to examine the average least control pool size required to match \( n_1 \) case observations with a caliper width of \( c \). As is usual, we will assume, without loss of generality, that \( X \sim N(0,1) \). To simplify our discussion, we will define a new random variable \( R = N_0/n_1 \), the ratio of the least control pool size to the number of matches desired. Thus, \( E(R) = 1/n_1 E(N_0) = 1/p_c \) and \( V(R) = q_c/n_1p_c^2 \), where \( q_c = (1-p_c) \). Using this transformation allows us to ignore the actual value of the mean least control pool size needed and to consider only the ratio of the mean least number of controls needed to the case sample size (i.e., the mean least number of controls per case). Table 3.4.1 presents the mean value of \( R \) for a series of values of \( \mu_{x_0} \) and \( \sigma_{x_0}^2 \), where \( c = \sqrt{1+\sigma_{x_0}^2}/2 \). There are several points we wish to emphasize about this table. First, as expected, the mean value of \( R \) decreases as \( \mu_{x_0} \) increases. This decrease is fastest when \( \mu_{x_0} \) is small and the decrease slows as \( \mu_{x_0} \) increases. Increasing the size of \( \sigma_{x_0} \) from a value of 0.2 to a value of 0.6 lowers the mean value of \( R \) to about 1/3 of its value at 0.2, while another increase in \( \sigma_{x_0} \) from 0.6 to 1.0 only decreases the value of \( R \) to about 2/3 of its value at \( \sigma_{x_0} = 0.6 \). We would soon reach a point of diminishing returns, where an increase in \( \sigma_{x_0} \) would produce no worthwhile change in the mean value of \( R \). A second trend to emphasize is the increase in the mean value of \( R \) as \( \mu_{x_0} \) increases. Note that this increase is accelerating, i.e., the rate of increase in the mean value of \( R \) is larger for large \( \mu_{x_0} \) than it is for small \( \mu_{x_0} \). However, this acceleration becomes less dramatic as \( \sigma_{x_0}^2 \) increases. Considered together, these trends indicate that when an investigator has a limited control pool from which to randomly sample,
<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>$\sigma_{x_0}$</th>
<th>2</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8.89</td>
<td></td>
<td>4.49</td>
<td>3.04</td>
<td>2.33</td>
<td>1.92</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8.89</td>
<td></td>
<td>4.49</td>
<td>3.04</td>
<td>2.33</td>
<td>1.92</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8.89</td>
<td></td>
<td>4.49</td>
<td>3.04</td>
<td>2.33</td>
<td>1.92</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8.98</td>
<td>2</td>
<td>4.53</td>
<td>3.07</td>
<td>2.35</td>
<td>1.94</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8.95</td>
<td></td>
<td>4.52</td>
<td>3.06</td>
<td>2.35</td>
<td>1.93</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8.94</td>
<td></td>
<td>4.51</td>
<td>3.06</td>
<td>2.34</td>
<td>1.93</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9.25</td>
<td></td>
<td>4.67</td>
<td>3.16</td>
<td>2.42</td>
<td>1.99</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9.13</td>
<td></td>
<td>4.61</td>
<td>3.12</td>
<td>2.39</td>
<td>1.96</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.07</td>
<td></td>
<td>4.58</td>
<td>3.10</td>
<td>2.38</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9.72</td>
<td>3</td>
<td>4.90</td>
<td>3.31</td>
<td>2.53</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9.44</td>
<td></td>
<td>4.76</td>
<td>3.22</td>
<td>2.46</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.30</td>
<td></td>
<td>4.69</td>
<td>3.17</td>
<td>2.43</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>1</td>
<td>5.25</td>
<td>3.54</td>
<td>2.70</td>
<td>2.20</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9.89</td>
<td></td>
<td>4.98</td>
<td>3.36</td>
<td>2.57</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.63</td>
<td></td>
<td>4.85</td>
<td>3.28</td>
<td>2.51</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>1</td>
<td>5.73</td>
<td>3.85</td>
<td>2.92</td>
<td>2.37</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10.50</td>
<td></td>
<td>5.28</td>
<td>3.56</td>
<td>2.71</td>
<td>2.21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10.07</td>
<td></td>
<td>5.07</td>
<td>3.42</td>
<td>2.61</td>
<td>2.14</td>
<td></td>
</tr>
</tbody>
</table>
and the mean difference between the two populations is large, it would be to the investigator's advantage to use a caliper of medium width ($a = 0.6$) or large width ($a = 1.0$), rather than a narrow caliper ($a = 0.2$). If Table 3.2.2 (on the proportional reduction in mean difference) is examined, it can be seen that about 90% of the mean difference is removed when $a = 0.6$ and that about 75% is removed when $a = 1.0$. When $a = 0.6$, approximately 99% is removed, so that while there is some sacrifice in the reduction in mean difference when $a$ is increased, it is not significant in comparison to the reduction in the mean value of the average least control pool size needed to attain such a reduction in mean difference. We can quantify the effect of the caliper width on the proportional reduction in mean difference relative to $E(R)$ by considering the ratio of the proportional reduction to $E(R)$. This ratio is then the "average proportional reduction per average number of controls sampled" and so gives us a rough idea of the effectiveness of a caliper width on a "cost" basis. Table 3.4.2 gives the values of this ratio for a series of $a$'s when $\mu_{X_0} = 1$ and $\sigma^2_{X_0} = 1, 2, 3$. A low value in this table indicates that the associated caliper width has a low proportional reduction relative to the mean value of $R$ necessary to attain that proportional reduction, while a higher value gives a better proportional reduction relative to the associated value of $E(R)$. The trends suggested in this table also hold for all $\mu_{X_0} < 1$ and the maximum value of this ratio was generally occurred for $a$ between 1.2 and 1.3. For $\mu_{X_0} > 1$, the maximum value of the ratio increased slowly with $\mu_{X_0}$, so that very large values of $\mu_{X_0}$ required larger caliper values than 1.3 to obtain the maximum of this ratio. In this sense, then, wide calipers perform better than narrow calipers. This confirms our remarks
TABLE 3.4.2

Proportional Reduction in Mean Difference Due to Caliper Matching

with a Caliper of \( a\sqrt{(1+\sigma^2_{X_0})/2} \) Divided by the Mean Ratio of the

Least Control Pool Size to the Case Sample Size Necessary to

Match All Cases with the Same Caliper, \( X_1 \sim N(0,1) \), \( X_0 \sim N(1, \sigma^2_{X_0}) \)

<table>
<thead>
<tr>
<th>( a )</th>
<th>( \sigma^2_{X_0} = 1 )</th>
<th>( \sigma^2_{X_0} = 2 )</th>
<th>( \sigma^2_{X_0} = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.0438</td>
<td>0.0476</td>
<td>0.0496</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0866</td>
<td>0.0944</td>
<td>0.0985</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1275</td>
<td>0.1394</td>
<td>0.1456</td>
</tr>
<tr>
<td>0.4</td>
<td>0.1658</td>
<td>0.1820</td>
<td>0.1903</td>
</tr>
<tr>
<td>0.5</td>
<td>0.2008</td>
<td>0.2215</td>
<td>0.2320</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2321</td>
<td>0.2572</td>
<td>0.2699</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2595</td>
<td>0.2890</td>
<td>0.3038</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2827</td>
<td>0.3163</td>
<td>0.3331</td>
</tr>
<tr>
<td>0.9</td>
<td>0.3016</td>
<td>0.3390</td>
<td>0.3577</td>
</tr>
<tr>
<td>1.0</td>
<td>0.3164</td>
<td>0.3571</td>
<td>0.3775</td>
</tr>
<tr>
<td>1.1</td>
<td>0.3271</td>
<td>0.3706</td>
<td>0.3924</td>
</tr>
<tr>
<td>1.2</td>
<td>0.3339</td>
<td>0.3796</td>
<td>0.4026</td>
</tr>
<tr>
<td>1.3</td>
<td>0.3371</td>
<td>0.3843</td>
<td>0.4082</td>
</tr>
<tr>
<td>1.4</td>
<td>0.3369</td>
<td>0.3850</td>
<td>0.4094</td>
</tr>
<tr>
<td>1.5</td>
<td>0.3276</td>
<td>0.3821</td>
<td>0.4067</td>
</tr>
<tr>
<td>1.6</td>
<td>0.3191</td>
<td>0.3758</td>
<td>0.4004</td>
</tr>
</tbody>
</table>
about caliper choice in our discussion of the implications of Table
3.4.1, i.e., when the control pool size to be examined is limited,
either due to cost or time constraints, the use of a wide caliper is
generally recommended.

3.5. Covariance Analysis and Caliper Matching

We will now examine some properties of the mean difference esti-
mator of the true case-control difference using caliper pair-matched
data. We will assume, as we did in Section 2.5., that the response
variable is related to the matching variable by the usual straight
line Ancova model (i.e., \( y_{ij} = \alpha_i + \beta x_{ij} + e_{ij} \)). After finding the
mean, the variance and the bias of the caliper matched mean difference
estimator under the assumed model, we shall compare the variance and
mean square error of this estimator to those of the usual covariance
analysis estimator, which is the traditional estimator used under the
assumed model.

In Section 2.5., we were able to employ an "unpaired" mean-differ-
ence estimator of the case-control difference when the matching vari-
able had been categorically matched. We had a number of observations
in each category and these categories were non-overlapping, so that we
could analyze the unpaired data by an approach similar to that used in
a two-way analysis of variance with more than one observation per cell,
and so obtain an estimate of the variance of the mean difference. This
approach cannot be used when the data has been caliper matched. If we
were to associate a "category" with each observed \( x_i \), each such cate-
gory having a width of twice the caliper size (i.e., from \( x_i - c \) to \( x_i + c \)),
these "categories" would overlap for any reasonably large sample size
and caliper width. Furthermore, even if there were no overlap, there
would still be no way to estimate the "within category" variance for each group, since there would be only one observation per "category" per group. This, then, would preclude the use of a two-way analysis of variance approach to estimate the variance of the mean-difference estimator of the case-control difference. Thus, we can either estimate the within group variance for each group and then combine these two separate estimates to form a pooled estimate of the variance of the mean difference having 2(n-1) degrees of freedom, or we can find the paired differences and then estimate the variance of the mean difference using the paired differences (this estimator would have (n-1) degrees of freedom). However, these two different variance estimators do not have the same expected value, since, as we showed in Section 3.2., the control matching variable is correlated with the case matching variable after caliper matching. Now, by the very nature of the method of caliper matching, this correlation will always have a positive value. Thus, the paired difference estimator of the variance will give a smaller estimate of variance; and, we will consider only the paired estimator in what follows.

We will assume that the usual analysis of covariance model holds namely

\[ y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}, \quad i=0,1; \ j=1,\ldots,n, \]  

(3.5.1.)

where \( e_{ij} \sim N(0, \sigma^2_e) \). When \( n \) controls have been caliper pair-matched, using a caliper of size \( c \), let \( x_{0j} \) denote the control matching variable that has been paired with \( x_{1j} \), the \( j \)-th case matching variable observation. The response variable values associated with these two observations are then \( y_{0j} \) and \( y_{1j} \), respectively. The within-pair difference
for the response variable is then \( d_j = (y_{1j} - y_{0j}) \) and, from (3.5.1.), it follows that

\[
d_j = y_{1j} - y_{0j} = \alpha_1 - \alpha_0 + \beta(x_{1j} - x_{0j}) + (e_{1j} - e_{0j}) .
\]

So, the conditional mean and variance of \( d_j \) given \( x_{0j} \) and \( x_{1j} \) are

\[
E(d_j | x_{1j}, x_{0j}) = (\alpha_1 - \alpha_0) + \beta(x_{1j} - x_{0j}) ,
\]

and

\[
\text{Var}(d_j | x_{1j}, x_{0j}) = 2\sigma_e^2 .
\]

Thus, the unconditional mean and variance of \( d_j \) are

\[
E(d_j) = (\alpha_1 - \alpha_0) + \beta(\mu_{x_1} - \mu(c)) ,
\]

and

\[
\text{V}(d_j) = 2\sigma_e^2 + \beta^2(\sigma_{x_1}^2 + \sigma^2(c)) - 2 \text{Cov}(X_1, X_0 | c) .
\]

The caliper-matched estimator of the true case-control difference

\( (\alpha_1 - \alpha_0) \) is \( \bar{d} = 1/n \sum_{j=1}^{n} d_j \); so, from the moments we have just found, we have

\[
E(\bar{d}) = \alpha_1 - \alpha_0 + \beta(\mu_{x_1} - \mu(c)) , \quad (3.5.2.)
\]
\[
\text{Var}(\overline{d}) = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \left( \sigma^2 \text{x}_1^2 + \sigma^2(c) - 2 \text{Cov}(x_1, x_0 \mid c) \right), \quad (3.5.3)
\]

so that

\[
\text{BIAS}(\overline{d}) = E(\overline{d}) - (\alpha_1 - \alpha_0)
\]

\[= \beta(\mu_{x_1} - \mu(c)) \quad . \quad (3.5.4)
\]

Now, given the further assumption that \(X_1 \sim N(\mu_{x_1}, \sigma^2_{x_1})\), we can now make use of the expressions for \(\mu(c), \sigma^2(c), \) and \(\text{Cov}(X_1, X_0 \mid c)\) given by (3.3.3), (3.3.4), and (3.3.5), respectively. Given this assumption, then, (3.5.3) and (3.5.4) become

\[
\text{BIAS}(\overline{d}) = \beta(\mu_{x_1} - \mu(c))
\]

\[= \beta \left\{ \mu_{x_1} - \mu_{x_0} - \sigma \int_{-\infty}^{\infty} J(x_1, c) \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \, dx_1 \right\} ,
\]

and

\[
\text{Var}(\overline{d}) = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \left\{ \sigma^2 \text{x}_1 + \sigma^2 \text{x}_0 + \mu^2(c) + 2\mu_{x_0} \int_{-\infty}^{\infty} J(x_1, c) \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \, dx_1 \right\}
\]
\[ + \sigma^2 \int_{-\infty}^{\infty} H(x_1, c) Z(x_1 - \mu_{x_1}) \frac{dx_1}{\sigma_{x_1}} - 2\sigma_{x_1} \int_{-\infty}^{\infty} (x_1 - \mu_{x_1}) J(x_1, c) Z \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \frac{dx_1}{\sigma_{x_1}} \] 

In the last chapter, we showed that the analysis of randomly matched or unmatched samples could be treated as special limiting cases of the general category matching approach. We can also show that these analyses are limiting cases of the caliper matching approach. Earlier, in Section 3.3., we found that \( \lim_{c \to 0} \mu(c) = \mu_{x_1} \), \( \lim_{c \to 0} \sigma^2(c) = \sigma^2_{x_1} \), and

that \( \lim_{c \to 0} \text{Cov}(X_1, X_0 | c) = \sigma^2_{x_1} \). If we then take the limits of (3.5.3.) and (3.5.4.) as \( c \to 0 \) and make use of the results just cited, we obtain

\[ \lim_{c \to 0} \text{BIAS}(\bar{d}) = \beta_{x_1} - \lim_{c \to 0} \mu(c) = 0, \]

and

\[ \lim_{c \to 0} \text{Var}(\bar{d}) = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \{ \sigma^2_{x_1} + \lim_{c \to 0} \sigma^2(c) - 2 \lim_{c \to 0} \text{Cov}(x_1, x_0 | c) \} \]

\[ = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} (\sigma^2_{x_1} + \sigma^2_{x_1} - 2\sigma^2_{x_1}) \]

\[ = \frac{2\sigma^2}{n}. \]
These results are the same as the results derived in the usual manner for exactly pair-matched data (e.g., Cochran (1954)). This is to be expected since a caliper of zero requires that $x_{1j} = x_{0j}$ for every $j$ (i.e., exact pair-matching). As $c$ approaches infinity, the choice of a control becomes more and more a completely random process. And, in fact, we have shown that $\lim_{c \to \infty} \mu(c) = \mu$, $\lim\sigma^2(c) = \sigma^2$, and $\lim_{c \to \infty} \text{Cov}(X_{1}, X_{0}|c) = 0$. Applying these results to expressions (3.5.3.) and (3.5.4.), we have

$$\lim_{c \to \infty} \text{BIAS} (\bar{d}) = \beta(\mu_{-} \mu_{0})$$

and

$$\lim_{c \to \infty} \text{Var}(\bar{d}) = \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \left( \sigma^2_{x_1} + \lim_{c \to \infty} \sigma^2(c) - 2 \lim_{c \to \infty} \text{Cov}(X_{1}, X_{0}|c) \right)$$

$$= \frac{2\sigma^2}{n} + \frac{\beta^2}{n} \left( \sigma^2_{x_1} + \sigma^2_{x_0} \right).$$

These results are identical to those based on using random samples from the case and control groups.

We will now compare the caliper pair-matched estimator $\bar{d}$ with the covariance analysis estimator $\hat{\psi}$. We showed in Section 2.5., for $X_{1} \sim N(0,1)$ and for $X_{0} \sim N(\mu_{0}, \sigma^2_{x_0})$, that

$$\text{BIAS} (\hat{\psi}) = 0, \quad (3.5.5.)$$
and that

\[
\text{Var}(\hat{\psi}) = \frac{2\sigma^2_e}{n} \left\{ 1 + \frac{(1+\sigma^2_{x_0})^2 + (1+\sigma^2_{x_0})n\mu^2_{x_0}}{2(n-1)(1+\sigma^2_{x_0})^2 - 4(1+\sigma^4_{x_0})} \right\},
\]

where \( \hat{\psi} \) is estimated using random samples of size \( n \) from both groups. As before, we will consider the relative efficiency and the relative mean square error of the caliper matched mean-difference estimator to the covariance analysis estimator. Note that

\[
\text{Var}(\bar{d}) = \frac{2\sigma^2_e}{n} + \beta^2 \left\{ \sigma^2_{x_1} + \sigma^2(c) - 2 \text{Cov}(X_1, X_0 | c) \right\}
\]

\[
= \frac{2\sigma^2_e}{n} \left\{ 1 + \frac{\rho^2_{x_1y_1}}{2(1-\rho^2_{x_1y_1})} \left[ 1 + \sigma^2(c) - 2 \text{Cov}(X_1, X_0 | c) \right] \right\},
\]

(3.5.7.)

since \( X_1 \sim N(0,1), \sigma^2_e = \sigma^2_{y_1} (1-\rho^2_{x_1y_1}), \) and \( \beta^2 = \sigma^2_{y_1} \rho^2_{x_1y_1} \). Also, under the same conditions,

\[
\text{MSE}(\bar{d}) = \text{Var}(\bar{d}) + \{\text{BIAS}(\bar{d})\}^2
\]

\[
= \frac{2\sigma^2_e}{n} \left\{ 1 + \frac{\rho^2_{x_1y_1}}{2(1-\rho^2_{x_1y_1})} \left[ 1 + \sigma^2(c) - 2 \text{Cov}(X_1, X_0 | c) \right] \right\}
\]

\[+ \beta^2 \mu^2(c) \]
\[
\frac{2\sigma^2_e}{n} \left\{ 1 + \frac{\rho_{x_1 y_1}^2}{2(1-\rho_{x_1 y_1}^2)} \left[ 1 + \sigma^2(c) - 2 \text{Cov}(X_1, X_0|c) + n\mu^2(c) \right] \right\}.
\]

(3.5.8.)

Recall our discussion in Section 2.5. concerning the degrees of freedom available for estimation of the variance of the estimators. Fisher has shown that the efficiency of \( \bar{x} \) is \((\nu+1)/(\nu+3)s^2\) when the estimator of the variance of \( \bar{x} \) involves using an \( S^2 \) with \( \nu \) degrees of freedom, as opposed to an efficiency of \( 1/\sigma^2 \) when the variance of \( x \) is known to be \( \sigma^2 \). The multiplier \((\nu+1)/(\nu+3)\) acts to correct the efficiency for the loss of information incurred when the variance of \( \bar{x} \) must be estimated. We used this factor throughout Section 2.5. to correct for the different error degrees of freedom associated with \( \hat{\psi} \) and with the category matched mean difference estimator. The correction was small in that case, since the error degrees of freedom ((2n-3) and (2n-2r), respectively) were quite close for those two estimators. The error degrees of freedom are quite different for the caliper matched estimator and the covariance analysis estimator. As before, the estimator of \( \text{Var}(\hat{\psi}) \) has (2n-3) degrees of freedom; but, the estimator of \( \text{Var}(\tilde{d}) \) under caliper matching has only (n-1) degrees of freedom, since it is computed from the n paired differences. For this reason, we will continue to use this correction factor. The correction factor for the relative efficiency and the relative mean square error is \( n^2/(n+2)(n-1) \). Values of this correction factor are shown below in Table 3.5.1.
TABLE 3.5.1

Some Values of the Correction Factor \( \frac{n^2}{(n+2)(n-1)} \)

<table>
<thead>
<tr>
<th>n</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>80</th>
<th>( \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{n^2}{(n+2)(n-1)} )</td>
<td>.924</td>
<td>.957</td>
<td>.977</td>
<td>.988</td>
<td>1</td>
</tr>
</tbody>
</table>

The factor only has an effect when \( n \) is fairly small. When corrected by this factor the relative efficiency of \( \hat{d} \) to \( \hat{\psi} \) is

\[
\frac{n^2}{(n+2)(n-1)} \frac{v(\hat{\psi})}{v(\hat{d})} = \frac{n^2}{(n+2)(n-1)} \frac{(1+\sigma^2_{x_0})^2 + (1+\sigma^2_{x_0})\mu^2_{x_0} - 4(1+\sigma^2_{x_0})}{\{1 + \frac{\rho^2_{x_0,y_1}}{2(n-1)(1+\sigma^2_{x_0}) - 4(1+\sigma^2_{x_0})}\}}
\]

\[
\frac{1 + \frac{\rho^2_{x_0,y_1}}{2(1-\rho^2_{x_0,y_1})}}{(1+\sigma^2_{x_1,y_1} - 2 \text{ Cov}(X_1,X_0|c))}
\]

(3.5.9.)

By careful examination of the expression (3.5.9.), we can spot several general trends. First, \( \frac{\rho^2_{x_1,y_1}}{(1-\rho^2_{x_1,y_1})} \) increases as \( \rho^2_{x_1,y_1} \) increases; so, since \( \frac{\rho^2_{x_1,y_1}}{(1-\rho^2_{x_1,y_1})} \) appears in the denominator of (3.5.9.) but not in the numerator, (3.5.9.) will decrease of \( \rho^2_{x_1,y_1} \), increases. Thus with other parameters held constant, caliper matching will perform best in terms of variance considerations when \( \rho^2_{x_1,y_1} \) is small. Secondly, since \( \mu^2_{x_0} \) appears in the numerator but not in the denominator, (3.5.9.) will increase as \( \mu^2_{x_0} \) increases, which indicates that caliper pair-matching
should perform better when \( \mu_{x_0}^2 \) is large than when \( \mu_{x_0}^2 \) is small.

Finally, \((1+\sigma^2(c)-2 \text{Cov}(X_1,X_0|c))\) will increase as \( c = \alpha \sqrt{1+\sigma_{x_0}^2}/2 \)
increases in \( a \) for \( \sigma_{x_0}^2 > 1 \), so that (3.5.9) should decrease as \( a \) increases. When \( \sigma_{x_0}^2 \approx 1 \), the increase or decrease will depend on whether
or not \( 2 \text{Cov}(X_1,X_0|c) \) changes faster in absolute amount than \( \sigma^2(c) \).

We are referring to the fact that for \( \sigma_{x_0}^2 \approx 1 \), we observed that for certain values of \( a \), \( \sigma^2(c) \) increased as \( a \) increased. Tables 3.5.2, 3.5.3,
and 3.5.4 give some computed values of (3.5.9) for \( a = 0.2, 0.6, 1.0 \),
respectively. As is usual, we have assumed that \( X_1 \sim N(0,1) \), and
\( X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2) \), and that each member of the case sample of size \( n \) has
been caliper pair-matched to a control using a caliper of size \( c = \alpha \sqrt{1+\sigma_{x_0}^2}/2 \) for the caliper matched mean difference estimator and that
the covariance analysis estimator is based on random samples of size \( n \).

All three tables reflect the trends suspected above; i.e., the relative
efficiency decreases as \( \rho_{x_1,y_1}^2 \) and \( a \) increase and increases as \( \mu_{x_0}^2 \)
increases. Also, two other trends are evident: the relative efficiency decreases as \( n \) increases and as \( \sigma_{x_0}^2 \) increases.

It can be seen that for small \( a \) (\( a=0.2 \)) the caliper pair-matched estimator is almost always as efficient as the covariance analysis estimator and is usually more efficient. The same holds for medium \( a \) (\( a=0.6 \)) when \( \rho_{x_1,y_1} < 0.4 \) and for large \( a \) (\( a=1.0 \)) when \( \rho_{x_1,y_1} < 0.2 \). For \( a = 0.6 \) or \( 1.0 \), when \( \rho_{x_1,y_1} \) is larger than 0.4 and 0.2, respectively,
then covariance analysis is as efficient for small \( \mu_{x_0}^2 \) and \( \sigma_{x_0}^2 \) and
TABLE 3.5.2
Relative Efficiency of the Caliper Pair-Matched Estimator to the Covariance

Estimator for the Caliper $c = 0.2\sqrt{1 + \rho_{x_0}^2}/2$, $X_1 \sim N(0,1)$, $X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2)$

(Expression (3.5.9))

<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>$\sigma_{x_0}^2$</th>
<th>n= Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>n= Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9997</td>
<td>0.9894</td>
<td>0.9832</td>
<td>0.9835</td>
<td>0.9987</td>
<td>0.9884</td>
<td>0.9823</td>
<td>0.9825</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9994</td>
<td>0.9892</td>
<td>0.9833</td>
<td>0.9851</td>
<td>0.9974</td>
<td>0.9872</td>
<td>0.9814</td>
<td>0.9832</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0097</td>
<td>0.9997</td>
<td>0.9939</td>
<td>0.9951</td>
<td>1.0087</td>
<td>0.9987</td>
<td>0.9929</td>
<td>0.9941</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>1.0044</td>
<td>0.9944</td>
<td>0.9887</td>
<td>0.9911</td>
<td>1.0024</td>
<td>0.9924</td>
<td>0.9868</td>
<td>0.9891</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.0397</td>
<td>1.0305</td>
<td>1.0258</td>
<td>1.0298</td>
<td>1.0386</td>
<td>1.0295</td>
<td>1.0247</td>
<td>1.0288</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>1.0194</td>
<td>1.0089</td>
<td>1.0049</td>
<td>1.0090</td>
<td>1.0174</td>
<td>1.0079</td>
<td>1.0029</td>
<td>1.0070</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.0897</td>
<td>1.0819</td>
<td>1.0789</td>
<td>1.0877</td>
<td>1.0886</td>
<td>1.0808</td>
<td>1.0778</td>
<td>1.0866</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>1.0444</td>
<td>1.0357</td>
<td>1.0318</td>
<td>1.0389</td>
<td>1.0423</td>
<td>1.0337</td>
<td>1.0298</td>
<td>1.0368</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.1596</td>
<td>1.1538</td>
<td>1.1533</td>
<td>1.1687</td>
<td>1.1585</td>
<td>1.1527</td>
<td>1.1522</td>
<td>1.1675</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.0794</td>
<td>1.0719</td>
<td>1.0696</td>
<td>1.0806</td>
<td>1.0772</td>
<td>1.0698</td>
<td>1.0674</td>
<td>1.0785</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.2496</td>
<td>1.2464</td>
<td>1.2490</td>
<td>1.2728</td>
<td>1.2484</td>
<td>1.2451</td>
<td>1.2477</td>
<td>1.2715</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>1.1243</td>
<td>1.1185</td>
<td>1.1181</td>
<td>1.1344</td>
<td>1.1221</td>
<td>1.1163</td>
<td>1.1159</td>
<td>1.1321</td>
</tr>
</tbody>
</table>
TABLE 3.5.2 continued...

\[ \rho_{x_1, y_1} = 0.6 \]

\[ \rho_{x_1, y_1} = 0.8 \]

<table>
<thead>
<tr>
<th>( \mu_{x_0} )</th>
<th>( \sigma^2_{x_0} )</th>
<th>n= Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>40</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9962</td>
<td>0.9860</td>
<td>0.9798</td>
<td>0.9801</td>
<td>0.9882</td>
<td>0.9781</td>
<td>0.9720</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9925</td>
<td>0.9924</td>
<td>0.9766</td>
<td>0.9783</td>
<td>0.9768</td>
<td>0.9669</td>
<td>0.9612</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0062</td>
<td>0.9962</td>
<td>0.9904</td>
<td>0.9917</td>
<td>0.9981</td>
<td>0.9882</td>
<td>0.9825</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>0.9975</td>
<td>0.9875</td>
<td>0.9819</td>
<td>0.9843</td>
<td>0.9817</td>
<td>0.9719</td>
<td>0.9664</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.0361</td>
<td>1.0269</td>
<td>1.0222</td>
<td>1.0262</td>
<td>1.0278</td>
<td>1.0187</td>
<td>1.0140</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>1.0124</td>
<td>1.0029</td>
<td>0.9980</td>
<td>1.0021</td>
<td>0.9964</td>
<td>0.9871</td>
<td>0.9822</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.0859</td>
<td>1.0782</td>
<td>1.0752</td>
<td>1.0839</td>
<td>1.0772</td>
<td>1.0695</td>
<td>1.0666</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>1.0372</td>
<td>1.0286</td>
<td>1.0247</td>
<td>1.0317</td>
<td>1.0808</td>
<td>1.0124</td>
<td>1.0086</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.1557</td>
<td>1.1499</td>
<td>1.1494</td>
<td>1.1646</td>
<td>1.1465</td>
<td>1.1407</td>
<td>1.1402</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.0719</td>
<td>1.0646</td>
<td>1.0622</td>
<td>1.0732</td>
<td>1.0550</td>
<td>1.0478</td>
<td>1.0455</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.2453</td>
<td>1.2420</td>
<td>1.2446</td>
<td>1.2683</td>
<td>1.2352</td>
<td>1.2320</td>
<td>1.2346</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.1166</td>
<td>1.1108</td>
<td>1.1104</td>
<td>1.1266</td>
<td>1.0990</td>
<td>1.0933</td>
<td>1.0929</td>
</tr>
</tbody>
</table>
TABLE 3.5.3

Relative Efficiency of the Caliper Pair-Matched Estimator to the Covariance Estimator for the Caliper $c=0.6\sqrt{(1+\sigma_x^2)/2}$, $X_1 \sim N(0,1)$, $X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2)$

(Expresssion (3.5.9))

$$\rho_{x_1, y_1} = 0.2$$

<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>$\sigma_{x_0}^2$</th>
<th>$n= Looks at the Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9975</td>
<td>0.9872</td>
<td>0.9811</td>
<td>0.9814</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9951</td>
<td>0.9049</td>
<td>0.9791</td>
<td>0.9809</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0074</td>
<td>0.9075</td>
<td>0.9917</td>
<td>0.9929</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0001</td>
<td>0.9901</td>
<td>0.9845</td>
<td>0.9868</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.0374</td>
<td>1.0282</td>
<td>1.0235</td>
<td>1.0276</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>1.0150</td>
<td>1.0055</td>
<td>1.0006</td>
<td>1.0047</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.0873</td>
<td>1.0796</td>
<td>1.0766</td>
<td>1.0853</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>1.0399</td>
<td>1.0313</td>
<td>1.0274</td>
<td>1.0344</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.1572</td>
<td>1.1514</td>
<td>1.1509</td>
<td>1.1662</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.0747</td>
<td>1.0674</td>
<td>1.0650</td>
<td>1.0760</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.2470</td>
<td>1.2438</td>
<td>1.2464</td>
<td>1.2702</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>1.1196</td>
<td>1.1137</td>
<td>1.1133</td>
<td>1.1296</td>
</tr>
</tbody>
</table>

$$\rho_{x_1, y_1} = 0.4$$

<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>$\sigma_{x_0}^2$</th>
<th>$n= Looks at the Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9975</td>
<td>0.9872</td>
<td>0.9811</td>
<td>0.9814</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9951</td>
<td>0.9049</td>
<td>0.9791</td>
<td>0.9809</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0074</td>
<td>0.9075</td>
<td>0.9917</td>
<td>0.9929</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0001</td>
<td>0.9901</td>
<td>0.9845</td>
<td>0.9868</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.0374</td>
<td>1.0282</td>
<td>1.0235</td>
<td>1.0276</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>1.0150</td>
<td>1.0055</td>
<td>1.0006</td>
<td>1.0047</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.0873</td>
<td>1.0796</td>
<td>1.0766</td>
<td>1.0853</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>1.0399</td>
<td>1.0313</td>
<td>1.0274</td>
<td>1.0344</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.1572</td>
<td>1.1514</td>
<td>1.1509</td>
<td>1.1662</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.0747</td>
<td>1.0674</td>
<td>1.0650</td>
<td>1.0760</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.2470</td>
<td>1.2438</td>
<td>1.2464</td>
<td>1.2702</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>1.1196</td>
<td>1.1137</td>
<td>1.1133</td>
<td>1.1296</td>
</tr>
<tr>
<td>$\mu_{x_0}$</td>
<td>$\sigma^2_{x_0}$</td>
<td>$n$= Large Sample</td>
<td>40</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>-----------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9675</td>
<td>0.9575</td>
<td>0.9716</td>
<td>0.9718</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9380</td>
<td>0.9284</td>
<td>0.9229</td>
<td>0.9246</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.9772</td>
<td>0.9675</td>
<td>0.9619</td>
<td>0.9631</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>0.9427</td>
<td>0.9333</td>
<td>0.9280</td>
<td>0.9302</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.0065</td>
<td>0.9976</td>
<td>0.9930</td>
<td>0.9969</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>0.9569</td>
<td>0.9480</td>
<td>0.9433</td>
<td>0.9471</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.0552</td>
<td>1.0477</td>
<td>1.0448</td>
<td>1.0533</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>0.9805</td>
<td>0.9724</td>
<td>0.9688</td>
<td>0.9753</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.1236</td>
<td>1.1180</td>
<td>1.1175</td>
<td>1.1323</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.0136</td>
<td>1.0067</td>
<td>1.0044</td>
<td>1.0148</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.2115</td>
<td>1.2084</td>
<td>1.2109</td>
<td>1.2340</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>1.0562</td>
<td>1.0507</td>
<td>1.0503</td>
<td>1.0656</td>
</tr>
</tbody>
</table>
TABLE 3.5.4

Relative Efficiency of the Caliper Pair-Matched Estimator to the Covariance Estimator for the Caliper \( c = 1.0/\sqrt{1+\sigma^2_{x_0}/2} \), \( X_1 \sim N(0,1) \), \( X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0}) \).

(Expression 3.5.9))

<table>
<thead>
<tr>
<th>( \mu_{x_0} )</th>
<th>( \sigma^2_{x_0} )</th>
<th>( n= ) Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>( n= ) Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9932</td>
<td>0.9830</td>
<td>0.9768</td>
<td>0.9772</td>
<td>0.9700</td>
<td>0.9600</td>
<td>0.9541</td>
<td>0.9543</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9871</td>
<td>0.9770</td>
<td>0.9712</td>
<td>0.9730</td>
<td>0.9436</td>
<td>0.9340</td>
<td>0.9285</td>
<td>0.9301</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0032</td>
<td>0.9933</td>
<td>0.9875</td>
<td>0.9887</td>
<td>0.9798</td>
<td>0.9701</td>
<td>0.9645</td>
<td>0.9657</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>0.9920</td>
<td>0.9821</td>
<td>0.9765</td>
<td>0.9789</td>
<td>0.9484</td>
<td>0.9389</td>
<td>0.9336</td>
<td>0.9358</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.0331</td>
<td>1.0240</td>
<td>1.0193</td>
<td>1.0233</td>
<td>1.0094</td>
<td>1.0005</td>
<td>0.9959</td>
<td>0.9998</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>1.0069</td>
<td>0.9875</td>
<td>0.9926</td>
<td>0.9966</td>
<td>0.9628</td>
<td>0.9538</td>
<td>0.9491</td>
<td>0.9530</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.0829</td>
<td>1.0752</td>
<td>1.0723</td>
<td>1.0810</td>
<td>1.0586</td>
<td>1.0511</td>
<td>1.0482</td>
<td>1.0567</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>1.0316</td>
<td>1.0231</td>
<td>1.0193</td>
<td>1.0262</td>
<td>0.9868</td>
<td>0.9786</td>
<td>0.9749</td>
<td>0.9816</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.1527</td>
<td>1.1470</td>
<td>1.1465</td>
<td>1.1617</td>
<td>1.1276</td>
<td>1.1220</td>
<td>1.1215</td>
<td>1.1364</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.0663</td>
<td>1.0590</td>
<td>1.0567</td>
<td>1.0676</td>
<td>1.0204</td>
<td>1.0134</td>
<td>1.0111</td>
<td>1.0216</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.2425</td>
<td>1.2392</td>
<td>1.2419</td>
<td>1.2655</td>
<td>1.2165</td>
<td>1.2133</td>
<td>1.2159</td>
<td>1.2391</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>1.1109</td>
<td>1.1052</td>
<td>1.1047</td>
<td>1.1209</td>
<td>1.0636</td>
<td>1.0581</td>
<td>1.0577</td>
<td>1.0731</td>
</tr>
<tr>
<td>$n=\text{Large Sample}$</td>
<td>$n=10$</td>
<td>$n=20$</td>
<td>$n=40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_{X_1,Y_1}=0.6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_{X_1,Y_1}=0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c^2_{X_0}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>1.0023</td>
<td>1.0023</td>
<td>1.0023</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.9552</td>
<td>0.9552</td>
<td>0.9552</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.9187</td>
<td>0.9187</td>
<td>0.9187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.8826</td>
<td>0.8826</td>
<td>0.8826</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.8479</td>
<td>0.8479</td>
<td>0.8479</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.8142</td>
<td>0.8142</td>
<td>0.8142</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{X_0}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>1.0023</td>
<td>1.0023</td>
<td>1.0023</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.9552</td>
<td>0.9552</td>
<td>0.9552</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.9187</td>
<td>0.9187</td>
<td>0.9187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.8826</td>
<td>0.8826</td>
<td>0.8826</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.8479</td>
<td>0.8479</td>
<td>0.8479</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.8142</td>
<td>0.8142</td>
<td>0.8142</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5.4 continued...
becomes more efficient as $\sigma^2_{x_0}$ increases. Note that when $\rho_{x_1,y_1} < 0.2$ the three values of a give about the same results, but as $\rho_{x_1,y_1}$ increases, the smaller values of $a$ must be used in order for caliper matching to be preferred to covariance analysis on efficiency grounds. This suggests that in practice a wide caliper is generally suitable when the correlation between the matching variable and the response is weak, as it often is in epidemiological applications. When this correlation is strong, however, we would recommend that a narrow caliper be used or, if there are limits on the number of controls available, that analysis of covariance be used.

These recommendations are based only on variance considerations, and so do not take into account the possible bias in the estimators. Under our assumed model, $\text{BIAS} (\hat{\psi}) = 0$ and $\text{BIAS} (\overline{d}) = \beta (\mu_{x_1} - \mu(c))$. If we focus on the mean square error criterion considered earlier in Chapter 2, we can take into account both the variance and bias of the estimators. Recall that the mean square error of an estimator $\hat{\theta}$ is the sum of its variance and squared bias, namely, $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{BIAS}^2(\hat{\theta})$. Using this definition, we have

$$\text{MSE}(\hat{\psi}) = \text{Var}(\hat{\psi}) + \text{BIAS}^2(\hat{\psi})$$

$$= \text{Var}(\hat{\psi}) ;$$

and from (3.5.8.), we have

$$\text{MSE}(\overline{d}) = \frac{2\sigma^2}{n} \left\{ 1 + \frac{\rho_{x_1,y_1}^2}{2(1-\rho_{x_1,y_1}^2)} \left[ 1 + \sigma^2(c) - 2 \text{Cov}(X_1X_0|c) + nu^2(c) \right] \right\} .$$
Thus, the relative mean square error of the covariance estimator to the caliper pair-matched estimator, after the adjustment for differing error degrees of freedom, is

$$\frac{n^2}{(n+2)(n-1)} \left[ 1 + \frac{(1+\sigma_x^2)^2 + (1+\sigma_x^2)n\mu_x^2}{2(n-1)(1+\sigma_x^2)^2 - 4(1+\sigma_x^4)} \right].$$

$$1 + \frac{\rho_{x_1y_1}^2}{2(1-\rho_{x_1y_1}^2)} \left[ 1 + \sigma_x^2(c) - 2 \text{Cov}(x_1, x_0 | c) + n\mu_x^2(c) \right]$$

(3.5.10.)

When we examined (3.5.9.), we located several general features of interest. In particular, (3.5.9.) increased as $\mu_x^2$, increased and decreased as each of $n$, $\sigma_x^2$, $a$, and $\rho_{x_1y_1}^2$ increased. Because of the similarity in structure of expressions (3.5.9.) and (3.5.10.), we would expect the same trends to be present for (3.5.10.). Note, however, that the decrease in (3.5.10.) as $n$ increases should be much faster than the corresponding decrease in (3.5.9.) when $\mu(c)$ is not zero, since the denominator of (3.5.10.) involves the factor $n\mu_x^2(c)$. Also, since $\mu_x^2(c)$ is now in the denominator, the relative mean square error will not increase as quickly with increasing $\mu_x$, as does the relative efficiency. Tables 3.5.5, 3.5.6, and 3.5.7, for $a = 0.2, 0.6,$ and $1.0$, respectively, present some computed values of (3.5.10.). As before, we have caliper pair-matched $n$ observations from the $N(\mu_x, \sigma_x^2)$ control
TABLE 3.5.5

Relative Mean Square Error of the Covariance Analysis Estimator to the Caliper

Pair-Matched Estimator, $c = 0.2 \sqrt{1 + \sigma^2_{x_0}/2}$, $X_1 \sim N(0,1)$, $X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$

(Expression (3.5.10))

<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>$\sigma^2_{x_0}$</th>
<th>n= Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>40</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9997</td>
<td>0.9894</td>
<td>0.9832</td>
<td>0.9835</td>
<td>0.9987</td>
<td>0.9884</td>
<td>0.9823</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9994</td>
<td>0.9892</td>
<td>0.9833</td>
<td>0.9851</td>
<td>0.9975</td>
<td>0.9872</td>
<td>0.9814</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.0082</td>
<td>0.9997</td>
<td>0.9939</td>
<td>0.9951</td>
<td>1.0021</td>
<td>0.9986</td>
<td>0.9929</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>1.0037</td>
<td>0.9944</td>
<td>0.9887</td>
<td>0.9911</td>
<td>0.9994</td>
<td>0.9924</td>
<td>0.9868</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.0337</td>
<td>1.0305</td>
<td>1.0258</td>
<td>1.0298</td>
<td>1.0120</td>
<td>1.0294</td>
<td>1.0247</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>1.0167</td>
<td>1.0099</td>
<td>1.0049</td>
<td>1.0080</td>
<td>1.0054</td>
<td>1.0078</td>
<td>1.0029</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.0757</td>
<td>1.0818</td>
<td>1.0789</td>
<td>1.0867</td>
<td>1.0277</td>
<td>1.0806</td>
<td>1.0777</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>1.0383</td>
<td>1.0357</td>
<td>1.0318</td>
<td>1.0388</td>
<td>1.0152</td>
<td>1.0336</td>
<td>1.0298</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.1336</td>
<td>1.1537</td>
<td>1.1533</td>
<td>1.1686</td>
<td>1.0484</td>
<td>1.1522</td>
<td>1.1519</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>1.0682</td>
<td>1.0719</td>
<td>1.0695</td>
<td>1.0806</td>
<td>1.0284</td>
<td>1.0696</td>
<td>1.0673</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>1.2061</td>
<td>1.2462</td>
<td>1.2489</td>
<td>1.2727</td>
<td>1.0717</td>
<td>1.2443</td>
<td>1.2473</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>1.1064</td>
<td>1.1184</td>
<td>1.1180</td>
<td>1.1344</td>
<td>1.0446</td>
<td>1.1160</td>
<td>1.1157</td>
</tr>
<tr>
<td>$\mu_{x_0}$</td>
<td>$\sigma_{x_0}^2$</td>
<td>n= Large Sample</td>
<td>40</td>
<td>20</td>
<td>10</td>
<td>n= Large Sample</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>--------</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9962</td>
<td>0.9860</td>
<td>0.9798</td>
<td>0.9801</td>
<td>0.9882</td>
<td>0.9781</td>
<td>0.9720</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9925</td>
<td>0.9924</td>
<td>0.9766</td>
<td>0.9783</td>
<td>0.9768</td>
<td>0.9669</td>
<td>0.9612</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.9870</td>
<td>0.9961</td>
<td>0.9904</td>
<td>0.9916</td>
<td>0.9409</td>
<td>0.9880</td>
<td>0.9824</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>0.9889</td>
<td>0.9875</td>
<td>0.9819</td>
<td>0.9843</td>
<td>0.9558</td>
<td>0.9718</td>
<td>0.9664</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0.9614</td>
<td>1.0266</td>
<td>1.0221</td>
<td>1.0262</td>
<td>0.8266</td>
<td>1.0177</td>
<td>1.0135</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>0.9782</td>
<td>1.0028</td>
<td>0.9979</td>
<td>1.0020</td>
<td>0.8988</td>
<td>0.9867</td>
<td>0.9820</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>0.9245</td>
<td>1.0774</td>
<td>1.0748</td>
<td>1.0837</td>
<td>0.6962</td>
<td>1.0672</td>
<td>1.0654</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>0.9617</td>
<td>1.0283</td>
<td>1.0246</td>
<td>1.0316</td>
<td>0.8204</td>
<td>1.0114</td>
<td>1.0081</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.8825</td>
<td>1.1485</td>
<td>1.1486</td>
<td>1.1643</td>
<td>0.5818</td>
<td>1.1363</td>
<td>1.1380</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>0.9406</td>
<td>1.0640</td>
<td>1.0619</td>
<td>1.0731</td>
<td>0.7356</td>
<td>1.0460</td>
<td>1.0445</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>0.8382</td>
<td>1.2396</td>
<td>1.2434</td>
<td>1.2677</td>
<td>0.4897</td>
<td>1.2245</td>
<td>1.2308</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>0.9168</td>
<td>1.1098</td>
<td>1.1099</td>
<td>1.1263</td>
<td>0.6549</td>
<td>1.0903</td>
<td>1.0914</td>
</tr>
</tbody>
</table>
TABLE 3.5.6

Relative Mean Square Error of the Covariance Analysis Estimator to the Caliper

Pair-Matched Estimator, $c = 0.6\sqrt{1+\frac{\sigma^2_{x_0}}{\mu_{x_0}}}/2$, $X_1 \sim N(0,1)$, $X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$

(Expression (3.5.10))

<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>$\sigma^2_{x_0}$</th>
<th>$n$= Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>$\rho_{x_1,y_1}$=0.2</th>
<th></th>
<th>$\rho_{x_1,y_1}$=0.4</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9975</td>
<td>0.9872</td>
<td>0.9811</td>
<td>0.9814</td>
<td>0.9887</td>
<td>0.9785</td>
<td>0.9725</td>
<td>0.9727</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.9951</td>
<td>0.9849</td>
<td>0.9791</td>
<td>0.9809</td>
<td>0.9781</td>
<td>0.9681</td>
<td>0.9624</td>
<td>0.9641</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.9188</td>
<td>0.9971</td>
<td>0.9915</td>
<td>0.9928</td>
<td>0.6949</td>
<td>0.9870</td>
<td>0.9821</td>
<td>0.9838</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>0.9540</td>
<td>0.9899</td>
<td>0.9844</td>
<td>0.9868</td>
<td>0.8076</td>
<td>0.9723</td>
<td>0.9672</td>
<td>0.9697</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0.7495</td>
<td>1.0267</td>
<td>1.0227</td>
<td>1.0272</td>
<td>0.3752</td>
<td>1.0122</td>
<td>1.0111</td>
<td>1.0168</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>0.8508</td>
<td>1.0048</td>
<td>1.0002</td>
<td>1.0045</td>
<td>0.5343</td>
<td>0.9850</td>
<td>0.9818</td>
<td>0.9867</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>0.5851</td>
<td>1.0759</td>
<td>1.0747</td>
<td>1.0844</td>
<td>0.2204</td>
<td>1.0539</td>
<td>1.0591</td>
<td>1.0718</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>0.6077</td>
<td>1.0641</td>
<td>1.0634</td>
<td>1.0752</td>
<td>0.2372</td>
<td>1.0350</td>
<td>1.0398</td>
<td>1.0542</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.4608</td>
<td>1.1445</td>
<td>1.1474</td>
<td>1.1644</td>
<td>1.1462</td>
<td>1.1112</td>
<td>1.1257</td>
<td>1.1484</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>0.6066</td>
<td>1.0641</td>
<td>1.0634</td>
<td>1.0752</td>
<td>0.2372</td>
<td>1.0350</td>
<td>1.0398</td>
<td>1.0542</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>0.3742</td>
<td>1.2323</td>
<td>1.2406</td>
<td>1.2672</td>
<td>0.1069</td>
<td>1.1834</td>
<td>1.2105</td>
<td>1.2464</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>0.5096</td>
<td>1.1084</td>
<td>1.1106</td>
<td>1.1282</td>
<td>0.1725</td>
<td>1.0719</td>
<td>1.0829</td>
<td>1.1046</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE 3.5.6 continued...

<table>
<thead>
<tr>
<th>$\mu_{X_0}$</th>
<th>$\sigma_{X_0}^2$</th>
<th>$n$</th>
<th>Large Sample</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>$\rho_{X_1,Y_1}=0.6$</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>$\rho_{X_1,Y_1}=0.8$</th>
<th>40</th>
<th>20</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td></td>
<td>0.9675</td>
<td>0.9575</td>
<td>0.9516</td>
<td>0.9518</td>
<td>0.9040</td>
<td>0.8947</td>
<td>0.8891</td>
<td>0.8894</td>
<td>0.8188</td>
<td>0.8140</td>
<td>0.8154</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td></td>
<td>0.9380</td>
<td>0.9284</td>
<td>0.9229</td>
<td>0.9246</td>
<td>0.8272</td>
<td>0.8188</td>
<td>0.8140</td>
<td>0.8154</td>
<td>0.8176</td>
<td>0.8157</td>
<td>0.8191</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td></td>
<td>0.4318</td>
<td>0.9626</td>
<td>0.9595</td>
<td>0.9619</td>
<td>0.1930</td>
<td>0.8909</td>
<td>0.8923</td>
<td>0.8967</td>
<td>0.8803</td>
<td>0.9016</td>
<td>0.9184</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td></td>
<td>0.5838</td>
<td>0.9310</td>
<td>0.9269</td>
<td>0.9287</td>
<td>0.3064</td>
<td>0.8176</td>
<td>0.8157</td>
<td>0.8191</td>
<td>0.8140</td>
<td>0.8209</td>
<td>0.8299</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td></td>
<td>0.1669</td>
<td>0.9779</td>
<td>0.9831</td>
<td>0.9919</td>
<td>0.0593</td>
<td>0.8803</td>
<td>0.9016</td>
<td>0.9184</td>
<td>0.8814</td>
<td>0.9294</td>
<td>0.9538</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td></td>
<td>0.2768</td>
<td>0.9387</td>
<td>0.9387</td>
<td>0.9448</td>
<td>0.1075</td>
<td>0.8140</td>
<td>0.8209</td>
<td>0.8299</td>
<td>0.8140</td>
<td>0.8209</td>
<td>0.8299</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td></td>
<td>0.0862</td>
<td>1.0026</td>
<td>1.0218</td>
<td>1.0146</td>
<td>0.0288</td>
<td>0.8652</td>
<td>0.9166</td>
<td>0.9538</td>
<td>0.8652</td>
<td>0.9166</td>
<td>0.9538</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td></td>
<td>0.1504</td>
<td>0.9514</td>
<td>0.9582</td>
<td>0.9700</td>
<td>0.0528</td>
<td>0.8083</td>
<td>0.8294</td>
<td>0.8476</td>
<td>0.8083</td>
<td>0.8294</td>
<td>0.8476</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td></td>
<td>0.0540</td>
<td>1.0359</td>
<td>1.0749</td>
<td>1.1103</td>
<td>0.0187</td>
<td>0.8481</td>
<td>0.9367</td>
<td>1.0017</td>
<td>0.8481</td>
<td>0.9367</td>
<td>1.0017</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td></td>
<td>0.0940</td>
<td>0.9687</td>
<td>0.9851</td>
<td>1.0050</td>
<td>0.0316</td>
<td>0.8012</td>
<td>0.8408</td>
<td>0.8712</td>
<td>0.8012</td>
<td>0.8408</td>
<td>0.8712</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td></td>
<td>0.0384</td>
<td>1.0766</td>
<td>1.1411</td>
<td>1.1974</td>
<td>0.0124</td>
<td>0.8316</td>
<td>0.9612</td>
<td>1.0609</td>
<td>0.8316</td>
<td>0.9612</td>
<td>1.0609</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td></td>
<td>0.0650</td>
<td>0.9903</td>
<td>1.0192</td>
<td>1.0496</td>
<td>0.0214</td>
<td>0.7931</td>
<td>0.8550</td>
<td>0.9029</td>
<td>0.7931</td>
<td>0.8550</td>
<td>0.9029</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 3.5.7
Relative Mean Square Error of the Covariance Analysis Estimator to the Caliper

Pair-Matched Estimator, $c = 1.0 / \sqrt{1 + \sigma_x^2 / \mu_x}$, $X_1 \sim N(0,1)$, $X_0 \sim N(\mu_x, \sigma_x^2)$

(Expression (3.5.10))

<p>| $\mu_x$ | $\sigma^2_x$ | $\rho_{x_1,y_1} = 0.2$ |  |  |  | $\rho_{x_1,y_1} = 0.4$ |  |  |  |
|---------|--------------|------------------------|-------------------------|-------------------------|-------------------------|------------------------|-------------------------|-------------------------|
|         | n= Large Sample | 40  | 20  | 10 | n= Large Sample | 40  | 20  | 10 |
| 0.0     | 1             | 0.9932 | 0.9830 | 0.9769 | 0.9772 | 0.9700 | 0.9600 | 0.9541 | 0.9543 |
| 0.0     | 3             | 0.9871 | 0.9770 | 0.9712 | 0.9730 | 0.9436 | 0.9340 | 0.9285 | 0.9301 |
| 0.2     | 1             | 0.6484 | 0.9911 | 0.9864 | 0.9882 | 0.2845 | 0.2845 | 0.9607 | 0.9598 | 0.9633 |
| 0.2     | 3             | 0.7537 | 0.9809 | 0.9759 | 0.9786 | 0.3982 | 0.3982 | 0.9338 | 0.9311 | 0.9346 |
| 0.4     | 1             | 0.3257 | 1.0152 | 1.0149 | 1.0211 | 0.0943 | 0.0943 | 0.9631 | 0.9769 | 0.9902 |
| 0.4     | 3             | 0.4453 | 0.9925 | 0.9901 | 0.9954 | 0.1478 | 0.1478 | 0.9332 | 0.9388 | 0.9478 |
| 0.6     | 1             | 0.1860 | 1.0549 | 1.0620 | 1.0758 | 0.0469 | 0.0469 | 0.9676 | 1.0049 | 1.0344 |
| 0.6     | 3             | 0.2697 | 1.0117 | 1.1035 | 1.0233 | 0.0739 | 0.0739 | 0.9325 | 0.9514 | 0.9196 |
| 0.8     | 1             | 0.1224 | 1.1096 | 1.1275 | 1.1520 | 0.0292 | 0.0292 | 0.9751 | 1.0429 | 1.0952 |
| 0.8     | 3             | 0.1779 | 1.0383 | 1.0462 | 1.0623 | 0.0447 | 0.0447 | 0.9319 | 0.9688 | 0.9997 |
| 1.0     | 1             | 0.0887 | 1.1786 | 1.2107 | 1.2495 | 0.0208 | 0.0208 | 0.9863 | 1.0904 | 1.1716 |
| 1.0     | 3             | 0.1271 | 1.0720 | 1.0879 | 1.1123 | 0.0305 | 0.0305 | 0.9318 | 0.9905 | 1.0379 |</p>
<table>
<thead>
<tr>
<th>$\mu_{x_0}$</th>
<th>$\sigma^2_{x_0}$</th>
<th>n= Large Sample</th>
<th>$\rho_{x_1,y_1}=0.6$</th>
<th>n= Large Sample</th>
<th>$\rho_{x_1,y_1}=0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1</td>
<td>0.9164</td>
<td>0.9069</td>
<td>0.9013</td>
<td>0.9016</td>
</tr>
<tr>
<td>0.0</td>
<td>3</td>
<td>0.8501</td>
<td>0.8414</td>
<td>0.8364</td>
<td>0.8379</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.1184</td>
<td>0.8924</td>
<td>0.8992</td>
<td>0.9064</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>0.1827</td>
<td>0.8337</td>
<td>0.8350</td>
<td>0.8401</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0.0340</td>
<td>0.8536</td>
<td>0.8934</td>
<td>0.9205</td>
</tr>
<tr>
<td>0.4</td>
<td>3</td>
<td>0.0554</td>
<td>0.8120</td>
<td>0.8311</td>
<td>0.8465</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>0.0164</td>
<td>0.8019</td>
<td>0.8857</td>
<td>0.9436</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>0.0262</td>
<td>0.7800</td>
<td>0.8250</td>
<td>0.8571</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.0100</td>
<td>0.7484</td>
<td>0.8783</td>
<td>0.9749</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>0.0156</td>
<td>0.7420</td>
<td>0.8175</td>
<td>0.8715</td>
</tr>
<tr>
<td>1.0</td>
<td>1</td>
<td>0.0071</td>
<td>0.7005</td>
<td>0.8733</td>
<td>1.0137</td>
</tr>
<tr>
<td>1.0</td>
<td>3</td>
<td>0.0105</td>
<td>0.7022</td>
<td>0.8094</td>
<td>0.8894</td>
</tr>
</tbody>
</table>
pool to \( n \) randomly selected case observations from an \( N(0,1) \) distribution, using the caliper \( c = \sqrt{n(1+\sigma^2)} \). The trends we discussed above are present in the tables. By comparison, it can be seen that the values in these tables are always lower than the corresponding values in Tables 3.5.3 and 3.5.5, when \( \mu \) is greater than zero. Note that for small \( a \) (\( a=0.2 \)) and small samples, \( \bar{d} \) has a mean square error that is almost always no larger than the covariance estimator's mean square error. If we examine corresponding entries in the three tables, it becomes apparent that for small values of \( |\rho_{x_1y_1}| \), the three values of \( a \) are comparable, and thus, in practice, we would recommend that a large value of \( a \) be used when it is known that the relationship between the matching variable and the response variable is weak. When the relationship is strong, the differences in mean square error values among the three values of \( a \) become large and a small value of \( a \) or covariance analysis should be used.
CHAPTER IV
NEAREST-NEIGHBOR PAIR-MATCHING AND A COMPARISON OF THE THREE MATCHING PROCEDURES

4.1. Introduction

In this chapter, we shall consider the results of a Monte Carlo study of nearest-neighbor pair-matching; and, we shall also present some mean square error comparisons of the three matching methods we have considered in this dissertation. This will complete our study of univariate matching methods for continuous variables, and should provide a basis for recommendations on the use of these matching methods. In the rest of this section, we will provide an explanation and discussion of the technique of nearest-neighbor matching.

The technique of nearest-neighbor matching is almost self-explanatory. The cases are ordered in some fashion, and, then, for each case, all available controls are examined to find that control which has a matching variable value that is closest (i.e., is the nearest-neighbor) to the matching variable value of the case under consideration. The case and the nearest available control are then considered to be matched and both are removed from further consideration in the matching process.

From this description of the matching procedure, we can see that we are faced with a slightly different problem in examining the properties of this matching method than we are concerned with in studying the properties of categorical matching and caliper matching. For these latter two methods, the choices of category structure and caliper size, respectively, determine the values of the post-matching mean, variance and
correlation, given the initial distributions, and also determines the expected minimum number of randomly sampled controls needed to be able to match all the cases. For nearest-neighbor matching, the available sample sizes for the two groups determine the values of the post-sampling parameters. This difference is crucial since it eliminates one problem that can plague the other two matching methods, i.e., the problem of incomplete matching of the cases. This can be particularly important when there are a small number of cases, as is often the situation when studying rare diseases. Using nearest-neighbor matching, we are always able to match all the cases when there are at least as many controls available as there are cases.

It is clear that when the case and control samples are considered a priori to the actual matching process, that a single control may be the nearest-neighbor to several case observations. When this occurs, that control observation is matched with the first of those several cases to be considered. Thus, the type of ordering of the cases that is used can affect the matches that are made. There are several possible orderings, only one of which will be considered in detail in this chapter. One of these orderings is "low-to-high", where the "smallest" available case (in terms of matching variable value) is matched at each step. Another is "high-to-low", where the "largest" available case is chosen to be matched at each step. A final ordering, which will be used throughout this chapter, is a random ordering, in which the next case to be matched is chosen randomly from the remaining candidate cases.

Each of these orderings has some advantages in certain situations. For example, suppose the matching variable is normally distributed in both populations, such that the mean of the controls is greater than
that of the cases. The "low-to-high" ordering considers the cases with
the lowest values first. The cases with the lowest values would be
those for which it is most difficult to obtain tight or close matches,
and so this ordering would then find the best available matches for
these "difficult to match" cases. On the other hand, the "high-to-low"
ordering will obtain very tight matches for the high values, but the
low-valued cases will not be matched nearly as well. The random order-
ing is simplest to use and would be expected to produce "average"
matches for the whole sample.

Rubin (1973) investigated these three orderings for normally dis-
tributed matching variables. Under the assumption that the mean of the
cases is less than that of the controls, his results indicate that the
"high-to-low" ordering will give the greatest reduction in mean differ-
ence, followed by the random ordering and then the "low-to-high" order-
ing. However, this ranking of the three ordering methods was reversed
when he studied the quality of the individual matches which was measured
by the average of the squared within-pair differences of the matching
variable.

The complex nature of nearest-neighbor matching is such that we
have been unable to develop analytical expressions for the post-matching
moments of the distribution of the matched controls. In order to study
this method further, we will thus need to use Monte Carlo methods. As
mentioned above, Rubin (1973a) has previously worked in this area and
some of the following material represents duplication of his work. In
the next two sections, we shall examine the post-matching distribution
of the controls and compare the nearest-neighbor pair-matched mean
difference estimator of the true case-control difference to the analysis
of covariance estimator.
4.2. Nearest-Neighbor Pair-Matching for Normally Distributed Variables

In this section, we shall examine the results of a study on pair-matching of normally distributed variables using a random order nearest-neighbor method for small case sample sizes (n=10, 20, 40). Since the exact determination of properties (e.g., mean and variance) of the post-matching distribution of the controls was found to be analytically intractable, Monte Carlo simulation techniques were used to obtain the results given below.

To simplify further discussion, we now introduce the following notation for certain post-nearest-neighbor matching moments. We will use $\mu(n,n_0)$ to denote the post-matching mean of a sample of controls of size $n$ selected from a pool of controls of size $n_0$ by nearest-neighbor matching. Likewise, we will use $\sigma^2(n,n_0)$, $\text{Cov}(x_1,x_0|n,n_0)$ and $\rho(n,n_0)$ to denote the post-matching variance of the controls, the covariance of the controls with the cases, and the correlation of the controls with the cases, respectively.

The first measure of interest that we shall examine is the proportional reduction in mean difference. This is defined, in the usual manner, as

$$
\frac{\mu(n,n_0) - \mu_{x_0}}{\mu_{x_1} - \mu_{x_0}},
$$

(4.2.1.)

where $\mu_{x_1}$ is the expected value of $X_1$ before matching. Recall that this is equal to the percent reduction in bias in $(\bar{Y}_1 - \bar{Y}_0)$ under the assumption that the model $y = \alpha + \beta x + e$ holds. The criterion (4.2.1.) has
been previously studied by Rubin (1970, 1973a, 1973b). Table 4.2.1, below, summarizes the results of his study on the percent reduction in bias due to nearest-neighbor matching randomly ordered cases. He has used the function \( B = (\mu_{x_1} - \mu_{x_0}) \left/ \left( \frac{\sigma^2_{x_1} + \sigma^2_{x_0}}{2} \right)^{1/2} \right. \), to define the initial bias in the estimators, where \( \sigma^2_{x_1} \) and \( \sigma^2_{x_0} \) are as before. The case sample sizes he used were \( n = 25, 50, 100 \) and he considered "control pool size to case sample size ratios" of \( r = 2, 3, 4 \). His table shows four trends in the proportional reduction in bias. First, the value of (4.2.1.) decreases as \( B \), his initial bias term, increases. For given \( \sigma^2_{x_1} \) and \( \sigma^2_{x_0} \), this means that (4.2.1.) decreases as \( |\mu_{x_1} - \mu_{x_0}| \) increases. Secondly, as \( \sigma^2_{x_1}/\sigma^2_{x_0} \) increases, (4.2.1.) decreases. Thus, the proportional reduction in mean difference would be less for a control population that has a small variance relative to the cases than it would for a control population that has a large variance relative to the cases. As \( r \), the ratio of the control pool size to the case sample size, increases, the proportional increase in mean difference also increases. As \( n \), the case sample size, increases, (4.3.1.) increases. Thus, this method seems to perform best for large values of \( r \), and \( n \), and for small values of \( \sigma^2_{x_1}/\sigma^2_{x_0} \) and \( |\mu_{x_1} - \mu_{x_0}| \).

To verify Rubin's results, we also ran a simulation using the same values of \( r \), \( n \), \( B \), and \( \sigma^2_{x_1}/\sigma^2_{x_0} \). Our results were essentially identical and are not presented here. In Table 4.2.2, we present the results of a simulation for \( r = 2, 3, 4; \sigma^2_{x_0} = 1, 3; \mu_{x_0} = 0.2, 0.4, 0.6, 0.8, 1.0, \) and \( n = 10, 20, 40 \). As is usual, we have assumed \( X_i \sim N(0,1) \). Our
### TABLE 4.2.1a

Percent Reduction in Mean Difference for Random Order
Nearest Available Matching; X Normal

<table>
<thead>
<tr>
<th>( \sigma_1^2 / \sigma_2^2 )</th>
<th>( \sigma_1^2 / \sigma_2^2 = \frac{1}{2} )</th>
<th>( \sigma_1^2 / \sigma_2^2 = 1 )</th>
<th>( \sigma_1^2 / \sigma_2^2 = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = )</td>
<td>( r = )</td>
<td>( r = )</td>
<td>( r = )</td>
</tr>
<tr>
<td>( B = \frac{1}{4} )</td>
<td>( \frac{1}{2} )</td>
<td>( \frac{3}{4} )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>2</td>
<td>97</td>
<td>94</td>
<td>89</td>
</tr>
<tr>
<td>3</td>
<td>99</td>
<td>98</td>
<td>97</td>
</tr>
<tr>
<td>4</td>
<td>99</td>
<td>99</td>
<td>97</td>
</tr>
<tr>
<td>( N=25 )</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>87</td>
<td>87</td>
<td>87</td>
<td>87</td>
</tr>
<tr>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>63</td>
<td>63</td>
<td>63</td>
<td>63</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>( N=50 )</td>
<td>92</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>84</td>
<td>84</td>
<td>84</td>
<td>84</td>
</tr>
<tr>
<td>87</td>
<td>87</td>
<td>87</td>
<td>87</td>
</tr>
<tr>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>69</td>
<td>69</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>53</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>51</td>
<td>51</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>( N=100 )</td>
<td>99</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>86</td>
<td>86</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>69</td>
<td>69</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>( N=200 )</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>99</td>
<td>99</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>86</td>
<td>86</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>69</td>
<td>69</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
</tbody>
</table>

*a* from Rubin (1973a), Table 5.2.
TABLE 4.2.2

Proportional Reduction in Mean Difference Due to Nearest-Neighbor Matching of n Randomly Ordered Cases from a Control Pool of Size $n_0=rn$

$x_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$, $x_1 \sim N(0,1)$

<table>
<thead>
<tr>
<th>$\sigma^2_{x_0}$</th>
<th>n</th>
<th>r</th>
<th>$\mu_{x_0}=0.2$</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td>.75</td>
<td>.74</td>
<td>.70</td>
<td>.67</td>
<td>.62</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3</td>
<td>.88</td>
<td>.83</td>
<td>.82</td>
<td>.81</td>
<td>.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>.93</td>
<td>.92</td>
<td>.90</td>
<td>.84</td>
<td>.85</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>.86</td>
<td>.82</td>
<td>.80</td>
<td>.70</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>3</td>
<td>.91</td>
<td>.91</td>
<td>.88</td>
<td>.85</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>.96</td>
<td>.95</td>
<td>.93</td>
<td>.92</td>
<td>.85</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>2</td>
<td>.91</td>
<td>.90</td>
<td>.82</td>
<td>.74</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>.97</td>
<td>.95</td>
<td>.93</td>
<td>.88</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>.98</td>
<td>.97</td>
<td>.95</td>
<td>.93</td>
<td>.89</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td></td>
<td>.94</td>
<td>.94</td>
<td>.88</td>
<td>.88</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>.98</td>
<td>.98</td>
<td>.97</td>
<td>.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>.98</td>
<td>.98</td>
<td>.98</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
<td>.97</td>
<td>.96</td>
<td>.95</td>
<td>.94</td>
<td>.93</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>3</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>.99</td>
<td>1.00</td>
<td>.99</td>
<td>1.00</td>
<td>.99</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td></td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
<td>.98</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>
simulation shows the same trends in the proportional reduction in mean differences as does the simulation of Rubin. We have also calculated $\sigma^2(n,n_0)$, the post-matching variance of $X_0$, and $\rho(n,n_0)$, the post-matching correlation between $X_1$ and $X_0$. The results are presented in Tables 4.2.3 and 4.2.4, respectively. The post-matching variance, $\sigma^2(n,n_0)$, approaches one (the variance of $X_1$) as $r$ increases, as the $\mu_{x_0}$ approaches zero, and as $n$ increases. For $\sigma^2_{x_0} = 3$, the post-matching variance tends to approach one from above, but it approaches one from below for $\sigma^2_{x_0} = 1$. The post-matching correlation increases towards one as $r$, $n$, and $\sigma^2_{x_0}$ increases, and as $\mu_{x_0}$ decreases towards zero.

Recall that, in the last two chapters, we were able to show that the post-sampling properties of randomly matched (or unmatched) and exactly matched control groups could be developed as special cases of the two matching methods we studied in those chapters. By examination of the three tables just discussed, we can see that as $r$ increases, the post-sampling correlation approaches one, as does the post-sampling variance, while the post-sampling mean approaches zero, thus leading (as expected) to the results that have been derived for the exact pair-matching case. When $r=1$, which would correspond to the randomly matched (or unmatched) situation, the post-sampling variance of the controls is, of course, $\sigma^2_{x_0}$ and the mean is $\mu_{x_0}$, since there is no subsampling of the original random sample selected from the control pool. However, the within-pair correlation will, in general, be greater than zero, since the matching method tends to order the control matching variable values according to the values of the cases (i.e., the association of pairs is not random). So, at first glance, it appears that the randomly matched
TABLE 4.2.3
Variance of n Control Values Chosen from a Pool of Size n_0 = r n
by Random Order Nearest-Neighbor Matching

\[ x_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2), \quad x_1 \sim N(0, 1) \]

<table>
<thead>
<tr>
<th>( \sigma_{x_0}^2 )</th>
<th>n</th>
<th>r</th>
<th>( \mu_{x_0} = 0.2 )</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td>.80</td>
<td>.75</td>
<td>.70</td>
<td>.65</td>
<td>.64</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td></td>
<td>.87</td>
<td>.85</td>
<td>.76</td>
<td>.73</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>.88</td>
<td>.91</td>
<td>.85</td>
<td>.75</td>
<td>.71</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>20</td>
<td>.88</td>
<td>.81</td>
<td>.79</td>
<td>.68</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.94</td>
<td>.89</td>
<td>.84</td>
<td>.76</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>.93</td>
<td>.94</td>
<td>.88</td>
<td>.83</td>
<td>.69</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>40</td>
<td>.95</td>
<td>.90</td>
<td>.80</td>
<td>.68</td>
<td>.59</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.97</td>
<td>.96</td>
<td>.90</td>
<td>.80</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>.95</td>
<td>.96</td>
<td>.92</td>
<td>.85</td>
<td>.76</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td></td>
<td>1.13</td>
<td>1.11</td>
<td>1.15</td>
<td>1.16</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>1.07</td>
<td>1.03</td>
<td>1.06</td>
<td>1.03</td>
<td>1.02</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>20</td>
<td>1.08</td>
<td>1.06</td>
<td>1.07</td>
<td>1.12</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>1.01</td>
<td>.99</td>
<td>.97</td>
<td>1.04</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.04</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>40</td>
<td>1.05</td>
<td>1.06</td>
<td>1.04</td>
<td>1.06</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>1.04</td>
<td>1.01</td>
<td>1.00</td>
<td>1.00</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>1.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.97</td>
</tr>
</tbody>
</table>
TABLE 4.2.4

Correlation of matched values when controls are selected from a pool of size $n_0 = r n$ by random order nearest-neighbor matching

$$X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2), \ X_1 \sim N(0,1)$$

<table>
<thead>
<tr>
<th>$\sigma_{x_0}^2$</th>
<th>n</th>
<th>r</th>
<th>$\mu_{x_0} = 0.2$</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>10</td>
<td>3</td>
<td>.95</td>
<td>.93</td>
<td>.89</td>
<td>.85</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>.98</td>
<td>.97</td>
<td>.94</td>
<td>.93</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
<td>.94</td>
<td>.93</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>3</td>
<td>.97</td>
<td>.96</td>
<td>.94</td>
<td>.87</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>.99</td>
<td>.99</td>
<td>.97</td>
<td>.95</td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>.99</td>
<td>.98</td>
<td>.95</td>
<td>.89</td>
<td>.81</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>1</td>
<td>1.00</td>
<td>.99</td>
<td>.98</td>
<td>.96</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.98</td>
<td>.95</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>3</td>
<td>.93</td>
<td>.94</td>
<td>.90</td>
<td>.89</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>.98</td>
<td>.98</td>
<td>.97</td>
<td>.97</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
<td>.95</td>
<td>.98</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>3</td>
<td>.97</td>
<td>.96</td>
<td>.96</td>
<td>.95</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
<td>.97</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
(or unmatched) situation cannot be treated as a special case of the nearest-neighbor matching method with r=1. We can resolve this apparent difficulty by noting that, since there is no subsampling of the control sample, but only an ordering of the values within the sample, the sample means of the two samples will always be uncorrelated, since they are independent of each other. This can be clarified by considering the following procedure. Suppose we draw two independent samples of the same size from two groups. The control samples are "matched" by the nearest-neighbor method to the case samples and the means are calculated for each group, as is the matched-pair correlation. Repeat this procedure till there is a series of means and matched-pair correlations. Now, if we calculate the correlation between the case and control means, we would expect this to be zero, since the samples were selected independently. However, the average-matched pair-correlation (over the samples) will not be zero, since we are ordering the controls with respect to the cases. From this point of view, we can then treat the randomly matched or unmatched controls as a special case of nearest-neighbor matching (with r=1).

4.3. Covariance Analysis and Nearest-Neighbor Pair-Matching

We will now examine the relative efficiency and relative mean square error of the pair-matched mean difference estimator of the case-control difference parameter to the covariance analysis estimator of this parameter. We will assume, as we did in Sections 2.5. and 3.5., that the response variable is related to the matching variable by the usual straight line Ancova model (i.e., \( y = \alpha + \beta x + e \)). After finding the mean, the variance, and the bias of the nearest-neighbor pair-matched estimator under the assumed model, we shall compare the variance and mean
square error of this estimator to those of the usual covariance estimator (based on random samples), which is the traditional estimator used under the assumed model.

We will assume that the usual analysis of covariance model holds, namely

$$y_{1j} = \alpha_1 + \beta x_{1j} + e_{1j} \quad i=0,1; j=1,\ldots,n, \quad (4.3.1.)$$

where $e_{1j} \sim N(0, \sigma^2_e)$. Given that $n$ controls, from a pool of size $n_0$, have been nearest-neighbor pair-matched to $n$ cases, let $x_{0j}$ denote the control-matching variable observation that has been paired with $x_{1j}$, the $j$-th case-matching variable observation. The response variable values associated with $x_{0j}$ and $x_{1j}$ are denoted by $y_{0j}$ and $y_{1j}$, respectively. The within-pair difference for the response variable is then $d_j = (y_{1j} - y_{0j})$, and, from (4.3.1.), it follows that $d_j = (\alpha_1 - \alpha_0) + \beta(x_{1j} - x_{0j}) + (e_{1j} - e_{0j})$. So, the conditional mean and variance of $d_j$ given $x_{0j}$ and $x_{1j}$ are

$$E(d_j|x_{1j},x_{0j}) = (\alpha_1 - \alpha_0) + \beta(x_{1j} - x_{0j}),$$

and

$$\text{Var}(d_j|x_{1j},x_{0j}) = 2\sigma^2_e.$$

Thus, the unconditional mean and variance of $d_j$ are

$$E(d_j) = (\alpha_1 - \alpha_0) + \beta(\mu_{x_1} - \mu(n,n_0)),$$

and

$$\text{Var}(d_j) = 2\sigma^2_e + \beta^2[\sigma^2_{x_1} + \sigma^2_{x_0}(n,n_0) - 2\text{Cov}(x_1,x_0|n,n_0)].$$
The nearest-neighbor pair-matched mean difference estimator of the case-control difference \( (\alpha_1 - \alpha_0) \) is \( \bar{d} = 1/n \sum_{j=1}^{n} d_j \); so, from the moments we have just found, we have

\[
E(\bar{d}) = \alpha_1 - \alpha_0 + \beta [\mu_{x_1} - \mu(n, n_0)] ,
\]

so that

\[
\text{BIAS } (\bar{d}) = E(\bar{d}) - (\alpha_1 - \alpha_0) = \beta [\mu_{x_1} - \mu(n, n_0)] , \quad (4.3.2)
\]

and

\[
\text{Var } (\bar{d}) = \frac{2\sigma_x^2}{n} + \frac{\beta^2}{n} \left\{ \sigma_{x_1}^2 + \sigma^2(n, n_0) - 2 \text{Cov}(X_1, X_0 | n, n_0) \right\} \quad (4.3.3)
\]

Now, given the further assumption that \( X_i \sim N(\mu_{x_i}, \sigma_{x_i}^2) \), we can use the Monte Carlo results shown in the last section to give specific values for (4.3.3.) at the points we examined in our simulation. We shall now compare the nearest-neighbor pair-matched estimator \( \bar{d} \) with the covariance-analysis estimator \( \hat{\psi} \). We showed in Section 2.5., for \( X_1 \sim N(0, 1) \) and for \( X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2) \), that \( \text{BIAS } (\hat{\psi}) = 0 \), and that

\[
\text{Var } (\hat{\psi}) = \frac{2\sigma_x^2}{n} \left\{ 1 + \frac{(1+\sigma_{x_0}^2)^2 + (1+\sigma_{x_0}^2)^2 \mu_{x_0}^2}{2(n-1)(1+\sigma_{x_0}^2)^2 - 4(1+\sigma_{x_0}^2)^4} \right\} , \quad (4.3.4)
\]
when \( \hat{\psi} \) is estimated using random samples of size \( n \) from both groups. As before, we will consider the relative efficiency and relative mean square error of the two estimators, \( \bar{d} \) and \( \hat{\psi} \). Note that, from (4.3.3.)

\[
\text{Var} (\bar{d}) = \frac{2e^2}{n} \left\{ 1 + \frac{\rho^2_{x_1y_1}}{2(1-\rho^2_{x_1y_1})} \right\} \left[ 1 + \sigma^2(n,n_0) - 2 \text{Cov}(X_1,X_0|n,n_0) \right],
\]

(4.3.5.)

since, when \( X_1 \sim \text{N}(0,1) \), \( \sigma_e^2 = \sigma^2 \rho^2_{x_1y_1} \), and \( \beta^2 = \sigma^2 \rho^2_{x_1y_1} \). Also, under the same conditions,

\[
\text{MSE} (\bar{d}) = \text{Var} (\bar{d}) + \{\text{BIAS} (\bar{d})\}^2
\]

\[
= \frac{2e^2}{n} \left\{ 1 + \frac{\rho^2_{x_1y_1}}{2(1-\rho^2_{x_1y_1})} \right\} \left[ 1 + \sigma^2(n,n_0) - 2 \text{Cov}(X_1,X_0|n,n_0) \right] + \beta^2 \sigma^2(n,n_0)
\]

\[
= \frac{2e^2}{n} \left\{ 1 + \frac{\rho^2_{x_1y_1}}{2(1-\rho^2_{x_1y_1})} \right\} \left[ 1 + \sigma^2(n,n_0) - 2 \text{Cov}(X_1,X_0|n,n_0) \right]
\]

\[
+ \beta^2 \sigma^2(n,n_0)
\]

(4.3.6.)

In Sections 2.5. and 3.5., we discussed the use of a correction factor, based on the degrees of freedom for error of an estimator, to adjust the relative efficiency and relative mean square error comparisons. We shall continue to use such a correction factor in this
TABLE 4.3.1

Relative Efficiency of Random Case Ordering Nearest-Neighbor Pair-Matched Mean Difference Estimator to the Covariance Analysis Estimator

\[ x_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2), \quad x_1 \sim N(0,1) \]

\[ \sigma_{x_0}^2 = 1 \]

\begin{array}{cccccccccc}
\rho_{x_1y_1} & \mu_{x_0} & r = 2 & 3 & 4 & 2 & 3 & 4 & 2 & 3 & 4 \\
0.2 & .99 & .99 & 1.00 & .99 & .99 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\
0.4 & 1.03 & 1.03 & 1.03 & 1.02 & 1.03 & 1.03 & 1.03 & 1.03 & 1.03 & 1.03 \\
0.6 & 1.08 & 1.09 & 1.09 & 1.08 & 1.08 & 1.08 & 1.08 & 1.08 & 1.08 & 1.08 \\
0.8 & 1.16 & 1.16 & 1.17 & 1.15 & 1.15 & 1.15 & 1.15 & 1.15 & 1.15 & 1.15 \\
1.0 & 1.26 & 1.27 & 1.27 & 1.24 & 1.24 & 1.25 & 1.24 & 1.24 & 1.24 & 1.24 \\
0.2 & .99 & .99 & .99 & .99 & .99 & .99 & 1.00 & 1.00 & 1.00 & 1.00 \\
0.4 & 1.02 & 1.02 & 1.03 & 1.02 & 1.02 & 1.03 & 1.03 & 1.03 & 1.03 & 1.03 \\
0.6 & 1.07 & 1.08 & 1.08 & 1.07 & 1.007 & 1.08 & 1.07 & 1.08 & 1.08 & 1.08 \\
0.8 & 1.14 & 1.15 & 1.16 & 1.13 & 1.14 & 1.15 & 1.13 & 1.14 & 1.14 & 1.14 \\
1.0 & 1.23 & 1.25 & 1.26 & 1.21 & 1.23 & 1.23 & 1.21 & 1.23 & 1.24 & 1.24 \\
0.2 & .97 & .98 & .99 & .98 & .99 & .99 & .99 & 1.00 & 1.00 & 1.00 \\
0.4 & .99 & 1.01 & 1.02 & 1.00 & 1.02 & 1.02 & 1.02 & 1.03 & 1.03 & 1.03 \\
0.6 & 1.03 & 1.05 & 1.07 & 1.04 & 1.06 & 1.07 & 1.05 & 1.07 & 1.08 & 1.08 \\
0.8 & 1.08 & 1.12 & 1.13 & 1.08 & 1.12 & 1.14 & 1.09 & 1.13 & 1.14 & 1.14 \\
1.0 & 1.15 & 1.20 & 1.23 & 1.14 & 1.19 & 1.20 & 1.14 & 1.19 & 1.21 & 1.21 \\
\end{array}
TABLE 4.3.1 continued...

\[ \sigma^2_{x_0} = 1 \]

\[
\begin{array}{cccccccccccc}
\rho_{x_1y_1} & \mu_{x_0} & r = 2 & 3 & 4 & 2 & 3 & 4 & 2 & 3 & 4 \\
0.2 & .91 & .96 & .97 & .95 & .98 & .98 & .98 & .99 & 1.00 \\
0.4 & .92 & .98 & 1.01 & .96 & 1.00 & 1.01 & 1.00 & 1.02 & 1.02 \\
0.8 & .92 & .99 & 1.04 & .97 & 1.00 & 1.05 & .99 & 1.05 & 1.06 \\
0.8 & .93 & 1.03 & 1.06 & .94 & 1.05 & 1.10 & .97 & 1.07 & 1.11 \\
1.0 & .94 & 1.06 & 1.14 & .96 & 1.08 & 1.12 & .95 & 1.09 & 1.15 \\
\end{array}
\]
<table>
<thead>
<tr>
<th>$\rho_{x_1 y_1}$</th>
<th>$\mu_{x_0}$</th>
<th>$r = 2$</th>
<th>3</th>
<th>4</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.4</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.03</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>0.6</td>
<td>1.03</td>
<td>1.04</td>
<td>1.04</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>0.8</td>
<td>1.08</td>
<td>1.08</td>
<td>1.08</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>1.0</td>
<td>1.13</td>
<td>1.13</td>
<td>1.13</td>
<td>1.11</td>
<td>1.11</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
</tr>
<tr>
<td>0.2</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.4</td>
<td>1.00</td>
<td>1.01</td>
<td>1.01</td>
<td>1.00</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>0.6</td>
<td>1.02</td>
<td>1.03</td>
<td>1.04</td>
<td>1.02</td>
<td>1.03</td>
<td>1.03</td>
<td>1.03</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>0.8</td>
<td>1.06</td>
<td>1.08</td>
<td>1.08</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>1.0</td>
<td>1.10</td>
<td>1.12</td>
<td>1.13</td>
<td>1.10</td>
<td>1.12</td>
<td>1.11</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
</tr>
<tr>
<td>0.2</td>
<td>0.95</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.4</td>
<td>0.97</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>0.6</td>
<td>0.98</td>
<td>1.02</td>
<td>1.03</td>
<td>1.01</td>
<td>1.02</td>
<td>1.03</td>
<td>1.02</td>
<td>1.03</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>0.8</td>
<td>1.01</td>
<td>1.06</td>
<td>1.07</td>
<td>1.04</td>
<td>1.06</td>
<td>1.07</td>
<td>1.05</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>1.0</td>
<td>1.04</td>
<td>1.11</td>
<td>1.12</td>
<td>1.06</td>
<td>1.11</td>
<td>1.11</td>
<td>1.10</td>
<td>1.11</td>
<td>1.11</td>
<td>1.12</td>
</tr>
</tbody>
</table>
TABLE 4.3.2 continued...

\[ \sigma^2 = 3 \]

\[ x_0 \]

\[ \begin{array}{cccccccc}
\rho_{x_1y_1} & \mu_{x_0} & r = 2 & 3 & 4 & 2 & 3 & 4 & 2 & 3 & 4 \\
0.2 & .88 & .96 & .97 & .94 & .98 & .98 & .97 & .99 & .99 \\
0.4 & .90 & .97 & .99 & .94 & .99 & 1.00 & .99 & 1.01 & 1.01 \\
0.8 & .87 & .98 & 1.01 & .96 & .99 & 1.02 & 1.00 & 1.03 & 1.03 \\
0.8 & .89 & 1.03 & 1.05 & .97 & 1.04 & 1.06 & 1.02 & 1.06 & 1.07 \\
1.0 & .89 & 1.05 & 1.10 & .96 & 1.07 & 1.10 & 1.02 & 1.11 & 1.11 \\
\end{array} \]
section. When we incorporate this correction factor, we find the adjusted relative efficiency of $\bar{d}$ to $\hat{y}$ to be

$$\frac{n^2}{(n+2)(n-1)} \frac{V(\hat{y})}{V(\bar{d})} = \frac{(1+\sigma^2_{x_0})^2}{\mu_{x_0}} + \frac{n\mu_{x_0}^2}{x_0} \frac{x_0}{x_0} \frac{x_0}{x_0}$$

$$\frac{2(n-1)(1+\sigma^2_{x_0})^2}{4(1+\sigma^4_{x_0})}$$

$$\frac{n^2}{(n+2)(n-1)} \frac{\rho_{x_1y_1}^2}{1 + \frac{\rho_{x_1y_1}^2}{2 (1-\rho_{x_1y_1}^2)} [1 + \sigma^2(n,n_0) - 2 \text{Cov}(X_1,X_0|n,n_0)]}$$

(4.3.7.)

By careful examination of the expression (4.3.7.), we can spot several general trends. First, since $\text{Cov}(X_1,X_0|n,n_0)$ and $\sigma^2(n,n_0)$ both approach $\sigma^2_{x_1}$ ($=1$) as $r = n_0/n$ increases, expression (4.3.7.) should also increase as $r$ increases. Secondly, since $\mu_{x_0}^2$ appears in the numerator but not in the denominator, (4.3.7.) should increase as $\mu_{x_0}^2$ increases. Finally, $\rho_{x_1y_1}^2/(1-\rho_{x_1y_1}^2)$ increases as $\rho_{x_1y_1}^2$ increases; so, since $\rho_{x_1y_1}^2/(1-\rho_{x_1y_1}^2)$ appears in the denominator but not in the numerator, (4.3.7.) decreases as $\rho_{x_1y_1}^2$ increases. Tables 4.3.1 and 4.3.2 (for $\sigma^2_{x_0} = 1$ and $\sigma^2_{x_0} = 3$, respectively), give computed values of (4.3.7.) for $\mu_{x_0}^2 = 0.2, 0.4, 0.6, 0.8, 1.0; n = 10, 20, 40; r = 2, 3, 4; \rho_{x_1y_1}^2 = 0.2, 0.4, 0.6, 0.8$. As is usual, we have assumed that $X_1 \sim N(0,1)$ and $X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$, that each member of the case sample of size $n$ has been nearest-neighbor pair-matched to one of the $rn = n_0$ controls based on a random ordering in the
case sample, and that the covariance-analysis estimator is based on random samples of size n from both groups.

By examination of the tables, we can see that the suspected trends just discussed do indeed occur. There are also two other trends evident in the tables, namely, (4.3.7.) decreases as n decreases and as \( \sigma^2_{x_0} \) increases. With regard to the actual values of (4.3.7.), we can see that the nearest-neighbor pair-matched mean difference estimator is very rarely less efficient than the covariance analysis estimator (i.e., (4.3.7.) is mostly larger than one). This estimator is generally less efficient only for large \( \rho_{x_1y_1} \) and \( r = 2 \), and occasionally when \( r = 3 \) and \( \mu_{x_0} \) is small. Thus, it appears that \( \bar{d} \) is to be preferred to \( \hat{\psi} \) as a more efficient estimator of \( (\alpha_1 - \alpha_0) \), assuming equal costs in obtaining the final sample of size n in both populations.

In order to take into account the fact that, under our assumed model, \( \bar{d} \) is biased while \( \hat{\psi} \) is not, we will make a mean square error comparison of the two estimators. Since the bias of \( \hat{\psi} \) is zero, the mean square error of \( \hat{\psi} \) is the same as its variance. We gave an expression for the mean square error of \( \bar{d} \) in (4.3.6.). Then, as in Sections 2.5. and 3.5., the relative mean square error (adjusted for differing error d.f.'s) would be

\[
\frac{n^2}{(n+2)(n-1)} \left( 1 + \frac{(1+\sigma^2_{x_0})^2}{2(n-1)(1+\sigma^2_{x_0})^2 - 4(1+\sigma^4_{x_0})} \right) \frac{\rho_{x_1y_1}^2}{2(1-\rho_{x_1y_1}^2)} \left[ 1 + \sigma^2(n,n_0) - 2 \text{Cov}(X_1,X_0|n,n_0) + \mu_{x_0}^2(n,n_0) \right] \]

(4.3.8.)
By comparison with (4.3.7.), we can see that we would expect the
same trends to be present in (4.3.8.) that were present in (4.3.7.). As
$2
\rho_{x_1y_1}$ increases, expression (4.3.8.) will decrease in value. As $r = n_0/n$
increases, the expression should increase, since $\sigma^2(n,n_0)$ and $\text{Cov}(X_1,X_0|n,n_0)$ will go to one and $\mu(n,n_0)$ will go to zero. However, since
$\mu(n,n_0)$ increases as $\mu_{x_0}$ increases, (4.3.8.) will not increase as quickly
with increasing $\mu_{x_0}^2$ as did (4.3.7.). Also, by observation of the be-
behavior of (4.3.7.), we would expect (4.3.8.) to decrease as $\sigma^2_{x_0}$ increases.
It is not clear what the effect of increasing $n$ will be, since $n$ appears
in both the numerator and the denominator. Tables 4.3.3 and 4.3.4 (for
$\sigma^2_{x_0} = 1$ and $\sigma^2_{x_0} = 3$) exhibit some computed values of (4.3.8.) for the
same entries as in Tables 4.3.1 and 4.3.2.

Note that for $\mu_{x_0} > 0.6$, (4.3.8.) tends to decrease as $n$ increases
For $\mu_{x_0} < 0.4$, (4.3.8.) either increases or remains fairly constant with
increasing $n$. Another trend to note is that (4.3.8.) does not always
increase as $\mu_{x_0}$ increases. When $\rho_{x_1y_1}$ is large (0.6 or 0.8), we find
that (4.3.8.) decreases as $\mu_{x_0}$ increases when $r = 2$. For $r = 3$ or $4$
(4.3.8.) either increases slightly and then decreases or increases slowly
with increasing $\mu_{x_0}$.

By inspection of Tables 4.3.3 and 4.3.4, we conclude that, for the
values examined, the nearest-neighbor pair-matched mean difference esti-
mator has a lower adjusted mean square error than the covariance estima-
tor only when $\rho_{x_1y_1}$ is small ($\rho_{x_1y_1} < 0.4$). By consideration of the
of the trends seen in these tables, we can suggest that for large $r$
nearest-neighbor matching can lead to a better mean square error esti-
mator than would covariance analysis even for large values of $\rho_{x_1y_1}$. 
TABLE 4.3.3

Relative Mean Square Error of the Covariance Analysis Estimator to the Random Case
Ordering Nearest-Neighbor Pair-Matched Mean Difference Estimator

\[ x_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2), \ x_1 \sim N(0,1) \]

\[ \sigma_{x_0}^2 = 1 \]

\[ n = 10 \quad n = 20 \quad n = 40 \]

<table>
<thead>
<tr>
<th>( \rho_{x_1 y_1} )</th>
<th>( \mu_{x_0} )</th>
<th>( r = 2 )</th>
<th>( n = 10 )</th>
<th>( n = 20 )</th>
<th>( n = 40 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.4</td>
<td>1.02</td>
<td>1.03</td>
<td>1.03</td>
<td>1.02</td>
<td>1.03</td>
</tr>
<tr>
<td>0.6</td>
<td>1.08</td>
<td>1.08</td>
<td>1.09</td>
<td>1.07</td>
<td>1.08</td>
</tr>
<tr>
<td>0.8</td>
<td>1.15</td>
<td>1.16</td>
<td>1.16</td>
<td>1.12</td>
<td>1.15</td>
</tr>
<tr>
<td>1.0</td>
<td>1.23</td>
<td>1.25</td>
<td>1.26</td>
<td>1.19</td>
<td>1.23</td>
</tr>
<tr>
<td>0.2</td>
<td>1.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.4</td>
<td>1.01</td>
<td>1.02</td>
<td>1.03</td>
<td>1.01</td>
<td>1.02</td>
</tr>
<tr>
<td>0.6</td>
<td>1.04</td>
<td>1.07</td>
<td>1.08</td>
<td>1.04</td>
<td>1.06</td>
</tr>
<tr>
<td>0.8</td>
<td>1.07</td>
<td>1.13</td>
<td>1.14</td>
<td>1.02</td>
<td>1.11</td>
</tr>
<tr>
<td>1.0</td>
<td>1.09</td>
<td>1.18</td>
<td>1.23</td>
<td>1.00</td>
<td>1.15</td>
</tr>
<tr>
<td>0.2</td>
<td>0.96</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>0.4</td>
<td>0.96</td>
<td>1.00</td>
<td>1.02</td>
<td>0.98</td>
<td>1.01</td>
</tr>
<tr>
<td>0.6</td>
<td>0.95</td>
<td>1.02</td>
<td>1.06</td>
<td>0.97</td>
<td>1.03</td>
</tr>
<tr>
<td>0.8</td>
<td>0.92</td>
<td>1.07</td>
<td>1.08</td>
<td>0.82</td>
<td>1.04</td>
</tr>
<tr>
<td>1.0</td>
<td>0.84</td>
<td>1.04</td>
<td>1.15</td>
<td>0.72</td>
<td>1.00</td>
</tr>
</tbody>
</table>
TABLE 4.3.3 continued...

\[ \sigma^2_{x_0} = 1 \]

<table>
<thead>
<tr>
<th>( \rho_{x_1 y_1} )</th>
<th>( u_{x_0} )</th>
<th>( n = 10 )</th>
<th>( n = 20 )</th>
<th>( n = 40 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r = 2 )</td>
<td>( 3 )</td>
<td>( 4 )</td>
<td>( 2 )</td>
</tr>
<tr>
<td>0.2</td>
<td>.89</td>
<td>.95</td>
<td>.97</td>
<td>.94</td>
</tr>
<tr>
<td>0.4</td>
<td>.85</td>
<td>.95</td>
<td>1.00</td>
<td>.88</td>
</tr>
<tr>
<td>0.8</td>
<td>.74</td>
<td>.91</td>
<td>1.01</td>
<td>.79</td>
</tr>
<tr>
<td>0.8</td>
<td>.62</td>
<td>.87</td>
<td>.93</td>
<td>.51</td>
</tr>
<tr>
<td>1.0</td>
<td>.49</td>
<td>.74</td>
<td>.96</td>
<td>.38</td>
</tr>
</tbody>
</table>
TABLE 4.3.4
Relative Mean Square Error of the Covariance Analysis Estimator to the Random Case Ordering Nearest-Neighbor Pair-Matched Mean Difference Estimator

\[ X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2), \quad X_1 \sim N(0,1) \]

\[ \sigma_{x_0}^2 = 3 \]

\( n = 10 \quad n = 20 \quad n = 40 \)

<table>
<thead>
<tr>
<th>( \rho_{x_1'y_1} )</th>
<th>( \mu_{x_0} )</th>
<th>( r = 2 )</th>
<th>( 3 )</th>
<th>( 4 )</th>
<th>( 2 )</th>
<th>( 3 )</th>
<th>( 4 )</th>
<th>( 2 )</th>
<th>( 3 )</th>
<th>( 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.6</td>
<td>1.01</td>
<td>1.04</td>
<td>1.01</td>
<td>1.04</td>
<td>1.01</td>
<td>1.01</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>0.8</td>
<td>1.08</td>
<td>1.08</td>
<td>1.08</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>1.0</td>
<td>1.12</td>
<td>1.13</td>
<td>1.13</td>
<td>1.11</td>
<td>1.12</td>
<td>1.12</td>
<td>1.11</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
</tr>
</tbody>
</table>

| 0.2             | 0.98           | 0.99   | 0.99   | 0.98   | 0.99   | 0.99   | 0.99   | 0.99   | 0.99   | 0.99   |
| 0.4             | 1.00           | 1.01   | 1.01   | 1.00   | 1.00   | 1.00   | 1.01   | 1.01   | 1.01   | 1.01   |
| 0.6             | 1.01           | 1.04   | 1.04   | 1.02   | 1.03   | 1.03   | 1.03   | 1.04   | 1.04   | 1.04   |
| 0.8             | 1.05           | 1.07   | 1.08   | 1.05   | 1.07   | 1.07   | 1.07   | 1.07   | 1.07   | 1.07   |
| 1.0             | 1.09           | 1.12   | 1.13   | 1.10   | 1.12   | 1.12   | 1.10   | 1.12   | 1.12   | 1.12   |

| 0.2             | 0.95           | 0.98   | 0.98   | 0.97   | 0.98   | 0.99   | 0.99   | 0.99   | 0.99   | 0.99   |
| 0.4             | 0.97           | 1.00   | 1.00   | 0.98   | 1.00   | 1.00   | 1.00   | 1.00   | 1.00   | 1.00   |
| 0.6             | 0.97           | 1.02   | 1.03   | 1.00   | 1.02   | 1.03   | 1.02   | 1.02   | 1.02   | 1.02   |
| 0.8             | 0.99           | 1.06   | 1.07   | 1.02   | 1.06   | 1.07   | 1.05   | 1.07   | 1.07   | 1.07   |
| 1.0             | 1.00           | 1.10   | 1.12   | 1.04   | 1.10   | 1.13   | 1.06   | 1.11   | 1.12   | 1.12   |
TABLE 4.3.4 continued...

\[ \sigma^2 = 3 \]

\( x_0 \)

<table>
<thead>
<tr>
<th>( \rho_{x_1y_1} )</th>
<th>( \mu_{x_0} )</th>
<th>( n = 10 )</th>
<th>( n = 20 )</th>
<th>( n = 40 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r = 2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>0.2</td>
<td>.88</td>
<td>.96</td>
<td>.97</td>
<td>.94</td>
</tr>
<tr>
<td>0.4</td>
<td>.80</td>
<td>.97</td>
<td>.99</td>
<td>.93</td>
</tr>
<tr>
<td>0.8</td>
<td>.84</td>
<td>.98</td>
<td>1.01</td>
<td>.95</td>
</tr>
<tr>
<td>0.8</td>
<td>.84</td>
<td>1.02</td>
<td>1.05</td>
<td>.94</td>
</tr>
<tr>
<td>1.0</td>
<td>.80</td>
<td>1.03</td>
<td>1.09</td>
<td>.89</td>
</tr>
</tbody>
</table>
However, for the small values of r we have examined, it can be seen that, for situations where there is a strong correlation, covariance analysis is recommended.

4.4. **Mean Square Error Comparisons of the Three Matching Methods**

In this section, we shall compare the three matching methods that we have previously considered individually. Our comparisons shall be based on the mean square error of the mean difference estimator of the case-control difference when the estimator is computed using observations chosen using each of the three matching methods. The methods shall be compared in such a manner as to equalize, as much as possible, the control pool sample size requirements.

Throughout this dissertation we have assumed that the simple linear Ancova model (i.e., \( y = \alpha + \beta x + e \)) adequately described the relationship between the matching variable and the response variable. We have compared the mean difference estimator for each of the three methods (category-matching, caliper-matching, and nearest-neighbor matching) to the covariance-analysis estimator, which was based on the use of randomly selected observations from the case and control groups. In general, when there is a low correlation between the response and the matching variable and the difference in the mean value of the matching variable in the two groups is large, then the mean-difference estimator under matching generally has a lower mean square error than the covariance-analysis estimator for small samples. This suggests that it would be useful to know which of the three matching methods is best, in terms of mean square error.

Before the three methods can be equitably compared we need to consider the different control group sample size requirements associated
with these matching methods. In Section 2.4., we studied the distribution of the least number of randomly sampled controls required to category-match a given number of cases. We also presented an approximate distribution, in Section 3.4., for the least number of controls required to caliper-match a given number of cases. However, nearest-neighbor pair-matching does not involve any such control group sample size considerations as these. Rather, it is the available control pool sample size, n₀, that completely dictates the performance of the nearest-neighbor matching procedure, as opposed to the other two methods, where additional considerations (i.e., caliper size and category structure) influence the performance of the methods for a given case and control group matching variable distribution. Thus, the cases are always completely matched using the nearest-neighbor matching method, but not necessarily with the other two procedures.

We can adjust for the differing control pool size effects in the following manner. For a given case sample size and given case and control matching variable distributions, we can find the expected least number of randomly sampled controls necessary to category match all the cases. Then using this expected value, we can find the caliper width, c*, say, associated with the same expected least number of controls required to completely caliper match all the cases. Finally, we can also use this expected value as the control pool size for nearest-neighbor matching. This approach "equalizes" the sample size requirements for caliper and category matching, so that we can compare these two methods on an equitable basis. The difficulty remains in equalizing the control pool sample size requirements of these two methods with the requirements for nearest-neighbor matching. This comparison will, in
fact, tend to penalize nearest-neighbor matching since the other two methods will not always obtain complete matches at the given sample size, whereas nearest-neighbor matching always will. However, we shall show that nearest-neighbor matching still performs best, even with this handicap.

We now introduce the following notation to facilitate comparison of the three mean difference estimators of the case-control difference based on the three matching methods. Let $\bar{d}_f$ denote the "category-matched" estimator, let $\bar{d}_c$ denote the "caliper-matched" estimator, and let $\bar{d}_n$ denote the "nearest-neighbor matched" estimator. Then from (2.5.24.),

$$\text{MSE} (\bar{d}_f) = \frac{2\sigma^2}{n} \left\{ 1 + \frac{\rho^2}{2(1-\rho^2)} \frac{1}{\chi_1 y_1} \left[ 1 + \sigma^2 \frac{x_0}{\sum_{j=1}^{r} p_j [1+H_{0,j} - J^2_{0,j}]} \right] \right. \right.$$ 

$$\left. + n\{\mu_{x_0} + \sigma_{x_0} \sum_{j=1}^{r} p_j J_{0,j}\} \right\}^2,$$

for $X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0})$ and $X_1 \sim N(0,1)$. Likewise, under the same assumptions, from (3.5.8.)

$$\text{MSE} (\bar{d}_c) = \frac{2\sigma^2}{n} \left\{ 1 + \frac{\rho^2}{2(1-\rho^2)} \frac{1}{\chi_1 y_1} \left[ 1 + \sigma^2 (c*) - 2 \text{Cov}(X_1, X_0 | c*) + n\mu^2(c*) \right] \right\},$$

where $c*$ is such that both the category-matching and the caliper-matching
both have the same expected least control pool sample size, \(n_0^*\). And for this particular control pool sample size, we have from (4.3.6.) that

\[
\text{MSE} \left( \bar{d}_n \right) = \frac{2\sigma^2}{n} \left\{ 1 + \frac{\rho^2_{x_1 y_1}}{2(1-\rho^2_{x_1 y_1})} \right\} \left[ 1 + \sigma^2(n_0, n_0^*) - 2 \text{Cov}(X_{1 o}, X_0 | n, n_0^*) + n \mu^2(n, n_0^*) \right].
\]

Note that all three mean square error expression are of the general form

\[
\frac{2\sigma^2}{n} \left\{ 1 + \frac{\rho^2_{x_1 y_1}}{2(1-\rho^2_{x_1 y_1})} \right\} (g)
\]

where \((g)\) is some function dependent on the particular matching method being used. Since we are only interested in comparing the matching methods, and not in the particular values of ratios of mean square errors, it follows that the parameter \(\rho^2_{x_1 y_1}\) can be considered to be a nuisance parameter. Thus, we can ignore this term in our comparisons.

So, the functions we will be comparing are

\[
g \left( \bar{d}_x \right) = 1 + \sigma^2 \left( \sum_{j=1}^{r} p_j (1 + H_{j,0}, -J_{0,j}^2) \right) + n (\mu_0 + \sum_{j=1}^{r} p_j J_{0,j})^2
\]

\[
g \left( \bar{d}_c \right) = 1 + \sigma^2(c^*) - 2 \text{Cov}(X_{1 o}, X_0 | c^*) + n \mu^2(c^*)
\]
\[ g(n) = 1 + \sigma^2(n, n^*) - 2 \text{Cov}(X_1, X_0 | n, n^*) + n \mu^2(n, n^*) \]

Our comparisons will be for both three and four category-matching structures. We will assume that both category structures are equiprobable for the case distribution and that there are the same number of cases in all categories. The case sample sizes we will consider are \( n = 12, 24, 48 \), values for which there can be an equal number of cases in each category. As usual, we will assume that \( X_0 \sim N(\mu_{x_0}, \sigma^2_{x_0}) \) and \( X_1 \sim N(0, 1) \). We will examine the three functions above at the values \( \mu_{x_0} = 0.2, 0.4, 0.6, 0.8, 1.0 \) and at \( \sigma^2_{x_0} = 1, 3 \). Table 4.4.1 presents values for these functions. As in the last two sections, the values for nearest-neighbor matching were computed using Monte Carlo methods.

There are several conclusions we can draw based on this table. The first is that category-matching has the highest mean square error, i.e., \( g(\overline{d}_f) \) always has the highest values. Secondly, when \( \sigma^2_{x_0} = 3 \), nearest-neighbor matching always gives the lowest values of mean square error. For \( \sigma^2_{x_0} = 1 \), the situation is more complicated. When \( \mu_{x_0} < 0.6 \), nearest-neighbor matching is always best. When \( \mu_{x_0} > 0.8 \), caliper-matching is always best. However, note that as \( n \) (the case sample size) increases, the point at which \( g(\overline{d}_c) > g(\overline{d}_n) \) increases in \( \mu_{x_0} \). This suggests that when the case sample size is very large, nearest-neighbor matching will always be better for moderate \( |\mu_{x_0} - \mu_{x_1}| \). Also, recall that this method of comparison tends to favor category-matching and caliper-matching over nearest-neighbor matching, since we are only
TABLE 4.4.1

Mean Square Error Function Comparisons of r-Category Matching \{g(d_f)\},
Caliper Matching \{g(d_c)\}, and Nearest-Neighbor Matching \{g(d_n)\}

\[ X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2), X_1 \sim N(0,1) \]

\[ r = 3 \]

<table>
<thead>
<tr>
<th>σ²</th>
<th>n</th>
<th>( g(d_f) )</th>
<th>( g(d_c) )</th>
<th>( g(d_n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( g(d_f) ) 1.520</td>
<td>( g(d_c) ) 0.622</td>
<td>( g(d_n) ) 0.214</td>
</tr>
<tr>
<td>σ²</td>
<td>24</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>x₀</td>
<td></td>
<td>( g(d_f) ) 2.040</td>
<td>( g(d_c) ) 0.870</td>
<td>( g(d_n) ) 0.191</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( g(d_f) ) 3.080</td>
<td>( g(d_c) ) 1.252</td>
<td>( g(d_n) ) 0.243</td>
</tr>
<tr>
<td>σ²</td>
<td>24</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>x₀</td>
<td></td>
<td>( g(d_f) ) 6.028</td>
<td>( g(d_c) ) 0.730</td>
<td>( g(d_n) ) 0.194</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( g(d_f) ) 11.057</td>
<td>( g(d_c) ) 0.829</td>
<td>( g(d_n) ) 0.104</td>
</tr>
<tr>
<td>σ²</td>
<td>24</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>x₀</td>
<td></td>
<td>( g(d_f) ) 21.113</td>
<td>( g(d_c) ) 0.914</td>
<td>( g(d_n) ) 0.047</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

\[ \mu_{x_0} \]
TABLE 4.4.2

Mean Square Error Function Comparisons of r-Category Matching \{g(\overline{d}_f)\},
Caliper Matching \{g(\overline{d}_c)\}, and Nearest-Neighbor Matching \{g(\overline{d}_n)\}

\[ X_0 \sim N(\mu_{x_0}, \sigma_{x_0}^2), \quad X_1 \sim N(0,1) \]

\[ r = 4 \]

<table>
<thead>
<tr>
<th></th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\mu_{x_0}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\sigma_{x_0}^2 = 1</td>
<td>g(\overline{d}_f)</td>
<td>1.236</td>
<td>1.246</td>
<td>1.262</td>
<td>1.285</td>
</tr>
<tr>
<td>n=12</td>
<td>g(\overline{d}_c)</td>
<td>0.445</td>
<td>0.300</td>
<td>0.218</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>g(\overline{d}_n)</td>
<td>0.159</td>
<td>0.173</td>
<td>0.252</td>
<td>0.249</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\mu_{x_0}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\sigma_{x_0}^2 = 1</td>
<td>g(\overline{d}_f)</td>
<td>1.473</td>
<td>1.492</td>
<td>1.524</td>
<td>1.571</td>
</tr>
<tr>
<td>n=24</td>
<td>g(\overline{d}_c)</td>
<td>0.654</td>
<td>0.556</td>
<td>0.342</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td>g(\overline{d}_n)</td>
<td>0.133</td>
<td>0.228</td>
<td>0.295</td>
<td>0.229</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\mu_{x_0}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\sigma_{x_0}^2 = 1</td>
<td>g(\overline{d}_f)</td>
<td>1.946</td>
<td>1.984</td>
<td>2.048</td>
<td>2.141</td>
</tr>
<tr>
<td>n=48</td>
<td>g(\overline{d}_c)</td>
<td>0.984</td>
<td>0.935</td>
<td>0.567</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>g(\overline{d}_n)</td>
<td>0.141</td>
<td>0.195</td>
<td>0.219</td>
<td>0.332</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\mu_{x_0}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\sigma_{x_0}^2 = 1</td>
<td>g(\overline{d}_f)</td>
<td>3.436</td>
<td>3.471</td>
<td>3.529</td>
<td>3.611</td>
</tr>
<tr>
<td>n=12</td>
<td>g(\overline{d}_c)</td>
<td>0.460</td>
<td>0.465</td>
<td>0.455</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>g(\overline{d}_n)</td>
<td>0.089</td>
<td>0.101</td>
<td>0.125</td>
<td>0.128</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\mu_{x_0}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\sigma_{x_0}^2 = 3</td>
<td>g(\overline{d}_f)</td>
<td>5.872</td>
<td>5.942</td>
<td>6.058</td>
<td>6.222</td>
</tr>
<tr>
<td>n=24</td>
<td>g(\overline{d}_c)</td>
<td>0.563</td>
<td>0.613</td>
<td>0.641</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td>g(\overline{d}_n)</td>
<td>0.045</td>
<td>0.048</td>
<td>0.050</td>
<td>0.076</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\mu_{x_0}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\sigma_{x_0}^2 = 3</td>
<td>g(\overline{d}_f)</td>
<td>10.744</td>
<td>10.883</td>
<td>11.116</td>
<td>11.444</td>
</tr>
<tr>
<td>n=48</td>
<td>g(\overline{d}_c)</td>
<td>0.682</td>
<td>0.832</td>
<td>1.020</td>
<td>1.143</td>
</tr>
<tr>
<td></td>
<td>g(\overline{d}_n)</td>
<td>0.022</td>
<td>0.025</td>
<td>0.033</td>
<td>0.037</td>
</tr>
</tbody>
</table>
looking at the "expected" minimum control pool size. If we were to view the 90% bounds, say, the comparison would also be very much more in favor of nearest-neighbor matching. In summary then, we suggest that nearest-neighbor matching will generally give the lowest mean square error of the three methods we have studied where we consider the control pool sample size requirements associated with each of the three methods.
CHAPTER V
SUMMARY AND PROBLEMS FOR FUTURE RESEARCH

5.1. Summary

This dissertation has been concerned with the effects of three different matching methods to be used for continuous variables. We have considered the following three methods: category (or frequency) matching, which treats a continuous variable as a discrete variable; caliper matching, which matches two individuals whenever their matching variable values fall within a certain distance of each other; and, nearest-neighbor matching, which attempts to find the closest match possible for each case from the whole pool of controls. For each matching method, we have assessed the effects of the matching on the post-matching distribution of a matching variable which is normally distributed before the matching method is applied. This has included examining the post-matching mean (in terms of the proportional reduction in the original mean difference of the matching variable in the two groups), the post-matching variance, and for caliper and nearest-neighbor matching, the correlation between the case and control matching variable values after matching. We showed that each of these matching methods can substantially reduce the pre-matching mean difference between the expected values of the matching variable in the two groups. All methods also reduce the post-matching variance of the control matching variable; and, for caliper and nearest-neighbor matching, we found that the
matching procedure can induce a substantial correlation between the post-
sampling matching variable values in the two groups.

We have also looked at the sample size requirements associated
with the use of categorical or caliper matching. For both of these
methods, we studied the distribution of the number of randomly sampled
controls required to match all cases for a given category structure or
caliper size. This allowed us to quantify requirements on the control
pool size, for a particular category structure or caliper width. These
results also formed the basis for a later comparison of the mean square
errors of the mean difference estimators of the case-control parameters
for the two methods.

Under the assumption that the linear model \( y = \alpha + \beta x + \epsilon \) ade-
quately describes the relationship between the response variable \( y \) and
the matching variable \( x \), we examined efficiency and mean square error
characteristics of the mean-difference estimator of the case-control
difference \((\alpha_1 - \alpha_0)\) for each of the three matching methods. We compared
these results to those for the covariance-analysis estimator of \((\alpha_1 - \alpha_0)\),
namely \((\bar{y}_1 - \bar{y}_0) - \hat{\beta}(\bar{x}_1 - \bar{x}_0)\), which was based on random samples from both
groups. In general, the mean difference estimator under matching was
more efficient than the covariance-analysis estimator when the correla-
tion between the response and the matching variable was low and when the
true mean difference between the case and control matching variables
was large. Further, the mean square error of the mean difference
estimator was smaller than that of the covariance analysis estimator
when the above conditions held and the sample size was small \((n \leq 20, \text{ say})\). We have pointed out several times in the text, and wish to
restate here, that these comparisons between the covariance analysis
estimator and the matched estimator are valid only when there is no cost associated with obtaining the large numbers of controls necessary to match all cases. If there is a cost, and if this is included in the relative efficiency or mean square error comparisons, then the covariance analysis is more "cost-efficient" than the matched analysis.

Finally, we have compared the three methods to determine which method gives the smallest mean square error for the same control pool size requirements. We found that the mean-difference estimator based on categorically matched cases and controls would always have a larger mean square error than that based on either caliper or nearest-neighbor matched data. For certain situations the mean square error of the caliper-matched estimator was less than that of the nearest-neighbor matched estimator. However, in general, the nearest-neighbor matched estimator will have the smallest mean square error. Thus, we would recommend using nearest-neighbor matching for situations where pair-matching on a continuous variable is desired.

The remaining sections of this chapter discuss some research areas that require further investigation, and outline some possible approaches for dealing with these problems.

5.2. Multiple Matching Variables

The techniques we have considered all pertain to matching on a single continuous variable. It is not at all uncommon for an investigator to be faced with having to consider several matching variables simultaneously. Thus, it would be useful to study techniques for simultaneously matching on more than one variable. The currently available approaches to multivariable matching can be broken down into two
types. The first type is a straightforward extension of the univariate methods to the multivariable situation. The second type reduces the multivariable problem to a univariate one through the use of some sort of scoring function. We will briefly discuss the multivariable generalizations of the matching techniques used in this dissertation, and will then consider the scoring techniques.

A simple extension of category matching to a multivariable framework would involve forming categories on each matching variable, and then matching on all categories (i.e., the cells formed by the intersecting category boundaries) simultaneously. If we assume that the matching variables have a multivariate normal distribution, the effects of this matching procedure can be examined using an approach similar to the one we used in Chapter II. Now, since it is likely that these variables will be correlated, there can be empty or nearly empty categories. There is also the further problem of different correlation structures in the two groups. A possible way to alleviate these problems would be to transform the case matching variables so that they are independent. Then, based on these independent variables, the categories can be formed and treated as above, although the problem of different correlation matrices will still be there. Another approach could be used when the matching variables are not of equal importance. If the matching variables can be ranked in order of importance, the categories can be formed on the variables and the following matching scheme can be used: The controls would initially be classified into categories based on the first (i.e., most important) matching variable. Then, within each of these categories, the cases are further subsampled according to the categories of the second matching variable, and so on.
for the lower-ranking matching variables. This method has the advantage that it matches almost completely on the more important variables; and, when the matching must be terminated, due to a lack of available controls, any remaining, unmatched, variables will be the less important ones. This method needs to be studied to assess the relative efficiency of matching to various "levels", and to quantify sample size requirements needed to match to various levels.

An obvious extension of caliper-matching would be to choose as a match for a given case with a vector of matching variables \((x_{1j}, \text{ say})\) any control \((x_{0j}, \text{ say})\) such that the joint inequality \(|x_{0j} - x_{1j}| \leq \xi\) holds, where \(\xi\) is a vector of calipers, the value \(c_k\) being the caliper for the \(k\)-th matching variable. A further refinement would be to transform the case-matching variables so that they are mutually independent. Also, we could allow for "partial" matches, where a control was required only to satisfy a certain subset of the caliper inequalities and is allowed to violate the remaining ones; an equivalent effect can be achieved by setting narrow calipers on the "important" matching variables and wider calipers on the "unimportant" matching variables.

A multivariable extension of nearest-neighbor matching would involve matched based on the smallest Euclidian distance between the case- and control-matching variable values. There are several possible ways that this can be done. First, we can simply compute the Euclidian distance as the sum of the squared distances for each variable (i.e.,
\[ D_{jj'}^2 = (x_{1j} - x_{0j})' (x_{1j} - x_{0j}) \]). This distance measure could be refined by adjusting for the variance-covariance structure in the case group and using
\[ D_{jj'}^2 = (x_{1j} - x_{0j})' \sum_{l=1}^{\infty} (x_{lj} - x_{0j}) \] (which is simply the Mahalanobis
distance between the two observations $x_{ij}$ and $x_{0j}$ with respect to $\bar{I}$, or by assigning weights to the distances, based on the relative importance of the variables in the matching scheme (e.g., by using $D_{ij}^2 = (x_{ij} - x_{0j})' A(x_{ij} - x_{0j})$). Some work has been done on these matching methods (Rubin (1976)), but further investigation is still needed.

As mentioned above, a second approach to the multivariable matching problem is to reduce the matching variable values to a single index via the use of a "scoring" function. Several approaches can be suggested along this line. Rubin (1970, 1976) has suggested using the linear discriminant function, computed with respect to the estimated control variance-covariance matrix. Miettinen (1974) has suggested the use of a discriminant function that includes the response as a discriminator and is computed with respect to the pooled variance-covariance matrix. Miettinen (1976) has also suggested the use of a fitted linear model function to score the individuals prior to matching. Another method that would reduce the dimensionality of the multivariable problem is a principal-components approach. All three of these approaches need further investigation to determine their properties.

5.3. **Tests for Choosing Matching Variables**

Currently, the choice of matching variables tends to be left to the subject matter expert. An area that has not been adequately investigated concerns the use of statistical tests to determine if a variable should be matched on. There have been some tests suggested for this purpose, which we shall discuss below, but there is still need for
further investigation to determine the properties of these tests and to develop more powerful tests. The characteristics of such tests will, of course, depend on what measure is to be affected by the matching procedure (e.g., bias, variance, or mean square error), and will also be model dependent.

Three types of tests have been proposed in the literature. Cochran (1968) has discussed the use of a t-test to determine if a continuous variable needs to be controlled for. This simple approach was discussed in our literature review, where we reported that Cochran showed that a confidence interval about the mean difference in the response variable can be quite unsatisfactory, in terms of actual confidence, when the test for mean-difference in the matching variable is insignificant. Cochran (1968) has also suggested the use of a $\chi^2$ test for homogeneity to determine if a categorical-matching variable needs to be matched on. Little is known about the effect of not matching on a variable which has an insignificant $\chi^2$ value. Miettinen has also suggested a procedure to determine if a categorical variable should be matched on. He suggests that the statistic of interest be calculated for each value of the putative matching variable and for the population pooled over all values of the putative matching variable (i.e., the "crude" statistic). If the value of the crude statistic lies outside of the range of a standardized statistic, he suggests that the variable be controlled for in the analysis. Although this procedure has some intuitive appeal, it has not yet been subjected to any rigorous investigation.
5.4. **Other Areas of Interest**

In addition to the topics mentioned above, we feel that there are several other areas in this field where research could profitably be directed. These include the following:

(a) **Other Response Functions**

In this dissertation we have limited our investigation to the simple model \( y = \alpha + \beta x + e \). Although this is a natural starting point for an investigation of a single response and single-matching variable, it is certainly a very special case. Other functions of interest could be power functions of the matching variable, different functions in the case and control groups, and non-linear functions of the matching variable.

(b) **Other Distributions for the Matching Variable**

Some other distributions that could be of interest are the gamma distribution, the beta distribution, and various forms of the triangular distribution. These can be studied in the same manner as we did for the normal distribution.

(c) **Further Refinement for Categorical-Matching**

In our study, we used a very simple form of categorical matching in which any control was eligible to be a match. In Chapter IV, we showed that this method performed quite poorly in comparison with the other two methods. Perhaps the performance of this method could be improved by altering the method somewhat. One possible refinement would be to exclude from consideration any control that falls outside the observed range of the case values. A second alternative would be to exclude any control that is greater than (or less than) a confidence interval on the greatest (or smallest) order statistic from the case
population. Using either one of these methods would decrease the mean-difference in the two boundary categories, thus decreasing the overall mean-difference in the two post-matching populations, and hence, the mean square error of the mean-difference estimator.

(d) Combining Covariance Analysis with Matched Sampling

Although we have studied only the mean-difference estimator of the case-control difference, there is no reason that a more refined analysis could not be applied to the matched samples. For instance, a preliminary study of the relative efficiency of covariance analysis of caliper matched data to the mean-difference estimator when \( \rho_{x_1y_1} \) is large showed the covariance analysis to be more efficient. This suggests that a more refined analysis might provide better estimates than the simple analysis we used.
APPENDIX

THE $\lambda$-TH FACTORIAL MOMENT OF THE LEAST CONTROL POOL SIZE FOR A FOUR CATEGORY-MATCHING SCHEME

In Section 2.4., we discussed the distribution of the least number of randomly sampled controls necessary to match all cases using category-matching. We showed that the $\lambda$-th factorial moment could be expressed in the form

$$
E \left\{ \frac{(N+\lambda-1)!}{(N-1)!} \right\} = \sum_{n=M}^{\infty} \frac{(n+\lambda-1)!}{(n-1)!} P(n)
$$

(2.4.2.)

where $N$ is the random variable denoting the least control pool sample size, and $P(n)$ is the probability that the $n$-th sample matches the final member of a case sample of size $M$.

In the material that follows, we will derive an explicit form of (2.4.2.) for the general four category-matching scheme and find a general form for (2.4.2.) for any three category-matching scheme as a special case of the four category-matching scheme.

We will use the following relationships several times, so we will state them here and use them as needed:

1. \[ \sum_{n=K}^{\infty} \binom{n}{w} p^{w}(1-p)^{n-w} = p^{-1} I_{1-p}(K-w,w+1) \]
2. \[ \sum_{i=a}^{b} \binom{b+c}{i} p^i (1-p)^{b+c-i} = \sum_{i=0}^{b} \binom{b+c}{i} p^i (1-p)^{b+c-i} - \sum_{i=0}^{a-1} \binom{b+c}{i} p^i (1-p)^{b+c-i} \]

3. \[ \sum_{i=0}^{b} \binom{b+c}{i} p^i (1-p)^{b+c-i} = 1 - \sum_{j=0}^{c-1} \binom{b+c}{j} (1-p)^j p^{b+c-j} \]

The first of these can be found in Young (1961, p. 335), the second is self-evident, and the third can be obtained by using the transformation \( j = (b+c) - i \).

In the general four category-matching scheme, we would want to match the \( n_k \) case observations in the \( k \)-th category with \( n_k \) control observations which are also in the \( k \)-th category. The probability that a randomly sampled control observation falls into the \( k \)-th category is \( p_k (\sum_{k} p_k = 1) \). Define \( M = \sum_{k} n_k \). For this four category-matching scheme, (2.4.2.) becomes,

\[
E \left[ \frac{(N^\omega - 1)!}{(N-1)!} \right] = \sum_{n=M}^{\infty} \frac{(n-\omega + 1)!}{(n-1)!} \sum_{k=1}^{n} \binom{n-1}{n_k} p_k \left( 1 - p_k \right)
\]

\[
\sum_{j=1}^{c} \sum_{i=n_k}^{n} \frac{(n-n_k)!}{(n-k)!} \frac{r}{s_j} \frac{p_j}{1-p_k} \]

\[
\sum_{j=1}^{r} \sum_{i=n-k}^{n} \frac{r}{s_j} \frac{p_j}{1-p_k} \]

\[
\sum_{j=1}^{r} \sum_{i=n-k}^{n} \frac{r}{s_j} \frac{p_j}{1-p_k} \]
\[
\sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \sum_{k=1}^{4} E_4(p_k, p_{i+k}, n-n_k),
\]

\[
= \sum_{k=1}^{4} \sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} E_4(p_k, p_{i+k}, n-n_k)
\]

where \(p_{i+k}\) denotes the vector of probabilities of every category but the \(k\)-th.

A particular case of \(E_4(p_k, p_{i+k}, n-n_k)\) would be when \(k=4\); then,

\[
E_4(p_k, p_{i+k}, n-n_k) = E_4(p_4, p_1, p_2, p_3, n-n_4). \quad \text{We will find the value of}
\]

\[
\sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} E_4(p_4, p_1, p_2, p_3, n-n_4).
\]

The other \(E_4\)'s can be found by appropriate permutation of indices.

Now

\[
E_4(p_4, p_1, p_2, p_3, n-n_4) = \left(\frac{n-1}{n_4-1}\right)^{p_4} (1-p_4) \sum_{s_1=n_1}^{n-n_4} \frac{n-M+n_1}{\sum_{s_2=n_2}^{s_1+n_2+n-M-s_1}} \frac{(n-n_4)!}{s_1!s_2!(n-n_4-s_1-s_2)!} \frac{p_1 p_2 p_3}{(1-p_4)}
\]

\[
= \left(\frac{n-1}{n_4-1}\right)^{p_4} \sum_{s_1=n_1}^{n_1+n-M} \sum_{s_2=n_2}^{n_2+n-M-(s_1-n_1)} \frac{(n-n_4)!}{s_1!s_2!(n-n_4-s_1-s_2)!} \frac{p_1 p_2 p_3}{(1-p_4)}
\]

\[
= \left(\frac{n-1}{n_4-1}\right)^{p_4} \frac{s_1 s_2 n-n_4-s_1-s_2}{p_1 p_2 p_3}
\]
\[ \left( \frac{n-1}{n_4-1} \right) \left( \frac{n}{n_4} \right) ^{n+M-1} \sum_{s_1} \left( \frac{s_1}{n-n_4} \right) \left( \frac{s_2}{p_2+p_3} \right) \left( \frac{s_3}{p_2+p_3} \right) \left( \frac{s_4}{p_2+p_3} \right) \]
This last step follows by noting that \( s_3 = (n-n_4-s_1-s_2) \) and then using the third relationship that we stated earlier. If we bring in the exterior summation over \( s_1 \) into the interior summations we then have

\[
\binom{n-1}{n_4-1} p_4 \sum \binom{n_1+n-M}{s_1} \binom{n-n_4-s_1}{s_1} p_1 (p_2+p_3) \]

\[
\sum_{s_2=0}^{n_3-1} \binom{n_1+n-M}{s_1} \binom{n-n_4-s_1}{s_2} s_1 s_2 (n-n_4-s_1)^{-s_3} p_1 p_2 p_3
\]

(A.1.1.)

We will now examine this expression term by term, Firstly,

\[
\binom{n-1}{n_4-1} p_4 \sum \binom{n_1+n-M}{s_1} \binom{n-n_4-s_1}{s_1} p_1 (p_2+p_3)
\]

\[
= \binom{n-1}{n_4-1} p_4 \left\{ \binom{n_1+n-M}{s_1} \binom{n-n_4-s_1}{s_1} p_1 (p_2+p_3) \right\}
\]

\[
- \sum_{s_1=0}^{n_1-1} \binom{n_1-n_4}{s_1} s_1 (n-n_4-s_1)^{-s_1} p_1 (p_2+p_3)
\]
\[
= \binom{n-1}{n_4-1} p_4^n \left\{ \binom{n-n_4}{s_1} p_1^{s_1} (p_2+p_3)^{n-n_4-s_1} \right. \\
- \sum_{s_1=n_1+n-M+1}^{n-n_4} \binom{n-n_4}{s_1} p_1^{s_1} (p_2+p_3)^{n-n_4-s_1} \\
- \sum_{s_1=0}^{n_1-1} \binom{n-n_4}{s_1} p_1^{s_1} (p_2+p_3)^{n-n_4-s_1} \right\}
\]

In the second summation, just above, let \( i = (n-n_4)-s_1 \). Then the first term of (A.1.1.) becomes

\[
\binom{n-1}{n_4-1} p_4^n \left\{ \binom{n-n_4}{s_1} p_1^{s_1} (p_2+p_3)^{n-n_4-s_1} \right. \\
- \sum_{i=0}^{n_2+n_3-1} \binom{n-n_4}{i} (p_2+p_3)^i p_1^{n-n_4-i} \\
- \sum_{s_1=0}^{n_1-1} \binom{n-n_4}{s_1} p_1^{s_1} (p_2+p_3)^{n-n_4-s_1} \right\}
\]

If we sum this expression over \( \sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \), we have the following results for each of the three terms: First,

\[
= \sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-n_4)! (n_4-1)!} p_4^n (1-p_4)^{n-n_4}
\]

\[
\begin{align*}
\frac{n_4}{(n_4-1)!} \cdot \frac{(n_4+\ell-1)!}{p_4} \sum_{n=M}^{\infty} \left( \frac{n+\ell-1}{n_4+\ell-1} \right) p_4^{n_4+\ell-1} (1-p_4)^{n-n_4} \\
= \frac{n_4}{(n_4-1)!} \cdot \frac{(n_4+\ell-1)!}{p_4} I (1-p_4)^{(M-n_4,n_4+\ell)}
\end{align*}
\]

from the first relation we stated. Next,

\[
\begin{align*}
\sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \left( \frac{n-l}{n_4-1} \right) \sum_{i=0}^{r_n-1} \left( \frac{n-n_4}{p_2+p_3} \right)^i (p_1)^{n-n_4-i}
\end{align*}
\]

\[
\begin{align*}
= \frac{n_4}{(n_4-1)!} \sum_{i=0}^{n_2+n_3-1} \frac{(p_2+p_3)^i}{i!} \cdot \frac{(n_4+\ell+i-1)!}{(1-p_1)^{n_4+\ell+i-1}}
\end{align*}
\]

\[
\begin{align*}
\sum_{n=M}^{\infty} \left( \frac{n+\ell-1}{n_4+\ell+i-1} \right) (1-p_1)^{n_4+\ell+i-1} p_1^{n-n_4-i}
\end{align*}
\]

\[
\begin{align*}
= \frac{n_4}{(n_4-1)!} \sum_{i=0}^{n_2+n_3-1} \frac{(p_2+p_3)^i}{i!} \frac{(n_4+\ell+i-1)!}{1-p_1} p_1^{n-n_4-i, n_4+\ell+i}
\end{align*}
\]

Similarly,

\[
\begin{align*}
\sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \left( \frac{n-1}{n_4-1} \right) \sum_{s_1=0}^{n_1=1} \left( \frac{n-n_4}{s_1} \right)^{s_1} p_1 \left( p_2+p_3 \right)^{n-n_4-s_1}
\end{align*}
\]
\[
\frac{n_4}{p_4} \frac{p_1}{(n_4-1)!} \sum_{s_1=0}^{n_4-1} \frac{s_1(n_4+s_1-1)!}{s_1!} \frac{n_4+s_1-1}{p_1} \frac{I(p_2+p_3)^{(M-n_4-s_1,n_4+s_1)}}{I(p_2+p_3)^{(M-n_4-s_1,n_4+s_1)}}.
\]

Thus, the sum over the first term of (A.1.1.) is

\[
\frac{n_4}{p_4} \frac{p_1}{(n_4-1)!} \left\{ \frac{(n_4+\ell-1)!}{p_4} I(1-p_4)^{(M-n_4,n_4+\ell)} \right. \\
- \sum_{i=0}^{n_2+n_3-1} \frac{(p_2+p_3)^{i}(n_4+\ell+i-1)!}{i!(1-p_1)} I(p_1)^{(M-n_4-i,n_4+\ell+i)} \\
- \sum_{s_1=0}^{n_1-1} \frac{s_1(n_4+\ell+s_1-1)!}{s_1!(p_1+p_4)} \frac{n_4+\ell+s_1}{p_1} \frac{I(p_2+p_3)^{(M-n_4-s_1,n_4+\ell+s_1)}}{I(p_2+p_3)^{(M-n_4-s_1,n_4+\ell+s_1)}} \right\}. \quad (A.1.2.)
\]

Now, consider the second term of (A.1.1.):

\[
\sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1+n-M} \frac{1}{p_1} \frac{1}{p_3} \frac{1}{p_2} \frac{(n-n_4)(n_n-n_4-s_1)}{s_1} \frac{s_1}{s_2} \frac{(n-n_4-s_1-s_3)}{s_3} p_1 p_3 p_2
\]

\[
= \left\{ \sum_{s_3=0}^{n_3-1} \frac{(n-n_4)}{s_3} p_3 \sum_{s_1=0}^{n_1+n-M} \frac{(n-n_4-s_3)}{s_1} p_1 p_2 \right\}
\]

\[
= \sum_{s_3=0}^{n_3-1} \frac{(n-n_4)}{s_3} p_3 \left\{ \frac{n-n_4-s_3}{n_2+n_3-1} \sum_{i=0}^{n_4-s_3} \frac{(n-n_4-s_3-i)}{i} p_2 p_1 \right\}
\]
\[ - \sum_{s_1=0}^{n_1-1} \binom{n-n_4-s_3}{s_1} s_1 \left( n-n_4-s_3 \right)^{s_1} \left( n-n_4-s_3 \right)^{-s_1} \}
\]

We will now sum each term of this expression over \[ \sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \left( \frac{n-1}{n_4-1} \right)^p_4 \]

First

\[ \sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \left( \frac{n-1}{n_4-1} \right)^n_4 \sum_{s_3=0}^{n_3-1} \binom{n-n_4}{s_3} p_3^s_3 \left( p_1+p_2 \right) \left( n-n_4-s_3 \right) \]

\[ = \frac{n_4}{p_4} \sum_{s_3=0}^{n_3-1} \frac{s_3 \left( n_4+\ell+s_3 \right)!}{s_3! \left( p_3+p_4 \right)} \left( n_4+\ell+s_3 \right)^{-1} \left( p_1+p_2 \right) \left( n-n_4 \right) \]

\[ = \frac{n_4}{p_4} \sum_{s_3=0}^{n_3-1} \frac{s_3 \left( n_4+\ell+s_3 \right)!}{s_3! \left( p_3+p_4 \right)} \left( n_4+\ell+s_3 \right)^{-1} \left( p_1+p_2 \right) \left( n-n_4 \right) \]

\[ I_{\left( p_1+p_2 \right)} \left( M-n_4-s_3, n_4+s_3+\ell \right) \]

Secondly,

\[ \sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \left( \frac{n-1}{n_4-1} \right)^n_4 \sum_{s_3=0}^{n_3-1} \sum_{i=0}^{n_2+n_3-s_3-1} \binom{n-n_4}{s_3} \]

\[ \left( n-n_4-s_3 \right)^{s_j} i \left( n-n_4-s_3 \right)^{-i} \left( p_3+p_2 \right) \left( p_1 \right) \]
\[
\begin{align*}
\frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{i=0}^{n_2+n_3-s_3-1} & \frac{p_1^n p_3^{s_3(n_4+i+s_3-1)!}}{i! s_3!(p_2+p_3+p_4)^{n_4+i+s_3-1}} \\
\sum_{n=M}^{n_4} & \binom{n+l-1}{n_4+l+s_3+i-1}(1-p_1)^{n_4+l+s_3+i-1} p_1^{n-n_4-s_3-1} \\
& = \frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{i=0}^{n_2+n_3-s_3-1} \frac{p_1^n p_3^{s_3(n_4+i+s_3-1)!}}{i! s_3!(1-p_1)^{n_4+l+s_3+i}} \\
\end{align*}
\]

Finally, by following the immediately preceding argument, we have

\[
\sum_{n=M}^{n_4} \frac{(n+l-1)!}{(n-1)!} \binom{n-1}{n_4-1} \frac{n_4^{n_3-1} n_1^{-1}}{p_4} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \binom{n-n_4}{s_3}^s \binom{n-n_4-s_3}{s_1}^s p_3 \binom{n-n_4-s_3}{s_1}^s p_1 p_2 \\
= \frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \frac{p_1^n p_3^{s_3(n_4+i+s_1+s_3-1)!}}{s_1! s_3!(1-p_2)^{n_4+i+s_1+s_3}} \\
\]

\[
\frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \frac{p_1^n p_3^{s_3(n_4+i+s_1+s_3-1)!}}{s_1! s_3!(1-p_2)^{n_4+i+s_1+s_3}} \\
\]

\[
\frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \frac{p_1^n p_3^{s_3(n_4+i+s_1+s_3-1)!}}{s_1! s_3!(1-p_2)^{n_4+i+s_1+s_3}} \\
\]

\[
\frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \frac{p_1^n p_3^{s_3(n_4+i+s_1+s_3-1)!}}{s_1! s_3!(1-p_2)^{n_4+i+s_1+s_3}} \\
\]

\[
\frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \frac{p_1^n p_3^{s_3(n_4+i+s_1+s_3-1)!}}{s_1! s_3!(1-p_2)^{n_4+i+s_1+s_3}} \\
\]

\[
\frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \frac{p_1^n p_3^{s_3(n_4+i+s_1+s_3-1)!}}{s_1! s_3!(1-p_2)^{n_4+i+s_1+s_3}} \\
\]

\[
\frac{n_4}{(n_4-1)!} \sum_{s_3=0}^{n_3-1} \sum_{s_1=0}^{n_1-1} \frac{p_1^n p_3^{s_3(n_4+i+s_1+s_3-1)!}}{s_1! s_3!(1-p_2)^{n_4+i+s_1+s_3}} \\
\]
Thus, when we sum the second summation term in (A.1.1.) over

\[ \sum_{n=M}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} \frac{(n-1)}{(n_4-1)} p_4^n, \]

it becomes

\[ \frac{p_4^{n_4}}{(n_4-1)!} \left\{ \sum_{s_3=0}^{n_3-1} \frac{p_3^{s_3} (n_4+s_3-1)!}{s_3! (p_3 p_4)} \right\} \sum_{i=0}^{s_3} \frac{p_2^{i} p_3^{(n_4^2+s_3^2+i-1)!}}{n_4^2+s_3^2+i! s_3! (1-p_1)} \]

\[ I_{p_1} (M-n_4-s_3-i, n_4+s_3+i) \]

Now, if we examine the third summation term in (A.1.1.), we can see

that it differs from the second summation term only in that \( n_2 \) is

interchanged with \( n_3 \) and \( p_2 \) is interchanged with \( p_3 \). Thus, we can see

that the sum of the third term in (A.1.1.) over the same factor as the

sum of the second term becomes

\[ \left\{ \sum_{s_3=0}^{n_3-1} \frac{p_1^i p_3^{s_3} (n_4^2+s_3^2+i-1)!}{n_4^2+s_3^2+i! s_3! (1-p_2)} \right\} \]

\[ I_{p_2} (M-n_4-s_3-i, n_4+s_3+i) \]
\[
\frac{p_{4n}}{(n_{4}+1)} \sum_{s_{2}=0}^{n_{2}-1} \frac{s_{2}^{2}}{s_{2}} \frac{(n_{4}+\ell+s_{2}+1)}{n_{4}+\ell+s_{2}} \frac{I_{(p_{4}+p_{3})}(M-n_{4}-s_{2},n_{4}+s_{2}+\ell)}{s_{2}!(p_{3}+p_{4})} \\
- \sum_{s_{3}=0}^{n_{2}-1} \sum_{i=0}^{n_{2}+n_{3}-s_{2}-1} \frac{s_{2}}{p_{2}p_{3}} \frac{(n_{4}+\ell+s_{2}+i-1)!}{n_{4}+\ell+s_{2}+i} s_{2}!i!(1-p_{1}) \\
I_{p_{1}} (M-n_{4}-s_{2}-i,n_{4}+s_{2}+\ell) \\
- \sum_{s_{2}=0}^{n_{2}-1} \sum_{s_{1}=0}^{s_{1}-1} \frac{s_{1}s_{2}}{p_{1}p_{2}} \frac{(n_{4}+\ell+s_{1}+s_{2}-1)!}{n_{4}+\ell+s_{1}+s_{2}} s_{1}!s_{2}!(1-p_{3}) \\
I_{p_{3}} (M-n_{4}-s_{1}-s_{2},n_{4}+s_{1}+s_{2}+\ell)
\]

(A.1.4.)

Now, if we combine all the terms in (A.1.2.), (A.1.3.), and (A.1.4.), we will have found the value of

\[
\sum_{n=0}^{\infty} \frac{(n+\ell-1)!}{(n-1)!} E_{4}(p_{4},p_{1},p_{2},p_{3},n-n_{4}) \equiv F_{4}(p_{4},p_{1},p_{2},p_{3},M-n_{4})
\]

Before we combine the three expressions, we will define the following notation: let

\[
SS(n^{*},p^{*},p_{4},M,n_{4}) = \sum_{i=0}^{n^{*}} \frac{p^{*i}(n_{4}+\ell+i-1)}{n_{4}+\ell+i} I_{(1-p^{*}-p_{4})}(M-n_{4}-i,n_{4}+\ell+i) \frac{1}{i!(p^{*}+p_{4})}
\]

and let
\[ DS(n_a, n_b, p_a, p_b, p_4, M, n_4, ) \]

\[ = \sum_{i=0}^{n_a} \sum_{j=0}^{n_b-i} \frac{p_a^i p_b^j (n_4 + \ell + i + j - 1)!}{i! j! (p_a + p_b + p_4)^{n_4 + \ell + i + j}} \]

\[ I(1 - [p_a + p_b + p_4])^{(M - n_4 - i - j, n_4 + \ell + i + j)} . \]

If we use this notation to combine the terms in (A.1.2.), (A.1.3.), and (A.1.4.), we get

\[ F_4(p_4, p_1, p_2, p_3, n_4) = \frac{p_4^{n_4}}{(n_4 - 1)!} \left\{ \frac{(n_4 + \ell - 1)!}{p_4^{n_4 + \ell}} I(1 - p_4)^{(M - n_4, n_4 + \ell)} \right\} \]

- SS(n_1 - 1, p_1, p_4, M, n_4, \ell) - SS(n_2 - 1, p_2, p_4, M, n_4, \ell)

- SS(n_3 - 1, p_3, p_4, M, n_4, \ell) - SS(n_2 + n_3 - 1, p_2 + p_3, p_4, M, n_4, \ell)

+ DS(n_3 - 1, n_2 + n_3 - 1, p_3, p_2, p_4, M, n_4, \ell)

+ DS(n_2 - 1, n_2 + n_3 - 1, p_2, p_3, p_4, M, n_4, \ell)

+ DS(n_1 - 1, n_1 - 1, p_3, p_1, p_4, M, n_4, \ell)

+ DS(n_2 - 1, n_1 - 1, p_2, p_1, p_4, M, n_4, \ell) \}

(A.1.5.)

Finally,
$$E \frac{(n+\ell-1)!}{(n-1)!} = F_4(p_1, p_2, p_3, p_4, M-n_1, \ell) + F_4(p_2, p_3, p_4, p_1, M-n_2, \ell)$$

$$+ F_4(p_3, p_4, p_1, p_2, M-n_3, \ell) + F_4(p_4, p_1, p_2, p_3, M-n_4, \ell). \quad (A.1.6.)$$

Note that if we set any $p_k = 0$ and $n_k = 0$, we get the three category model as a special case of the four category model. This is assuming that terms involving $\sum_{i=0}^{1} a_i$ are interpreted to be zero and that $F_4(0, \ldots)$ is defined to be zero. For example, if we take $p_4 = n_4 = 0$, then (A.1.6.) becomes

$$E \frac{(N+\ell-1)!}{(N-1)!} = F_4(p_1, p_2, p_3, 0, M-n_1, \ell) + F_4(p_2, p_3, 0, p_1, M-n_2, \ell)$$

$$+ F_4(p_3, 0, p_1, p_2, M-n_3, \ell).$$

These three $F_4(\ldots)$ functions are each composed of a sum of $SS(\ldots)$ functions and $DS(\ldots)$ functions. To show how the three category model falls out of the four category case, examine $F_4(p_3, 0, p_1, p_2, M-n_3, \ell)$. From (A.1.5.), with $p_4 \rightarrow p_3$, $p_1 \rightarrow p_4$, $p_2 \rightarrow p_1$, $p_3 \rightarrow p_2$, and similarly for the $n_k$'s, we have

$$F_4(p_3, 0, p_1, p_2, M-n_3, \ell) = \frac{n_3^{p_3}}{(n_3-1)!} \left\{ \begin{array}{c}
\frac{(n_3+\ell-1)!}{n_3^{p_3+1}} \left(1-p_3\right)^{M-n_3-n_3+\ell} \\
- SS(-1, 0, p_3, M, n_3, \ell) - SS(n_1^{-1}, p_1, p_3, M, n_3, \ell) - SS(n_2^{-1}, p_2, p_3, M, n_3, \ell) \\
- SS(n_1^{-1}, n_2^{-1}, p_1, p_2, p_3, M, n_3, \ell) + DS(n_2^{-1}, n_2^{-1}, p_2, p_1, p_3, M, n_3, \ell) \\
+ DS(n_1^{-1}, n_2^{-1}, p_1, p_2, p_3, M, n_3, \ell) + DS(n_2^{-1}, p_2, 0, p_3, M, n_3, \ell)
\end{array} \right\}$$
\[ + DS(n_1-1, -1, p_1, 0, p_3, M, n_3, \ell) \] (A.1.7.)

The first SS function and the last two DS functions are all zero since they involve a summation of the form \( \sum_{i=0}^{n_1-1} \) which is defined to be zero. Now the definition at the incomplete beta function ratio is

\[ I_{p}(a, N, a+1) = \sum_{i=a}^{N} p^i (1-p)^{N-i} \]

If \( p=0 \), then \( I_{p}(a, N, a+1) = 0 \) if \( a > 0 \). The functions SS \((n_1+n_2-1, p_1+p_2, \ldots)\), DS\((n_2-1, n_2+n_1-1, p_2 p_1, \ldots)\), and DS\((n_1-1, n_1+n_2-1, p_1, p_2, \ldots)\) are all zero, because the probability \( p \) in their respective incomplete beta function ratios is equal to \( (1-(p_1+p_2+p_3)) = p_4 = 0 \). When we drop the zero expressions from (A.1.7.), it becomes

\[ F_x(p_3, 0, p_1, p_2, M-n_3, \ell) = \frac{p_3^{n_3}}{(n_3-1)!} \left\{ \frac{(n_3+\ell-1)!}{n_3+1} I_{p_3}(M-n_3, n_3+\ell) \right. \]
\[ - SS(n_1-1, p_1, p_3, M, n_3, \ell) - SS(n_2-1, p_2, p_3, M, n_3, \ell) \]
\[ = \frac{n_3}{p_3^{n_3}} \left\{ \frac{(n_3+\ell-1)!}{n_3+1} I_{p_3}(M-n_3, n_3+\ell) \right. \]
\[ - \sum_{i=0}^{n_1-1} \frac{p_1^i (n_3+\ell+i-1)!}{i! (p_1+p_3)^{n_3+\ell+1}} I_{p_2}(M-n_3-1, n_3+\ell+i) \]
\[ - \sum_{i=0}^{n_2-1} \frac{s_2^i (n_3+\ell+i-1)!}{i! (p_2+p_3)^{n_3+\ell+1}} I_{p_1}(m-n_3-1, n_3+\ell+i) \]
\[ \equiv F_3(p_3, p_1, p_2, M-n_3, \lambda) \]  \hspace{1cm} (A.1.8.)

So, for any three category-matching scheme, we have

\[ E \left\{ \frac{(N+\lambda-1)!}{(N-1)!} \right\} = F_3(p_1, p_2, p_3, M-n, \lambda) + F_3(p_2, p_3, p_1, M-n_2, \lambda) \]

\[ + F_3(p_3, p_1, p_2, M-n_3, \lambda) \]  \hspace{1cm} (A.1.9.)

This expression is identical to the expression that can be derived by starting with (2.4.2.) for a three category-matching scheme and performing manipulations similar to those used above.
REFERENCES


REFERENCFS continued...


REFERENCES continued...

Smith, A.E., (1976), "Minimum Distance Case-Control Matching," unpublished manuscript.