1

PM-1 3 1/2" x 4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Vous trouverez ci-dessous :

Chaque notice est précédée de :

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

**Statistical Methods to Overcome Nonindependence
of Coupled Data in Significance Testing**

by

Bruno Domenico Zumbo, B.Sc., M.A.

**A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of**

**Doctor of Philosophy
Department of Psychology**

Carleton University

Ottawa, Ontario

December, 1992

© copyright

1992, Bruno Domenico Zumbo

**National Library
of Canada**

**Acquisitions and
Bibliographic Services Branch**

**395 Wellington Street
Ottawa, Ontario
K1A 0N4**

**Bibliothèque nationale
du Canada**

**Direction des acquisitions et
des services bibliographiques**

**395, rue Wellington
Ottawa (Ontario)
K1A 0N4**

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-84031-5

Canada

The undersigned hereby recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis,

"Statistical Methods to Overcome Nonindependence
of Coupled Data in Significance Testing"

submitted by

Bruno Domenico Zumbo, B.Sc., M.A.

in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

Bill Jones, Ph.D.

Chair, Department of Psychology

Donald W. Zimmerman, Ph.D.

Thesis Supervisor

Jane F. Gentleman, Ph.D.

External Examiner

Carleton University

ABSTRACT

The normal statistical methods used in the social and behavioral sciences assume independent, identically distributed random variables. In some experimental or quasi-experimental settings, the assumption of independence within groups is violated by the experimental procedure. For example, in perceptual or cognitive research, lateral measures (e.g., measures on both ears) may be obtained for each subject. If there is no interest in the differential effect of the lateral measures and each is treated as though independent, one will violate the above assumption, as the two measures are indeed correlated (Coren & Hakstian, 1990). These correlated or, as I will refer to them, coupled data create a problem only in the context of statistical inference, as they will result in an inordinate increase in the type I error rate when analyzed using commonly-used significance tests.

The purpose of the present dissertation was to develop and explore alternative methods for data analysis when the existence of coupled data violated the assumption of independence. This was done in three main parts: (a) the examination and development of alternative methods arising from statistical and psychological research, (b) the derivation of correction formulas, and (c) an investigation of the statistical properties of the above methods using a series of Monte Carlo studies.

The results of the mathematical analysis and Monte Carlo simulation would suggest that researchers use the statistical methods derived in Chapter 3. That is, the Chapter 3 statistics (a) maintain their level, (b) are as powerful if not more powerful than some alternatives, (c) can be used in a meta-analytic setting, and (d) fit into the mold of commonly used statistical techniques in psychology.

Acknowledgements

I would like to thank Dr. Donald W. Zimmerman for all of his support and patience. My years under his guidance and tutelage have been wonderful! I would also like to thank Dr. Richard H. Williams who in the last few years has also been a great support, friend, and mentor. In addition, I would like to acknowledge the tremendous support of Dr. Roland Thomas and Dr. Jane F. Gentleman. Roland, in particular, has been a good friend and a wonderful sounding-board for my often crazed ideas in statistics and psychometrics. A special note of gratitude to Dr. Bob Hoffmann and Dr. Jo Tombaugh who, over the years I have been at Carleton, have been exceptionally kind and giving to me.

As a note, I would like to acknowledge my thesis examiners: Drs. Jane Gentleman, Roland Thomas, Bob Hoffmann, Bill Petrusic, Bill Jones, & Don Zimmerman for clarifying and improving the dissertation through their queries.

On a personal level, I would like to thank my wife Anita Hubley who has been a wonderful support both academically and personally. Without her and the support of wonderful friends like Doreen Zimmerman, Allison Kulak, Ron McMorran, Leslie Boucher, and Elliot Loh this dissertation would never have been completed. A special note to Tim Pychyl who has always been there for me. Finally, I would like to thank my family for their support and encouragement throughout my studies. Look Ma, I did it!

Table of Contents

ABSTRACT	ii
Acknowledgements	iii
Table of Contents.....	iv
List of Tables	vi
List of Figures	vii
INTRODUCTION	1
CHAPTER 2	
ALTERNATIVE METHODS.....	7
2.1 Approaches arising from the statistical literature	7
2.1.1 Approach based on survey sampling.....	8
2.1.2 Approach based on mixed effects models	13
2.1.3 Approach based on generalized least squares	14
2.2 Approaches based on the psychological literature.....	24
2.2.1 Coren and Hakstian's approach	24
2.2.2 Approach based on randomization tests	26
2.2.3 Approach based on nonparametric rank tests	31
2.3 Summary	32
CHAPTER 3	
DERIVATION OF CORRECTION FORMULAS	34
3.1 Fundamental equations.....	34
3.2 One-group case	38
3.3 Two-group case.....	39
3.4 k-group case.....	41
3.5 Analysis.....	44
3.5.1 Mathematical results of the type I error rate.....	45
CHAPTER 4	
MONTE CARLO INVESTIGATION	51
4.1 Purpose	51
4.2 Assumptions and Limitations	52
4.3 Methodology	53
4.3.1 Computer simulation methodology	54
4.3.2 Performance indices	55

4.3.2 Study 1	57
4.3.3 Study 2	58
4.3.4 Study 3	59
CHAPTER 5	
RESULTS & DISCUSSION	61
5.1 Study 1	61
5.2 Study 2	74
5.3 Study 3	84
5.4 Summary of the findings.	86
CHAPTER 6	
EXAMPLES	88
Example 1	
Fictitious data.	88
Example 2	
Effect of sleep deprivation on cognitive performance.....	90
CHAPTER 7	
CONCLUSIONS.....	92
REFERENCES.....	95

List of Tables

Table 1. Type I error rates of the as a function of the correlation among the elements of the couple, $\alpha = .05$.	48
Table 2. Type I error rates for the t-test for various values of alpha.	61
Table 3. Type I error rates for the F-test for various values of alpha.	62
Table 4. Type I error rates and power of the two independent samples t test, $\alpha = .05$.	79
Table 5. Type I error rates and power of the F test for $k=4$, $n=12$, $\alpha=.05$.	81
Table 6. Type I error rates and power of the F test for $k=4$, $n=30$, $\alpha=.05$.	82
Table 7. Type I error rates and power of the F test for $k=3$, $n=30$, $\alpha=.01$.	83
Table 8. The empirical type I error rate for the randomization and couplet referenced randomization test at various value of the within couple correlation, $\alpha=.05$.	85
Table 9. Type I error rates and power of the t' and couplet-randomization tests, $\alpha=.05$.	86
Table 10. Data for Example 1.	89

List of Figures

Figure 1. Coupled data example.	4
Figure 2. Variance-covariance matrices for independent and coupled data.	16
Figure 3. Example of Standard and Couplet-Referenced Randomization Tests, and the Randomization version of the Coren and Hakstian Procedure.....	30
Figure 4. Type I error rates of the as a function of the correlation among the elements of the couple, $n = 4$ and 10,000.	49
Figure 5. Power functions of Student's t-test for various values of correlation ($n=8$, $\alpha=.01$).	62
Figure 6. Power functions of Student's t-test for various values of correlation ($n=8$, $\alpha=.05$).	65
Figure 7. Power functions of Student's t-test for various values of correlation ($n=8$, $\alpha=.10$).	66
Figure 8. Power functions of Student's t-test for various values of correlation ($n=32$, $\alpha=.01$).	67
Figure 9. Power functions of Student's t-test for various values of correlation ($n=32$, $\alpha=.05$).	68
Figure 10. Power functions of Student's t-test for various values of correlation ($n=32$, $\alpha=.10$).	69
Figure 11. Power functions of Student's t-test for various values of correlation ($n=500$, $\alpha=.05$).	70
Figure 12. Power functions for the F-test for various values of correlation ($k=4$, $n=30$, $\alpha=.05$).	71
Figure 13. Power functions for the F-test for various values of correlation ($k=3$, $n=30$, $\alpha=.01$).	72
Figure 14. Power functions for the F-test for various values of correlation ($k=4$, $n=12$, $\alpha=.05$).	73
Figure 15. Type I error rates of the t-test as a function of the within couplet correlation ($n=12$).	75
Figure 16. Type I error rates of the t-test as a function of the within couplet correlation ($n=20$).	76
Figure 17. Type I error rate of the F-test as a function of the within couplet correlation ($k=3$, $n=32$).	77

Figure 18. Type I error rates of the F-test as a function of the within couplet correlation ($k=3$, $n=16$).....	78
---	-----------

INTRODUCTION

Frequently, research in psychology (or more generally, the social and behavioral sciences) makes use of normal theory statistical methods. These are the standard realizations of the general linear model, such as regression, analysis of variance (ANOVA), and Student's t-test. These normal theory methods are founded on two basic assumptions: (a) a normally distributed (i.e., Gaussian or bell-shaped) underlying distribution, and (b) sample observations that are independently and identically distributed. That is, considering the data for one group of individuals, each datum is thought of as a random variable and each of these random variables is independent (i.e., unrelated) to the other and identically distributed. What this implies is that each datum comes from a normal distribution (specified by a mean and variance parameter or location-shift parameter) and that each datum is probabilistically independent [i.e., $\text{Prob}(X_2 | X_1) = \text{Prob}(X_2)$].

How are these assumptions realized? First, normality is typically assumed. Next, independence is assured by a proper randomization mechanism such as random assignment and/or random sampling (or probability sampling). There are various modern developments in statistics, such as robust estimation and nonparametric rank statistics, which deal with violations of normality (for a review, see Lind & Zumbo, in press; Zimmerman & Zumbo, 1993). As well, several studies have shown that independence is crucial for the F-test (Cochran, 1947; Lana & Lubin, 1963; Scariana & Davenport, 1987; Scheffé, 1959; Zimmerman & Zumbo, in press a). Recently, Kenny & Judd (1986) have presented several scenarios from psychological research wherein independence could be violated. They focused on the impact of autocorrelation on the bias of the mean square error and its eventual inflation of type I error rate during statistical significance testing.

The purpose of the present study is to develop and explore alternative methods for data analysis when the assumption of independence is violated by paired or coupled data. This development involves two interrelated parts to this study: (a) an analytical (mathematical) derivation of some possible solutions, and (b) a Monte Carlo investigation of some of the proposed alternatives. The primary purpose of this dissertation is to present (in Chapter 3) an alternative way of dealing with the problem of paired or coupled data within the context of traditional psychological approaches to data analysis (i.e., least squares methods such as ANOVA and t-tests).

The formulas found in Chapter 3 were derived with the objective of making the corrections a generalization of normal theory statistics. The advantage to this method is two-fold. First, these formulas easily can be implemented by correcting results that come from commonly used programs such as SPSSX[®], BMDP[®], SAS[®], or a host of many other programs on the market. Secondly, many years of research do not have to be discarded. That is, with the new formulas in hand one can examine the literature in a meta-analytic context, and the reported F and t statistics can be corrected to see if the conclusions are altered.

At this juncture a more formal description of the new term 'coupled data' is necessary. Coupled data situations are those cases in which the observations in a study are not independent random variables, but rather are pairwise related. Coupled data situations are best described with an example from the area of auditory perception, which, in fact, was our motivation for studying this problem. Coren and Hakstian (1990) recently identified an important statistical problem that arises in research that assesses the efficiency of limited-frequency testing and screening for auditory sensitivity (see, for example, Wiley, Oviatt, & Block, 1987). They stated that until recently researchers would randomly assign subjects to groups and then obtain measurements on both ears. The data then were

analyzed at the "ear" level, effectively doubling the sample size per group. The standard statistical methods used, of course, assumed that the ears were independent. First, Coren and Hakstian established empirically that there is a substantial non-zero correlation between auditory sensitivities of the two ears, sometimes as high as .80, and demonstrated that the correlation is not an artifact of bilateral hearing loss occurring with age. They then noted that the common practice of treating measures of a subject's individual ears as independent, when in fact they are not, can lead to serious inflation of the type I error rate during statistical inference. Given that our motivation was the Coren and Hakstian two-ear problem, wherein they treated continuous observations, the focus of this paper will be confined to continuous data as well (e.g., tests involving t and F distributions).

To further clarify the problem, an example of the structure of coupled data from a two-group completely randomized design is given in Figure 1. In terms of the problem in Coren and Hakstian, this example has group 1, a control group consisting of six subjects, and group 2, an experimental group consisting of six subjects. The researchers are not interested in between ear differences; therefore, they have twelve observations per group. Traditionally, in the auditory perception literature, this has been envisioned as a two-group completely randomized design with twelve observations per group. It is usually analyzed with an independent samples t -test to test the hypothesis of a mean difference between groups. As Coren and Hakstian noted, this is incorrect because in this case they are considering the two ears independent when in fact they are not.

The same problem as Coren and Hakstian's two-ear scenario arises in visual research that treats the measures obtained on a subject's two eyes as independent (Ederer, 1973). Other examples could conceivably occur in the problem-solving literature where two individuals work in collaboration, yet respond independently, or in the area of human performance where the pair of hands, eyes, or feet are the subject of investigation. Furthermore, coupled data can result from nonindependent sampling. The sampling

Figure 1. Coupled data example.

Group 1		Group 2	
<u>subj. #</u>	<u>X</u>	<u>subj. #</u>	<u>X</u>
1	X ₁	7	X ₁₃
1	X ₂	7	X ₁₄
2	X ₃	8	X ₁₅
2	X ₄	8	X ₁₆
3	X ₅	9	X ₁₇
3	X ₆	9	X ₁₈
4	X ₇	10	X ₁₉
4	X ₈	10	X ₂₀
5	X ₉	11	X ₂₁
5	X ₁₀	11	X ₂₂
6	X ₁₁	12	X ₂₃
6	X ₁₂	12	X ₂₄

procedure selects pairs of subjects, rather than individual subjects. For example, a sample of married couples is obtained and measurements are obtained on both partners. The general scenario wherein an implicit pairing occurs is termed the "coupled data" situation. The phrase paired-data situation is avoided due to its possible confusion with the within-subject design. It is important to note that in the coupled data situation being considered here, the researcher usually has no interest in studying the difference between the elements of the couple. For example, in the case of the two-ear problem the researcher is usually not interested in testing whether the treatment had a differential effect on the ears¹.

¹ Since the researcher is not interested in differential ear effects the natural option of conducting a mixed model (with between and within factors) is not used. One could potentially ignore the within groups tests; however, this results in an unnecessary loss of $k - 1$ degrees of freedom, where k is the total number of within group levels. Along with this loss of degrees of freedom comes a loss in statistical power.

As is noted by Coren and Hakstian (1990), little has been published regarding statistical inference with coupled data. Therefore, Chapter 2 will present several alternative methods of treating this situation. Some of these methods involve newly derived procedures while others are an application of known methodology. Examples of some of the known methodologies are: (1) an application of inference with complex sampling designs (for a review, see Thomas, in press), (2) the use of a mixed effects ANOVA, and (3) a solution based on the generalized least squares, or Aitken, estimator, after its discoverer (Aitken, 1934). In addition, several solutions coming from psychology will be discussed: (a) averaging across the two elements of the couple (Coren & Hakstian, 1990), (b) the use of a randomization test (Edgington, 1966; 1987) and the presentation of a new randomization test referred to as the couplet referenced randomization test, and (c) other nonparametric methods. Chapter 3 discusses the derivation of formulas for the one-group, two-group, and multi-group completely randomized designs. Unlike the bulk of Chapter 2, the results of the third chapter are more in tune with traditional psychological approaches to data analysis. Chapters two and three represent the analytical investigation while chapters four and five represent the Monte Carlo investigation into the statistical methods that overcome nonindependence of coupled data in inference.

Therefore, based on the results of Chapter 3, the remainder of the study will focus on the use of Monte Carlo methods to:

- (a) Investigate the behavior of the normal theory tests with coupled data. That is, we are further investigating what happens to the type I and II error rates.
- (b) Investigate the type I error rate and power values for the modified t and F formulas, presented in Chapter 3, and compare their power to Coren and Hakstian's (1990) proposal.
- (c) Investigate the performance of the modified t and F formulas when sample estimates of the correlation are used instead of the known population values.

(d) Investigate the type I and II error properties of the randomization and the couplet referenced randomization test with coupled data.

The statement of the problem, limitations of the study, and discussion of the methodology will be the focus of Chapter 4. The remaining chapters will discuss the interpretation and implications of the results.

CHAPTER 2: ALTERNATIVE METHODS OF SOLVING THE COUPLED DATA PROBLEM

This chapter will present several methods for dealing with coupled data. An early proposed solution to the coupled data problem in ophthalmology was Ederer's (1973) suggestion² that the data be analyzed separately for the right and left eye and the results compared to establish a general conclusion. Unfortunately, Ederer's solution spuriously increases the type I error rate and halves the amount of information (and power) available for the analysis, and thus has not been widely used. The spurious increase in type I error rate stated above is analogous to the spurious increase found when conducting a series of t-tests rather than an ANOVA for a one-way layout, or the use of multiple³ ANOVAs rather than a Multivariate ANOVA. Also, if one were interested in differential ear effects, then a mixed model (i.e., split-plot) ANOVA would be more appropriate.

This chapter is divided into those solutions arising from recent developments in (a) the statistical literature, and (b) the psychological literature.

2.1 Approaches arising from the statistical literature

Approaches to the solution of the general problem of nonindependence in the statistical literature will be reviewed. As well, how these approaches can be applied to the coupled data situation will be discussed. Statisticians have explored the issue of nonindependence from survey sampling and experimental design (i.e., mixed effects

² Divenyi and Haupt (1992) in a response to Coren and Hakstian (1990) also suggest analyzing the ears separately.

³ Bock (1975) denotes this as a variablewise inflation in error rate.

models). The origin of the survey sampling approach is the use of complex survey results for statistical inference whereas mixed-effect (or hierarchical) models stem from biometrics where there is an interest in the analysis of familial data and modeling of change and growth. A third approach presented herein is an extension of generalized-least squares methodology to the coupled data situation.

2.1.1 Approach based on survey sampling

Survey sampling has not had as great an impact on data analysis conducted in psychology as analysis of experimental designs. However, a considerable amount of social research conducted in sociology and education involves the use of a sampling plan with stratification, clustering or multi-stage sampling for drawing the observations. In the statistical literature it is well known that the structure in the sampling plan and the structure in the population result in a violation of the independence assumption. In particular, these sampling plans result in a homogeneity of the observations within the lowest level of the sampling plan. That is, an intra-class correlation is introduced by the structured sampling plan.

Historically, sample surveys have been used solely for descriptive purposes where point estimates were primarily obtained. Recently, however, an interest has been shown in using the data from structured surveys for inferential comparative purposes. Therefore, only recently has the problem of statistical inference from complex surveys attracted the attention of statisticians. The question of interest is the degree to which the violation of the fundamental assumption of independence affects the outcomes of commonly used hypothesis testing procedures. In the statistical literature, an in depth investigation of the impact of structured sampling plans has been conducted for chi-square tests and regression. Given that we are concerned with continuous random variables we will focus on the results

from regression. A review of the results from the analysis of complex samples with discrete random variables (chi-square tests and loglinear models) including small samples can be found in Rao & Thomas (1988), Thomas & Rao (1987), and Thomas (in press).

As stated previously, one of the results of sampling from a structured population, especially when clustering is used, is a positive intra-class correlation⁴, which invalidates the independence assumption. Two of the earliest studies to examine the impact of this in the context of regression were Kish and Frankel (1974) and later Scott and Holt (1982)⁵. The results of this work show that when an intra-class correlation exists, the ordinary least squares⁶ (OLS) procedures are unbiased for the estimation of regression coefficients. However, the variance of the OLS estimators for the regression coefficients can be larger than those yielded when the observations are independent. This leads to underestimation of variances. The underestimation of variances, in turn, has consequences for confidence intervals and hypothesis tests. Scott and Holt further showed that the amount of influence on the estimation of the variances is a function of the number of observations in a cluster and the magnitude of the intra-class correlation. Wu, Holt and Holmes (1988) have developed a correction to the F statistic that results in the actual type I error rate (α) of the

⁴ This is generally defined as an aggregate measure of the correlation between population members. A very similar term, intra-cluster correlation, is often seen in the sampling literature. An intra-cluster correlation is an aggregate measure of the correlation between population members in the same cluster. When the class, in intraclass correlation, is a cluster then intra-class and intra-cluster correlations can be used interchangeably. For a review and discussion regarding the estimation of the intra-class correlation see Keen (1987).

⁵ In this review we are avoiding the extensive debate regarding what exactly are the parameters about which inference is required. For the purpose of this review we are recommending a balance between the assumption that the only relevant inference concerns finite population parameters, such as the population regression coefficient, and the other extreme position that envisions all inference as relation to the parameters of a probability distribution (i.e., a superpopulation) of which the finite population represents a realization. See Nathan (1981) for an excellent review of our position on this issue. Furthermore, the overall conclusions, for all present purposes, are much the same for the two extreme positions.

⁶ Ordinary least squares (OLS) is the commonly found least squares solution in psychological research. In the psychological literature, the term least squares refers to ordinary least squares as opposed to generalized least squares or weighted least squares.

test being very close to the nominal α . A brief review of the results of Wu et al. are given below.

Our notation follows standard statistical notation (Timm, 1975), wherein a bold-faced character represents a matrix, Greek letters are population values and Roman letters represent sample statistics. The results of Wu et al. as well as Scott and Holt (1982) are based on two-stage sampling. Cluster sampling may be modified or extended to cope with the specific demands of very complicated situations. One example arises when the selected clusters in the primary cluster are themselves sampled, rather than fully inspected. Therefore, if a simple random sample of size m of the M clusters which comprise the population is chosen, we may then take simple random samples of sizes $n_1, n_2, \dots, n_m$ of the secondary elements in the chosen clusters. Often we have $n_i < N_i$, where n_i is the sample size of the clusters due to subsampling, and N_i is the sample size of the clusters in one-stage cluster sampling. In contrast, in one-stage cluster sampling $n_i = N_i$. The process of drawing samples from the selected clusters is called sub-sampling. The resultant total sample size,

$$n = \sum_{i=1}^c n_i \quad (2.1.1-1)$$

is called a two-stage cluster sample, where n is the total sample size.

For the problem of intracluster correlation the model

$$Y = XB + e$$

is similar to the formulation under independence; however the n observations form a two-stage sample with c clusters drawn at the first stage and m elements drawn from the C th sampled cluster and in the second stage n is defined in (2.1.1-1) above. The covariance matrix, V , of the errors, or likewise of observations has a block diagonal form.

Under OLS with the assumption that ρ is zero (i.e., V is an identity matrix),

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X' Y, \\ \text{Var}(\hat{\beta}) &= \hat{\sigma}^2 (X'X)^{-1}, \\ \hat{\sigma}^2 &= \frac{Y'[I - X(X'X)^{-1}X']Y}{(n - k)}.\end{aligned}\tag{2.1.1-2}$$

Where $\hat{\beta}$ is the vector of estimated weights (also called regression coefficients), $\text{Var}(\hat{\beta})$ is the covariance of the coefficients, $\hat{\sigma}^2$ is the estimated error variance, and k is the number of explanatory variables (x-variables). The boundary of the joint confidence region for all k of the regression parameters β_i ($i = 0, 1, \dots, k - 1$) is the solution for β_i of

$$\frac{(\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta) / k}{\hat{\sigma}^2} = F(1 - \alpha; k, n - k),\tag{2.1.1-3}$$

with confidence coefficient $1 - \alpha$. Where $F(1 - \alpha; k, n - k)$ denotes the F value, with k and $n - k$ degrees of freedom, exceeded by probability $1 - \alpha$. However, the region defined by this boundary is often difficult to interpret. A slightly modified form of (2.1.1-3) can also be used to test the hypothesis that the coefficients are zero at an α significance level. It should be recalled that (2.1.1-2) and (2.1.1-3) are true given that standard OLS assumptions including independence of the errors are true.

Wu, Holt, and Holmes (1988) examined the effect of intra-class correlation on F , defined in (2.1.1-3), by approximating its numerator and denominator by constant multiples of chi-square random variables, $\chi^2(k)$ and $\chi^2(n - k)$. By matching first moments, they derived that the constants are $\text{tr}(PV)/k$ and $[n - \text{tr}(PV)] / (n - k)$, respectively, where,

$$P = X(X'X)^{-1}X',$$

tr is the trace of a matrix, and V is defined above. As a check for this solution, if $\rho = 0$ then $V = I$ and $\text{tr}(PV) / k = 1.0$, which is true in the case of simple random sampling (i.e.,

independence). Scott and Holt (1982) stated that even if the OLS estimator given in (2.1.1-2) is unbiased and reasonably efficient under two-stage sampling, the true covariance matrix of OLS regression coefficients is given by $C = \sigma^2 (X'X)^{-1} D$, where $D = (X'VX) (X'X)^{-1}$. Since the true covariance matrix of the regression coefficients under the independence of the errors is $C = \sigma^2 (X'X)^{-1}$, D represents the inflation factor needed to correct standard results for the effect of intracluster correlation. As Scott and Holt state "the matrix D represents the effect of wrongly omitting the intracluster correlation from the model and is conditional on the observed X " (p. 850), and they go on further to label D the misspecification effect. A misspecification effect that is larger than I , the identity matrix, will result in an inflated probability of committing a type I error.

These results were used by Wu et al. (1988) to derive a modified F statistic, analogous to (2.1.1-3), that would compensate for the misspecification. Assuming that the value of ρ is known, the modified statistic is

$$F^* = F \left\{ \frac{k[n - \text{tr}(PV)]}{\text{tr}(PV)(n - k)} \right\} \quad (2.1.1-4)$$

where F is defined in (2.1.1-3). If one is interested in some subset, q , of the regression weights (for example, one may want to test the reduced model without a constant) then the modified F statistic is

$$F_{\text{sub}}^* = F_{\text{sub}} \left\{ \frac{q[n - \text{tr}(P_{\text{sub}} V)]}{\text{tr}(P_{\text{sub}} V)(n - k)} \right\} \quad (2.1.1-5)$$

$$P_{\text{sub}} = X_{\text{sub}} (X_{\text{sub}}' X_{\text{sub}})^{-1} X_{\text{sub}}'$$

where P_{sub} is the projection matrix onto the reduced column space of X_{sub} , which is of dimension q , where $q < k$. The applicability of the (2.1.1-4) and (2.1.1-5) depends upon the availability of the value of ρ .

Wu et al. provided simulation results for the two x-variables case. The simulation conditions were ten clusters with ten observations per cluster and various values of ρ . They found that hypothesis testing using (2.1.1-4) and (2.1.1-5) resulted in values of the type I error rate ranging from .05 to .07 in all cases (the nominal α was set at .05). They also investigated the effect of unequal ρ within the clusters. They found that the F test was a bit more distorted but that the main effect was due to $\rho \neq 0$ rather than the unequal values of ρ . They concluded that the adjustments to the F statistic based upon $\text{tr}(\text{PV})$ were both simple and effective in maintaining the nominal type I error rate when the observations were sampled in a two-stage plan.

Applying the above results to the case of the coupled data situation is, in principle, quite simple. In a general coupled data situation, where there is a sampling of people and of the elements of the pair, the two-stage model assumption is that individuals would be considered the primary sampling unit and the two elements of the coupled data would be the secondary sampling units. The results of Wu et al. could then be applied. One issue that is yet to be considered is whether there would be a bias in the results if the secondary sampling units are not the result of simple random sampling (i.e., in the two-ear case subjects can be considered randomly sampled, however, given that there are only two ears they are fully inspected).

2.1.2 Approach based on mixed effects models

Mixed effects models (also referred to in the statistical and psychological literature as ANOVA Model III) are those which have both fixed and random factors. These models have traditionally been explored in experimental settings; however, in biometrics they have also been used for the analysis of change and growth (for an example see, Bock, Wainer, Petersen, Thissen, Murray, & Roche, 1973) and in edumetrics they have been used to examine school effectiveness (for an example see, Goldstein, 1986).

Rosner (1982) suggested the use of a nested mixed effects analysis of variance -- ANOVA-- (Neter, Wasserman, & Kutner, 1985) model of the form,

$$\begin{aligned}
 y_{ijk} &= \mu + \alpha_i + \beta_{ij} + e_{ijk}, \\
 i &= 1, \dots, g, j = 1, \dots, P_i, \text{ and } k = 1, 2, \dots, P_i \quad (2.1.2-1) \\
 \beta_{ij} &\sim N(0, \sigma_\beta^2), e_{ijk} \sim N(0, \sigma^2), \text{ and } \mu, \{\alpha_i\} \text{ are constants.}
 \end{aligned}$$

That is, one wishes to compare g groups of people on some measure, y , where there are P_i persons in the i th group, and $N = \sum \sum N_{ij}$ is the total number of eyes summed over all persons over all groups. Therefore, the group effect (represented by $\{\alpha_i\}$) is considered to be a fixed effect, while the effect of persons within groups (represented by $\{\beta_{ij}\}$) and the eyes within persons (represented by $\{e_{ijk}\}$) are considered to be random effects. An appropriate F statistic can then be computed with $(g - 1)$ and $(P - g)$ degrees of freedom for the numerator and denominator, respectively. It should be noted that the standard assumptions of ANOVA, such as equality of variances and normal errors, are also true of the random and mixed effects ANOVA. Unlike regression analysis, when fitting ANOVA models, the assumption of normality is seldom examined and its effect on the mean, variance, and statistics used for inferences is often underestimated (Lind & Zumbo, in press; Zumbo, in press).

2.1.3 Approach based on generalized least squares

This section will present the use of generalized least squares as a solution of the coupled data problem. This solution of the coupled data problem is novel to the literature, and therefore a detailed presentation will be given. The presentation will begin with a non-general Gauss-Markov set-up (i.e., with independent errors) then the more general model will be given. Many of the concepts in this section have been motivated by Timm (1975).

Non-General Gauss-Markov: The model and estimation

In the non-general Gauss-Markov set-up the linear model is used for the analysis of experimental designs by ANOVA (Timm, 1975). It also accommodates analysis of covariance and regression modelling methods. The model is commonly written as

$$\mathbf{Y}_1 = \mathbf{X}_q \beta_1 + \epsilon_1 \quad (2.1.3-1)$$

Throughout this section matrices and/or vectors are noted in bold letters and their order is given in subscripts. For simplicity, order of the matrices will only be given when they are first presented to the reader. The vector of observed scores is denoted $\mathbf{Y}$, and the design matrix is denoted $\mathbf{X}$ (which is a general form of a design matrix containing coding indicators for group membership, covariates, or regressors). The model coefficients are contained in β and ϵ is the matrix of model errors.

In the above model the expectation, $E(\epsilon)$, of the vector of model errors is assumed to be a zero matrix, $\mathbf{0}$. For statistical inference, it is assumed that ϵ follows a multivariate normal distribution where the variance, $V(\epsilon) = \Sigma$, is unknown. In the model commonly used in social science research it is further assumed that $\Sigma = \sigma^2 \mathbf{I}$, where σ^2 is an unknown variance. This implies that the errors are uncorrelated.

Figure 2 demonstrates the notions of independence and the special dependence in the coupled data case. The matrices are symmetrical covariance matrices for two subjects, where X_1 and X_2 are the random variables for the two ears for subject 1, and X_3 and X_4 are the random variables for subject 2. Matrix $\mathbf{A}$ represents the independence

Figure 2. Variance-covariance matrices for independent and coupled data.

Matrix A				
	X_1	X_2	X_3	X_4
X_1	σ_1^2			
X_2	0	σ_2^2		
X_3	0	0	σ_3^2	
X_4	0	0	0	σ_4^2

Matrix B				
	X_1	X_2	X_3	X_4
X_1	σ_1^2			
X_2	Cov(2,1)	σ_2^2		
X_3	0	0	σ_3^2	
X_4	0	0	Cov(4,3)	σ_4^2

assumption in the completely randomized design. It is clear that all the covariance terms are zero, indicating mutual independence.

The estimation of the coefficients (or weights) for the model is usually performed by ordinary least squares methods. The ordinary least squares estimates are obtained by

$$\hat{\beta} = (X'X)^{-1} X'Y, \quad (2.1.3-2)$$

$$\hat{\sigma}^2 = \frac{Y'[I - X(X'X)^{-1}X']Y}{N_{\text{total}} - R(X)} = \text{MSE}. \quad (2.1.3-3)$$

Where (2.1.3-2) is the unbiased sample estimate of the coefficients of the model and (2.1.3-3), in the univariate case, is the estimate of the error variance, also referred to as the mean squared error. The total sample size is denoted by N_{total} and the rank of the design matrix is denoted $R(X)$, the number of linear model coefficients being estimated.

Testing hypotheses

Now let us denote C as a contrast matrix and we are interested in testing the hypothesis $H_0: C\beta = 0$ and the alternative $H_1: C\beta \neq 0$. The F-ratio for the test of C is

$$F = \frac{(C' \hat{\beta})' [C'(X'X)^{-1}C]^{-1} (C' \hat{\beta}) / r}{\text{MSE}}, \quad (2.1.3-4)$$

where the degrees of freedom for the numerator is the rank of the contrast matrix, C , also denoted as r in (2.1.3-4). The denominator is defined in (2.1.3-3) and has a degrees of freedom equal to the total sample size minus the rank of X .

The general Gauss-Markov setup: Model and estimation

The general Gauss-Markov setup can be used to model the coupled data problem. The primary issue is that the covariance matrix of the error terms is not proportional to an identity matrix, I . It is important to note that algebraically the general Gauss-Markov model is the same as in (2.1.3-1) where the expected value of the errors is zero. However, the covariance of the errors is now V . For the general Gauss-Markov setup $\Sigma = \sigma^2 V$ rather than $\Sigma = \sigma^2 I$.

In stating that the covariance of the errors is not proportional to I but some specified matrix, proportional to V , we are invoking a generalized approach to model (2.1.3-1). The covariance matrix, V , is similar in form to Matrix B of Figure 2. Therefore, Matrix B represents the coupled data situation, where the two measures on a particular subject are

dependent on each other but independent of the measures of other subjects. The observed scores and the error scores, under the independence assumption, are patterned as in Matrix A of Figure 2 and they are patterned as in Matrix B of Figure 2 if the observations are from a coupled data situation.

The appropriate estimator of β , in the linear model given in (2.1.3-1), when the errors have covariance structure V is not the ordinary least squares estimator given in (2.1.3-2) but rather

$$\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}Y, \quad (2.1.3-5)$$

where the notation is the same as above. It should be noted that the generalized least squares estimator in (2.1.3-5) is an unbiased estimator of β , and that if we assume that the errors are normally distributed then (2.1.3-5) is a maximum likelihood estimator of the weights. Finally, if we assume that $V = \sigma^2 I$, as in the non-general case above, the $\hat{\beta}$ reduces to (2.1.3-2).

The generalized least squares analogue to the ordinary least squares estimate of the error variance, (2.1.3-3), is given by

$$\hat{\sigma}^2 = \frac{1}{N_{\text{total}} - R(X)} Y'[V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}]Y = \text{MSE}. \quad (2.1.3-6)$$

The degrees of freedom are the same as in the ordinary least squares estimator.

The generalized least squares estimation technique is useful when the variance and covariance of the errors can be specified. For the case of the coupled data the variances (i.e., the diagonal of the variance-covariance matrix of errors) are standardized at 1.0. The within-group correlations (the standardized covariances, $\text{Cov}(i,j)$, $i \neq j$, in Figure 2) can be calculated from the sample or from bench mark values (Coren & Hakstian, 1990).

Testing hypotheses

As in the non-general case above, let us denote a contrast matrix by C . We now will use an F-ratio to test the generalized least squares model. That is, to test the hypothesis $H_0: C\beta = 0$ and the alternative $H_1: C\beta \neq 0$, we use the following F-ratio which is a simple analogue of (2.1.3-4)

$$F = \frac{(C' \hat{\beta})' [C'(X'V^{-1}X)^{-1}C]^{-1} (C' \hat{\beta}) / r}{MSE} \quad (2.1.3-7)$$

It should be noted that the MSE for (2.1.3-7) comes from (2.1.3-6) and that the degrees of freedom for F are the same as in (2.1.3-4).

Transformational approach to computing GLS

Rather than use the formulas (2.1.3-5) to (2.1.3-7), another approach to conducting a GLS analysis is to take advantage of the following principle: For any positive definite matrix V it is possible to find a nonsingular symmetric matrix, T , such that $TT = T^2 = V$ (see, for example Timm, 1975). Since T is nonsingular, it has an inverse, T^{-1} . Now, premultiplying (2.1.3-1) by T^{-1} results in

$T^{-1}Y = T^{-1}X\beta + T^{-1}\epsilon$, and with some rearranging the resultant is

$Y^* = X^*\beta^* + \epsilon^*$ where $Y^* = T^{-1}Y$, $X^* = T^{-1}X$, $\epsilon^* = T^{-1}\epsilon$, and β^* is the GLS estimator given in (2.1.3-5). With this transformation $\text{Var}(\epsilon^*) = T^{-1}VT^{-1}\sigma^2 = I\sigma^2$ and ordinary least squares is appropriate for Y^* and X^* .

Many of the statistical programs used in the social and behavioral sciences are not designed to compute GLS for ANOVA or t-tests. It is possible, however, to make the transformation of the observations, discussed above, and use OLS methods, or to resort to a matrix algebra computer program and use (2.1.3-5) to (2.1.3-7).

It is important to note that in this chapter we have presented a one-step solution for the GLS results. However, iterative methods are sometimes used with GLS. That is, when the elements of V are not parameterized -- i.e., contains sample estimates rather than parameters --, the estimation of (2.1.3-5) proceeds in several steps:

- (i) an OLS regression as a starting point,
- (ii) an estimation of the covariance matrix from the residuals,
- (iii) a GLS regression using the estimated covariance matrix,
- (iv) an estimation of the covariance matrix from the residuals.

If the covariance matrix of the residuals differs beyond some prespecified value from step (ii) then the algorithm continues with steps (ii) and (iii) until it converges given the prespecified limit.

The generalized least squares solution has several interesting features:

1. The covariance matrix of the errors is specified in a general form; therefore issues of unequal variances among groups can be easily handled, as well as unequal correlation among the units of the coupled data. It should be noted that it is possible to specify submatrices of the covariance matrix such that the correlations differ for each group, for a subset of subjects, or for each subject. Therefore, it is clear that the GLS solution could potentially deal with a broad range of experimental and non-experimental scenarios.

2. All of the degrees of freedom associated with the individual observations are preserved, unlike the Coren & Hakstian (1990) solution of averaging. This, of course, results in a maintenance of maximal statistical power for hypothesis testing.

3. Although the above results are presented as an intuitive derivation, it should be noted that, given a prespecified V , the GLS solution is an exact test. If the elements of V

are not population values then the result is an approximate test. The approximation to the GLS tests depends on the accuracy (i.e., deviation) of the estimate of p .

4. In its approximate or exact form the GLS methodology can test the same hypotheses commonly found in psychological research.

5. The above solution is rather general and therefore many other models, such as simple and multiple regression, analysis of covariance, as well as multivariate techniques such as discriminant analysis or multivariate analysis of variance could be derived using it.

6. In the following section, it is argued that the GLS solution can be utilized in the single-subject context. In some research settings, for example psychophysics or operant conditioning, there is an interest in conducting single-subject experiments. This implies, of course, that one subject participates in all of the experimental conditions and is the replications in the design. For example, an experiment could be designed to investigate the effect of three distinct experimental conditions on the perception of the equality of line lengths⁷. The subject's task is to decide whether the line lengths are equal. The subject will be presented with 25 unique pairs of line lengths in each condition. The dependent variable of interest for this hypothetical study is not only the correct judgement, but also response time in making a decision. For each dependent variable, the experimenter would perform a one-factor within-subjects ANOVA with 25 observations per cell. It is this author's contention that this is inappropriate because of the possible within-groups correlation resulting from using one subject as the 25 replications per group⁸.

It could conceivably be argued that there is no violation of within-groups independence in a single-subject context. That is, one could conceivably define a universe for the random variable under investigation wherein independence could be correctly

⁷ This example comes from discussions with Dr. Joe Baranski, Carleton University.

⁸ However, inference to the single subject is appropriate.

assumed. A distribution of this nature is found in psychometric theory. Lord and Novick (1968) labeled such a universe a propensity distribution. That is, if a person's true score is a constant over replications then the resulting theoretical distribution is the propensity distribution. For the above example, the subject has a propensity distribution under each experimental condition. That is, the subject has a propensity distribution for deciding equality of the line lengths.

As Zimmerman, Zumbo, and Lalonde (in press) stated, the notion that two error-scores (or two scores) from one individual are uncorrelated is not a tautology, but an assumption that may not be realistic in practice. In the above example, it is possible that the observations from the subject are independent, but it is not necessarily true. In fact, the possibility of correlated errors is precisely the motivation for using a repeated measures analysis in the above example.

In discussing the notion of independence in psychometrics, Rozeboom (1966) stated,

Distressingly, however, this is a most unrealistic assumption. When the individual items on a test are administered at one sitting it is inevitable -- or at least we have no reason to think otherwise -- that a substantial proportion of the uncontrolled extraneous conditions which are presumably responsible for measurement error will persist from one item to the next, thus inducing positive error correlations which may well be substantial. (p. 415)

Rozeboom also goes on to state that for practical applications, assuming independence is akin to "putting on a clean shirt to rattle a hog" (p. 415). A proper randomization mechanism is sufficient to guarantee independence. It is not a necessary condition for independence because independence may occur without it. In the practice of statistics, if the sufficient condition for randomization is met, then one can make the assumption of independence of the data.

The issue of necessary conditions for independence is very crucial to the proper application of statistics derived from independent random variables. The randomization mechanism assures the independence. For example, cluster sampling could result in independent observations at the lowest level of sampling, however, it is not necessary that they are independent.

Therefore, given that we have established that for the single-subject design it is impractical to assume independence, an alternative statistical method is required. The GLS formulas presented in this section can also be used for the analysis of single-subject data⁹. Again, due to the fact that the covariance matrix is presented in a general form in (2.1.3-5) to (2.1.3-7), these equations can be used in a large variety of contexts. For the single-subject design, V should be specified with a common intra-class covariance for each experimental condition. Standard statistical hypotheses could then be tested with (2.1.3-7) or through the transformational approach.

At this point it is important that a caveat be given. As Rawlings (1988) states when V is not parameterized (i.e., contains all sample estimates and not known parameters), the GLS solution is no longer the minimum variance solution; therefore, in part, compromising replicability of the results. In the worst cases, where there is limited information with which to estimate V , the GLS estimates have substantially larger variances than the OLS estimates. Furthermore, if the off-diagonal elements of V (i.e., the covariance between the two elements of the couplet) are sufficiently large absolute values then the GLS estimates may be quite misleading. This is due to the fact that the inverse of V may be difficult to numerically ascertain. If the GLS regression line is plotted on the original scale all of the data points could lie on one side of the regression line and hence not be representative of

⁹The approach in Zimmerman & Zumbo (in press b) or Zimmerman, Williams, & Zumbo (1992) could also be used in this context.

the data (in data analysis terminology the regression line "missed" the data). This possible missing of the data is true irrespective of whether V contains only parameters.

However, estimation of V will probably cause the problem to occur more frequently.

Rawlings recommends plotting the GLS regression line on the original scale (i.e., if a transformational approach is utilized plot the line on the original scores) to check if the data are missed.

2.2 Approaches based on the psychological literature

To this point, we have seen three possible solutions from the statistical literature for the analysis of nonindependent observations. Given that the focus of this paper is the analysis of psychological data, we will now focus on three approaches to the analysis of nonindependent data that are currently advocated in the psychological literature: (a) Coren and Hakstian's (1990) solution, (b) randomization tests, and (c) nonparametric rank tests. It is important to note that even though these methods are advocated in the psychological literature they could be used in many other fields.

2.2.1 Coren and Hakstian's approach

As a solution to the coupled data situation in auditory perception research, Coren and Hakstian (1990) suggested calculating the standard analyses on a reduced sample obtained by averaging the two measures on each subject. The Coren and Hakstian recommendation result from a methodology which Darlington and Carlson (1987) refer to as clustering¹⁰.

¹⁰ The problem of clustering is also found in the educational methodology literature wherein it is referred to as the units of analysis problem. See Glass and Stanley (1970) and Blair, Higgins, Topping, and Mortimer (1983) for a discussion of the units of analysis problem and its impact on type I error rates during significance testing.

As a methodology clustering is a procedure by which to redesign the statistical analysis. Clustering procedures vary considerably in complexity, however, the basic process is the following:

- (a) form groups of observations (i.e., clusters) so that there is independence between clusters but nonindependence within clusters,**
- (b) compute an index number which summarizes the data in each cluster, and**
- (c) apply a statistical significance test to the index numbers.**

In the case of coupled data, Coren and Hakstian (1990) note that given the high interear correlation it makes most sense to sum or average the measures across the two ears for each individual in order to obtain a single number that is representative of the data for each individual. Zumbo and Zimmerman (1991) state that the results of the statistical significance test would not be different for simply calculating a difference or summing across the two ears. In fact, the power and type I error rates will be the same for summing, averaging, or any other linear function of the two ear measures.

Two points are noteworthy at this juncture. First, the type I error rate is maintained at its nominal value by clustering (Zumbo & Zimmerman, 1991). Second, this protection of the type I error rate does not come without a cost. That is, implicitly the clustering methodology reduces the sample size, results in a loss of degrees of freedom, and hence diminishes the statistical power to detect a true mean difference. The reduction in power should be more pronounced for small sample sizes than large sample sizes.

2.2.2 Approach based on randomization tests

A randomization test is a permutation test¹¹ which has the randomization mechanism as the basis of its permuting. In the case of randomization tests, if the randomization mechanism is as nearly perfectly balanced as physically possible, it is legitimate grounds for distributing equal degrees of certainty over the elementary outcomes of the probability space. A randomization test is not a statistical test in the usual sense but a way of determining significance. Therefore, these tests are often considered alternatives to statistical tables or computer determined p-values via the integral of a probability function. As such, Edgington (1980) states that the randomization tests do not rely on the assumptions that the standard sampling distributions and tabled values require, such as normality.

Randomization tests are carried out, with the aid of a computer, in the following manner. First, a test statistic is computed for the experimental data. Then the data are permuted (divided or rearranged) repeatedly in a manner consistent with the random assignment procedure, and finally the test statistic is computed for each of the resulting data permutations. These data permutations, including the one representing the obtained results, constitute the conditional empirical sampling distribution. The proportion of data permutations in the empirical sampling distribution with test statistics greater than or equal to (or for certain test statistics less than or equal to) the experimentally obtained results is the probability (or significance) level.

The preceding definition is broad enough to include randomization procedures that allow random sampling as well as random assignment (Kempthorne, 1955). Edgington (1980) represents the modern conception of a randomization test as one which is based on

¹¹ A permutation test is a test whose conditional empirical sampling distribution is obtained by a complete reordering (i.e., permutations) of the observations. Approximate permutation tests are those based on a less than complete reordering usually a subsampling.

random assignment alone, where it does not matter how the sample is selected. It is Edgington's conception of a randomization test that is the concern of this paper. Edgington (1966; 1980; 1987) built on the notion that the resulting randomization tests are not dependent on how the sample is selected to advocate their use in single-subject designs. As was stated above, these designs are those in which correlated errors are possible.

Given the above, randomization tests are a potential solution to the coupled data problem. Using randomization tests, thus ignoring the sampling structure, implies randomizing at the basic unit level. Therefore, a randomization test would be conducted on the elements of Figure 1, ignoring the pairing. It seems intuitively appealing that if one were to permute a set of objects, as is done in randomization tests, the sampling structure would not be relevant. A randomization test could then be performed ignoring the coupled data structure.

However, the sampling structure is not relevant only if another randomization mechanism, such as random assignment, is present. It is important to note that for a randomization test to be valid the test must maintain a uniform distribution over the space of permutations. This uniformity may not be true with nonindependent data (Thomas, in press; Zumbo & Zimmerman, in press) such as that evident with coupled data. Furthermore, Conover (1980) clearly states that independence within groups is an assumption of randomization tests. That is, the null-hypothesis of randomization tests is defined in the following:

H_0 : Hypothesis of randomness: The random variables $X_1, \dots, X_N$ satisfy the hypothesis of randomness if they are independent and governed by a common distribution function, $F(x)$, which is continuous but otherwise arbitrary.

The independence is guaranteed only with a randomization mechanism such as random assignment. If the hypothesis of randomness is rejected (using Fisherian terminology) then the groups are different over and above the randomization mechanism.

Very little empirical investigation has been devoted to examining the effect of nonindependence within groups on the statistical properties of the randomization test. A recent Monte Carlo study by Harwell (1991) shows that the empirical type I error rate for the randomization test is inflated as much as that of the t-test for the case where all of the possible pairings of observation within a group are equally correlated. This is a far more extreme case of nonindependence than that with coupled data; however, it suggests that the randomization test will perform as poorly as the t-test with coupled data. Although Harwell's results are suggestive, more empirical studies need to be conducted investigating the effect of coupled data on the empirical type I error rate of randomization tests.

Given that the standard permutation test, as described above, should not work properly for coupled data, an alternative will be proposed. Given that randomization tests are permutation tests which have the randomization mechanism as the basis of their permuting, it seems intuitively appealing that if one were to randomize on the basis of the coupled data then the permutations should be uniformly distributed over the sample space. That is, in studies involving coupled data the experimental unit that is being randomized is the individual subject. Therefore, for the randomization test to be valid one must permute the observations to obtain a conditional couplet referenced sampling distribution. This couplet referenced sampling distribution is obtained by permuting at the couplet level not at the individual observation level. Hence, this new test statistic is referred to as a couplet referenced randomization test.

Two important points should be noted. First, Edgington (1980) denotes coupled data situations as those involving complex experimental units. Edgington (1980) and

Thomas (in press) state that given that the complex experimental units are randomly assigned to treatment conditions (e.g., subjects assigned to conditions) the standard randomization approach can be applied to the mean or some other linear function of the experimental units. In the case of coupled data, one would calculate the average among the elements of the couple and then randomize that average. This would result in a randomization version of the Coren and Hakstian (1990) procedure.

Second, the randomization version of the Coren and Hakstian procedure is different than the couplet referenced randomization test. This will be demonstrated with the aid of Figure 3. The data structure shown in Figure 3 represents the observations from a two group completely randomized design involving four subjects, wherein X_1 and X_2 are from subject one, X_3 and X_4 are from subject two, X_5 and X_6 are from subject three, and X_7 and X_8 are from subject four. The upper part of Figure 3 shows the sampling distribution for the standard randomization test. The eight observations are permuted on the basis of group membership. The resulting randomization test would involve 70 t-tests with four observations per group. The middle part of Figure 3 shows the sampling distribution for the couplet referenced randomization test. The eight observations are permuted on the basis of group membership, however, the observations from each subject are not separated. The resulting randomization test would involve 6 t-tests with four observations per group. The bottom part of Figure 3 shows the sampling distribution for the randomization version of the Coren and Hakstian procedure. The average of the data points for each subject is represented by the subject number so that $1 = (X_1 + X_2)$, $2 = (X_3 + X_4)$, and so on. The resulting randomization test would involve 6 t-tests with two observations per group. Figure 3, therefore, clearly shows that the sampling distributions for the various randomization tests are different.

Figure 3. Example of Standard and Couplet-Referenced Randomization Tests, and the Randomization version of the Coren and Hakstian Procedure.

<u>Standard Randomization test</u>		
<u>Permutation</u>	<u>Group 1</u>	<u>Group 2</u>
1	X ₁ , X ₂ , X ₃ , X ₄	X ₅ , X ₆ , X ₇ , X ₈
2	X ₁ , X ₂ , X ₃ , X ₅	X ₄ , X ₆ , X ₇ , X ₈
3	X ₁ , X ₂ , X ₃ , X ₆	X ₄ , X ₅ , X ₇ , X ₈
4	X ₁ , X ₂ , X ₃ , X ₇	X ₄ , X ₅ , X ₆ , X ₈
5	X ₁ , X ₂ , X ₃ , X ₈	X ₄ , X ₅ , X ₆ , X ₇
6	X ₁ , X ₂ , X ₅ , X ₄	X ₃ , X ₆ , X ₇ , X ₈
7	X ₁ , X ₂ , X ₆ , X ₄	X ₃ , X ₅ , X ₇ , X ₈
.	.	.
.	.	.
.	.	.
70	X ₅ , X ₆ , X ₇ , X ₈	X ₁ , X ₂ , X ₃ , X ₄

<u>Couplet-Referenced Randomization test</u>		
<u>Permutation</u>	<u>Group 1</u>	<u>Group 2</u>
1	X ₁ , X ₂ , X ₃ , X ₄	X ₅ , X ₆ , X ₇ , X ₈
2	X ₁ , X ₂ , X ₅ , X ₆	X ₃ , X ₄ , X ₇ , X ₈
3	X ₁ , X ₂ , X ₇ , X ₈	X ₃ , X ₄ , X ₅ , X ₆
4	X ₃ , X ₄ , X ₅ , X ₆	X ₁ , X ₂ , X ₇ , X ₈
5	X ₃ , X ₄ , X ₇ , X ₈	X ₁ , X ₂ , X ₅ , X ₆
6	X ₅ , X ₆ , X ₇ , X ₈	X ₁ , X ₂ , X ₃ , X ₄

Randomization version of the Coren and Hakstian Procedure

<u>Permutation</u>	<u>Group 1</u>	<u>Group 2</u>
1	1, 2	3, 4
2	1, 3	2, 4
3	1, 4	2, 3
4	2, 4	1, 3
5	2, 4	1, 3
6	3, 4	1, 2

In summary, recent findings by Harwell (1991) would suggest that the standard randomization test would result in an inflated type I error rate for coupled data. Unfortunately, unaware of Harwell's findings, Zumbo and Zimmerman (1991) suggested the standard randomization test as a possible statistical method to treat coupled data. Instead of the standard randomization test, I now propose the couplet referenced randomization test. The couplet referenced randomization test is intuitively valid since it permutes the the experimental unit which was initially randomized, the coupled data or couplet. However, empirical data need to be gathered as to the statistical properties of the couplet referenced randomization test. Finally, the Edgington (1980) and Thomas (in press) suggestion of obtaining an average across the coupled data and then permuting the average score $\bar{y}_i$ similar to the Coren and Hakstian recommendations. In fact, given that standard tests such as the t-test are approximations to the results from randomization tests, the Coren and Hakstian procedure is a computationally simpler approximation to the randomization version of the Coren and Hakstian procedure and hence suffers from its limitations.

2.2.3 Approach based on nonparametric rank tests

It should be noted that randomization tests are examples of nonparametric statistical methods (also referred to as distribution-free tests). That is, randomization tests are derived without an assumption as to the shape of the population distribution. In the area of hypothesis testing and confidence intervals, the significance level and confidence coefficient of nonparametric tests remain constant over a broader class of distributions than for parametric tests. Therefore, nonparametric tests are valid for a broader class of distributions than parametric tests. Most nonparametric tests are based on ranks; however, this is not necessarily true of randomization tests. What makes randomization tests unique among nonparametric tests is that the randomization test's referent distribution is a

conditional empirical sampling distribution rather than a theoretical sampling distribution based on the order statistics.

Hence, what is conspicuously missing from our discussion of alternatives for the coupled data situation are nonparametric rank tests. The reason for their exclusion as viable options is two-fold. First, nonparametric rank tests are known to be equivalent to parametric tests performed on the ranks of the original scores (Conover & Iman, 1981; Zimmerman & Zumbo, 1993). For example, a Wilcoxon-Mann-Whitney (WMW) test can be conducted by pooling the raw data, performing a rank transformation, and conducting a Student's t-test on the ranks. Zimmerman and Zumbo (1993; in press a) also have demonstrated that the WMW as well as the Kruskal-Wallis test fail to be robust under the same conditions that their parametric counterparts also fail to be robust. That is, the type I error rate of parametric and nonparametric tests of shift is severely altered by: (a) unequal variances with unequal sample sizes, or (b) nonindependence between and/or within groups.

The results of Zimmerman and Zumbo (in press a) clearly demonstrate that nonparametric rank tests are not a solution to violation of independence. These results have led this author to conclude that nonparametric rank tests will not correct the inflation in type I error rate found in the coupled data situation. Therefore, nonparametric rank tests are not appropriate for coupled data.

2.3 Summary

Several possible solutions of the coupled data problem were presented. The problem of nonindependence of error terms is pervasive in many research settings. In applied statistics, the problem occurs in applying standard methods to the analysis of complex survey data, and in biometrics with familial research. In psychology, the problem

arises with coupled data, single-subject analyses, and with nonrandomness in social science research (Maguire & Rogers, 1989).

There are two limitations to the solutions offered in this chapter. First, it can be demonstrated, by investigating the history and the frequency of the use of nonparametric and Bayesian methods, that psychological researchers do not like to stray from their comfortable ground of normal theory statistics. As well, the choice of statistics is often based on the availability and ease of use of statistical software. Therefore, the methods coming from developments in statistical research will probably not be put to use for some time. The second issue is that the methods presented above cannot be used easily in a meta-analytic context to see if the incorrect analysis will change substantive findings.

CHAPTER 3: DERIVATION OF CORRECTION FORMULAS FOR COUPLED DATA

In this chapter formulas are derived based on statistical methods commonly used in psychology. The objective of the derivations is to produce a statistical methodology that is a generalization of the standard normal theory statistics (from commonly used programs such as SPSSX[®], BMDP[®], SAS[®]). Where possible the new formulas will be presented as functions of the commonly found statistics.

The chapter is organized in the following manner: First, the basic equations will be derived. These basic equations serve as a foundation for the derivations of the new formulas. Next, the one-group, two-group, and k-group problems will be treated separately. Beyond the derivation of the new formulas, a goal of this chapter is to show how these new results are generalizations of the normal theory statistics.

3.1 Fundamental equations

Even though the formulas for the variance of the sample mean and the mean of the sample variance can be found in many mathematical statistics textbooks, they are presented herein to highlight the importance of the assumption of independent, identically distributed (i.i.d.) random variables for the variance of the sampling distribution of the mean.

In the commonly found model, a random sample of size n is a sequence of observations of i.i.d. random variables, $X_1, X_2, \dots, X_n$. The variance of the sampling distribution of the mean is given by

$$\sigma_X^2 = \sigma^2 \left[\frac{X_1 + X_2 + \dots + X_n}{n} \right]. \quad (3.1-1)$$

Since the random variables are independent¹², (3.1-1) simplifies to

$$\sigma_X^2 = \frac{1}{n^2} [\sigma^2(X_1) + \sigma^2(X_2) + \dots + \sigma^2(X_n)]. \quad (3.1-2)$$

Furthermore, because the random variables are identically distributed (3.1-2) simplifies to

$$\sigma_X^2 = \frac{1}{n^2} [n \sigma_X^2] = \frac{\sigma_X^2}{n}, \quad (3.1-3)$$

where, σ_X^2 denotes the population variance for the sample observations.

Next we turn to the mean of the sample variance. It is denoted $E(S^2)$, where

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}. \quad (3.1-4)$$

Note that given $\sigma_X^2 = E(\bar{X}^2) - \mu^2$ and therefore $E(\bar{X}^2) = \sigma_X^2 + \mu^2$ so that

$E(S^2) = \sigma_X^2 + \mu^2 - \sigma_X^2 - \mu^2 = \sigma_X^2 - \sigma_X^2$, S^2 is biased. The degree of bias is $-\sigma_X^2$.

We know that in an i.i.d. setting the degree of bias is given in (3.1-3). Thus

$$E(S^2) = \sigma_X^2 - \frac{\sigma_X^2}{n} = \frac{n-1}{n} \sigma_X^2, \quad (3.1-5)$$

and therefore S^2 is too small by a factor of $(n-1)/n$. Multiplying S^2 by the inverse of this factor corrects for its bias. Since in our derivation n is known, we have

¹² This also follows from $\sigma^2(bX) = b^2 \sigma^2(X)$, where in our setting $b=1/n$.

$$\hat{\sigma}_X^2 = \frac{n}{n-1} S^2 = \frac{SS_X}{n-1}, \quad (3.1-6)$$

where $SS_X = \sum_{i=1}^n (X_i - \bar{X})^2$ and $\hat{\sigma}_X^2$ denotes the unbiased estimate of σ_X^2 .

To this point we have focussed on the i.i.d. case. We now turn to the coupled data problem. It is important to note that the above derivation was simplified by the fact that the random variables were independent. That is, the covariances, $\text{Cov}(X_i, X_j)$, $i \neq j$, were all zero. If these covariances were not zero then (3.1-2) would be the more complex expression

$$\sigma_X^2 = \frac{1}{n^2} \left[\sigma^2(X_1) + \sigma^2(X_2) + \dots + \sigma^2(X_n) + \sum_{i \neq j} \text{Cov}(X_i, X_j) \right]. \quad (3.1-7)$$

This expression follows from the fact that if $S_n = X_1 + X_2 + \dots + X_n$, then

$$\begin{aligned} \sigma^2(S_n) &= \sigma^2(X_1 + X_2 + \dots + X_n) \\ &= \sigma^2(X_1) + \sigma^2(X_2) + \dots + \sigma^2(X_n) + \sum_{i \neq j} \text{Cov}(X_i, X_j). \end{aligned} \quad (3.1-8)$$

It is important to note, at this juncture, that for coupled data only n of the $n(n-1)$ covariance terms in (3.1-8) are nonzero. This can be seen from the symmetric lower triangle of a matrix (Matrix B) in Figure 2. Therefore, given that we have identically distributed random variables (3.1-7) simplifies to

$$\sigma_X^2 = \frac{1}{n^2} [n \sigma_X^2 + n \text{Cov}(X_i, X_j)], \text{ where } i \neq j. \quad (3.1-9)$$

Furthermore, if $\rho = \frac{\text{Cov}(X_i, X_j)}{\sigma_X^2}$, $i \neq j$, and ρ is the same for all i and j , then for

coupled data (3.1-9) simplifies to

$$\sigma_X^2 = \frac{\sigma_X^2}{n} [1 + \rho]. \quad (3.1-10)$$

From statistical theory, we know that the mean of sample variances is biased by the amount $-\sigma_X^2$, hence

$$\begin{aligned} E(S^2) &= \sigma_X^2 - \frac{\sigma_X^2}{n} [1 + \rho] \\ &= \sigma_X^2 \frac{[n - (1 + \rho)]}{n}. \end{aligned} \quad (3.1-11)$$

We now know that $E(S^2)$ is too small by $\frac{n - (1 + \rho)}{n}$. Multiplying S^2 by the inverse of this factor corrects for its bias. Therefore, the unbiased estimate of the variance in coupled data is

$$\hat{\sigma}_X^2 = \frac{n S^2}{n - (1 + \rho)} = \frac{SS_X}{n - (1 + \rho)}, \quad (3.1-12)$$

where SS_X is defined above. As a check, if ρ equals zero, then (3.1-10) equals (3.1-3), (3.1-11) equals (3.1-5), and (3.1-12) equals (3.1-6) as expected.

In summary, with the assumption of independent identically distributed (i.i.d.) random variables at hand the variance of the sample mean, mean of the sample variance, and the unbiased estimate of the variance were derived. Then replacing the i.i.d. assumption with one of identically distributed couplet random variables, the variance of the sample mean, mean of the sample variance, and the unbiased estimate of the variance were derived in (3.1-10), (3.1-11), and (3.1-12), respectively. These formulas will be used in lieu of their corresponding i.i.d. expressions to derive the formulas for inference in the one-group, two-group, and k-group cases.

3.2 One-group case

Having equations (3.1-10), (3.1-11), and (3.1-12) at hand it is easy to obtain a modified t formula for coupled data. Since the mean is not affected by coupled data, we begin with the denominator of the t formula. The denominator is the standard error of the mean, and is given by $\sigma_{\bar{X}} = \sqrt{\sigma_X^2}$. Now substituting (3.1-10) into the formula for the

standard error results in $\sigma_{\bar{X}} = \sqrt{\frac{\sigma_X^2}{n} (1 + \rho)}$. Finally, substituting into the formula for the standard error, the estimate of the population variance given in (3.1-12) and some simple algebraic rearranging results in

$$\sigma_{\bar{X}} = \sqrt{\frac{SS_X}{n(n-1)} \times \frac{(n-1)(1+\rho)}{n-(1+\rho)}}. \quad (3.2-1)$$

Finally, using (3.2-1) as the denominator for the modified t formula gives us

$$t' = \frac{t}{\sqrt{\frac{(n-1)(1+\rho)}{n-(1+\rho)}}}, \quad v = \text{degrees of freedom} = n - 1, \quad (3.2-2)$$

where t denotes the standard normal theory t -test. As a check, if $\rho = 0$ then $t' = t$.

Equation (3.2-1) can be used in constructing a $100(1-\alpha)$ percent confidence interval for the mean, where α is customarily used to denote the risk of type I error. The confidence interval is constructed by using the following formula,

$$\bar{X} \pm t_{v: \alpha/2} \sqrt{\frac{SS_X}{n(n-1)} \times \frac{(n-1)(1+\rho)}{n-(1+\rho)}}, \quad (3.2-3)$$

where $t_{v: \alpha/2}$ equals the $100(1-\alpha/2)$ percentile value of the t distribution with v degrees of freedom

3.3 Two-group case

In this section we will derive a corrected independent samples t formula, t' , for : (a) the case of equal sample size and equal correlations for each group, and (b) the more general formula allowing for unequal sample sizes and unequal correlations. The former was presented without derivation in Zumbo and Zimmerman (1991).

Regarding notation, ρ_1 , ρ_2 , and ρ will denote the within couplet correlation for groups one, two, and the common correlation for both groups, respectively. Furthermore, n_1 , n_2 , and n will denote the sample size for groups one, two, and the common sample size, respectively.

In sampling theory, the standard deviation of a difference between two independent sample means is denoted by

$$\sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2}.$$

Combining this with (3.1-10) and with some algebraic rearranging the resulting equation is

$$\sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{(\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2) \frac{1+\rho}{n}}. \quad (3.3-1)$$

Substituting the appropriate unbiased estimate of the variance obtained from (3.1-12) into (3.3-1) the result is

$$\sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{SS_{X_1} + SS_{X_2}}{n(n-1)} \times \frac{(n-1)(1+\rho)}{n-(1+\rho)}}. \quad (3.3-2)$$

Therefore, the corrected t formula for coupled data with equal sample sizes and correlations is,

$$t' = \frac{t}{\sqrt{\frac{(n-1)(1+\rho)}{n-(1+\rho)}}}, \quad v = 2(n-1). \quad (3.3-3)$$

Again, t denotes the standard Student's t -test computed from the coupled data. Note that the correction formulas in (3.2-2) and (3.3-3) are the same.

Next, for the case where the sample sizes and the correlation values are allowed to differ for the two groups, the standard deviation of the difference is

$$\sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\sigma_{\bar{X}_1}^2 \left(\frac{1+\rho_1}{n_1} \right) + \sigma_{\bar{X}_2}^2 \left(\frac{1+\rho_2}{n_2} \right)}. \quad (3.3-4)$$

Substituting into (3.3-4) the appropriate expressions obtained from (3.1-12) and some algebraic rearranging one obtains the general formula for the standard deviation of the difference between two independent means of coupled data. Hence, the resulting corrected t formula is

$$t' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{SS_{X_1}}{n_1} \left(\frac{1+\rho_1}{n_1 - (1+\rho_1)} \right) + \frac{SS_{X_2}}{n_2} \left(\frac{1+\rho_2}{n_2 - (1+\rho_2)} \right)}}, \quad (3.3-5)$$

$v = n_1 + n_2 - 2$. Again, as a check, if $\rho_1 = \rho_2 = \rho$ then (3.3-5) simplifies to (3.3-3), and if $\rho_1 = \rho_2 = 0$ then (3.3-5) simplifies to

$$t' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{SS_{X_1}}{n_1(n_1-1)} + \frac{SS_{X_2}}{n_2(n_2-1)}}},$$

which is the standard unpooled version of Student's t statistic for the two independent samples cases.

Finally, we note that one can place a two-sided confidence interval around $(\mu_1 - \mu_2)$ by using

$$(\bar{X}_1 - \bar{X}_2) \pm t_{v: \alpha/2} \sqrt{\frac{SS_{X_1} + SS_{X_2}}{n(n-1)} \times \frac{(n-1)(1+\rho)}{n-(1+\rho)}}, \quad (3.3-6)$$

for the equal correlation and equal sample size case. Or for the more general case the confidence interval can be obtained with

$$(\bar{X}_1 - \bar{X}_2) \pm t_{v: \alpha/2} \sqrt{\frac{SS_{X_1}}{n_1} \left(\frac{1+\rho_1}{n_1-(1+\rho_1)} \right) + \frac{SS_{X_2}}{n_2} \left(\frac{1+\rho_2}{n_2-(1+\rho_2)} \right)}. \quad (3.3-7)$$

3.4 k-group case

The corrected F formula for the k -group case will be derived in a similar manner as the two-group t -test: equal sample sizes and correlations and then the more general formula. The derivations will focus on the one-way completely randomized design.

The notation is as follows:

$\bar{X}_k$ denotes the sample mean for the k th group,

SS_k denotes the sum of squared deviations from the mean of the k th group,

n_k denotes the sample size for the k th group,

$\hat{\sigma}_{X_k}^2$ denotes the variance estimate for the k th group,

ρ_k denotes the within couplet correlation for the k th group,

n denotes the common sample size, and

ρ denotes the common within couplet correlation.

First, the sums of squares between is given by

$$SS_B = n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2 + \dots + n_k(\bar{X}_k - \bar{X})^2, \quad v = k - 1,$$

where $\bar{X}$ denotes the weighted average

$$\bar{X} = \frac{n_1(\bar{X}_1) + n_2(\bar{X}_2) + \dots + n_k(\bar{X}_k)}{n_T}, \text{ where } n_T = n_1 + n_2 + \dots + n_k.$$

Because it does not incorporate any within-groups deviations, the sums of squares between is not affected by coupled data.

For the equal sample size and equal correlation case, the sums of squares within is

$$SS_w = \frac{(n-1) \left[\sum_{i=1}^k (SS_{X_k}) \right] (1+\rho)}{n - (1+\rho)}, \quad v = n_T - k. \quad (3.4-1)$$

Then, bringing together the means squares between and (3.4-1) and with some algebraic rearranging the corrected F for coupled data is

$$F' = \frac{F}{\left(\frac{(n-1)(1+\rho)}{n - (1+\rho)} \right)}, \quad (3.4-2)$$

where F is the standard F statistic calculated from the coupled data. As a check, if $\rho = 0$ then $F' = F$. Furthermore, it is interesting to note that if

$$c = \frac{(n-1)(1+\rho)}{n - (1+\rho)},$$

then the correction formula for Student's t test is $t' = t / \sqrt{c}$, and the correction for the F test is $F' = F / c$. Since $t_v^2 = F_{1, v}$, the last observation provides a check on the algebra.

Next, we will derive the general formula for the corrected F. Again, the mean square between is not altered by the coupled data; however, the sum of squares within is denoted by

$$SS_w = (n_1 - 1) \hat{\sigma}_{X_1}^2 + (n_2 - 1) \hat{\sigma}_{X_2}^2 + \dots + (n_k - 1) \hat{\sigma}_{X_k}^2, \text{ with } v = n_T - k. \quad (3.4-3)$$

Substituting the appropriate variance estimate -- from combining (3.1-10) and (3.1-12) -- into (3.4-3) the resulting formula for the sums of squares within for coupled data is

$$SS_w^* = \sum_{m=1}^k \left[(n_m - 1) \frac{SS_{X_m} (1 + \rho_m)}{n_m - (1 + \rho_m)} \right], v = n_T - k. \quad (3.4-4)$$

Therefore, the corrected F formula is,

$$F = \frac{MS_B}{SS_w / v}, \quad (3.4-5)$$

where SS_w and v are defined in (3.4-4).

Two points are noteworthy at this juncture. First, a commonly used measure of effect size for the k-group case is the eta-squared statistic. This is computed by dividing the sums of squares between by the sums of squares total. In the case of coupled data the sums of squares total can be computed by summing the sums of squares between and either (3.4-1) or (3.4-4). Second, it is important to note that, although derived for the one-way completely randomized design, the results derived in this section can also be used with completely randomized factorial, repeated measures, and split-plot designs. In all of these cases, when creating an F ratio to test for main effects or interactions the researcher simply uses the correct error term. If that error term involves the mean squares within, then (3.4-4), or some variant of it, can be used.

3.5 Analysis

In this section, the distributional theory for the formulas for coupled data will be examined, and then used to mathematically investigate the resulting type I error rate when Student's t-test is used on coupled data.

In order to investigate whether the formulas derived in this chapter are distributed according to an F distribution, we will apply a theorem from linear algebra and statistics regarding the distribution of quadratic forms. That is, the test statistics in this chapter involve ratios of mean squares or, more generally, quadratic forms that are independently distributed as chi-squared variates with their respective degrees of freedom. Cochran's theorem (1934), as expanded upon by Searle (1971, Sections 2.5 and 2.6; 1982, p. 356), dictates that for any variable $X \sim N(\mu, \Sigma)$: (i) the quadratic form $X'AX \sim \chi^2(r, \delta)$, where $\delta = \mu'A\mu$ and the rank of A equals r , if $A\Sigma$ is idempotent; and (ii) the quadratic forms $X'AX$ and $X'BX$ are distributed independently if $A\Sigma B = 0$. Applying Searle's results to the formulas in this chapter it can be shown that (3.4-2) and (3.4-5) are not distributed as F and since $t_v^2 = F_{1,v}$ hence (3.2-2), (3.3-3), and (3.3-5) are not distributed as t ¹³. These formulas are then approximate t or F tests¹⁴.

¹³ The author would like to thank Dr. D. Roland Thomas for working through Searle's conditions for the quadratic forms, for reminding the author that Walsh's (1947) results regarding the intra-class correlation were, in fact, exact and not asymptotic, and further for bringing to my attention how the approximation works.

¹⁴ The fact that these formulas are approximate tests is rather interesting in light of the fact that the general case of intra-class correlation results in exact tests (Walsh, 1947). Seduced by the principle of deductive logic I had assumed that the results from the general case of intra-class correlation would be applicable for coupled data. A note of gratitude to Drs Jane F. Gentleman and D. Roland Thomas for the probing questions that clarified this matter.

One can gain insight into how the approximation works by noting that the expected values of the variance, (3.1-11), can be expressed as

$$E(S^2) = \left(\sigma_X^2 \times \frac{n-1}{n} \right) \times \left(1 - \frac{\rho}{n-1} \right),$$

and is asymptotically unbiased. This clearly can be seen from the fact that the second term in the above product approaches zero as n approaches infinity. Furthermore, the utility of the corrections to the standard equations outlined in this chapter can be illustrated by considering the two sample case. The strength of the correction comes from adjusting the variance of the difference between means and not by adjusting the estimate of the variance (as is done when correcting for unequal sample sizes in the sample t -test by pooling the variance to improve its estimate).

In summary, the formulas presented in this chapter are approximate tests. What needs to be investigated then is the adequacy of the approximations. Furthermore, all of the derivations in this chapter have assumed a known ρ . The impact of using the Pearson product moment correlation coefficient as an estimate of ρ and substituting this into the approximate tests needs to be investigated. Both of these issues will be addressed in the following Monte Carlo simulation studies.

3.5.1 Mathematical results of the type I error rate

Before proceeding to the Monte Carlo, let me make use of the above results to investigate the effect on the type I error rate of using the standard t -test with coupled data. Given that the formulas presented in this chapter are not exact t or F tests, the following results are only approximations. However, given the good performance of these tests in the Monte Carlo simulation the results in this section should be indicative of the behavior of the type I error rates. For expository purposes, without loss of generality, let us consider

the one-sided confidence intervals computed for the population mean, μ . Given the i.i.d. assumption, the one-sided confidence interval for large samples is denoted by

$$\left(\bar{X} + Z_{\epsilon} \sqrt{\frac{SS_X}{n(n-1)}}, \infty \right),$$

where the only new notation is Z_{ϵ} which denotes the standard normal deviate exceeded by probability ϵ . Thus, the one-sided confidence interval can also be denoted by

$$\Pr\left(\bar{X} + Z_{\epsilon} \sqrt{\frac{SS_X}{n(n-1)}} < \mu\right) = \epsilon. \quad (3.5.1-1)$$

For small samples, then, the one-sided confidence interval is denoted by

$$\Pr\left(\bar{X} + t_{v:\epsilon} \sqrt{\frac{SS_X}{n(n-1)}} < \mu\right) = \epsilon, \quad (3.5.1-2)$$

where $t_{v:\epsilon}$ equals the 100(ϵ) percentile of the t distribution with v degrees of freedom.

Now, given coupled data, (3.5.1-1) and (3.5.1-2) can be rewritten as

$$\Pr\left(\bar{X} + Z_{\alpha} \sqrt{\frac{SS_X}{n(n-1)} \times \frac{(n-1)(1+\rho)}{n-(1+\rho)}} < \mu\right) = \alpha, \quad (3.5.1-3)$$

and,

$$\Pr\left(\bar{X} + t_{v:\alpha} \sqrt{\frac{SS_X}{n(n-1)} \times \frac{(n-1)(1+\rho)}{n-(1+\rho)}} < \mu\right) = \alpha, \quad (3.5.1-4)$$

where Z_{α} denotes the standard normal deviate, and $t_{v:\alpha}$ denotes the t value exceeded by probability α . It should be noted that α is the nominal level of the test and ϵ is the actual level achieved due to not accounting for covariance.

Setting (3.5.1-1) equal to (3.5.1-3), and (3.5.1-2) equal to (3.5.1-4) results in,

$$Z_{\alpha} = Z_{\epsilon} \sqrt{\frac{n - (1 + \rho)}{(n - 1)(1 + \rho)}}, \quad (3.5.1-5)$$

and

$$t_{v : \alpha} = t_{v : \epsilon} \sqrt{\frac{n - (1 + \rho)}{(n - 1)(1 + \rho)}}, \quad (3.5.1-6)$$

respectively. If the n observations are i.i.d., then $\rho = 0$ and $\alpha = \epsilon$. It is evident thus, that if $\rho \neq 0$, then α can be quite different than ϵ . The contents of Table 1 are values of the type I error rate for the one-sided confidence interval applied to coupled data, with the nominal alpha set at .05. The type I error rate was evaluated from (3.5.1-5) for $n = 4, 8, 16, 32, 64$, and 127 , and (3.5.1-6) for $n = 10,000$. These values of n were chosen to represent (a) a broad range of small sample sizes and (b) a very large sample size whose results are very close to the asymptotic values. The values of the within couple correlation, ρ , range from $-.90$ to $+.90$ in increments of $.10$. Figure 4 is a graphical depiction of the relationship between the type I error rate and ρ for a small and for a very large sample size.

It should be noted that the elements of Table 1 are the complement of the cumulative density function of the central standard normal (or central t with v degrees of freedom) for the resulting Z_{α} (or $t_{v : \alpha}$) from (3.5.1-5) [or (3.5.1-6) for the t]. The values of the cumulative density function for the central distributions were computed from an algorithm in the MINITAB® computer program.

Figure 4 depicts the Table 1 values for sample sizes of 4 and $10,000$ (the asymptotic values) and will be used in conjunction with Table 1. In Figure 4, the horizontal line traces the nominal type I error rate and the vertical line traces $\rho = 0$. The intersection of these horizontal and vertical lines in Figure 4 is the type I error rates for the i.i.d. case. Figure 1 shows that a positive correlation results in an increase in the type I error rate whereas a negative correlation results in a decrease in the type I error rate.

Table 1. Type I error rates of the as a function of the correlation among the elements of the couple, $\alpha = .05$.

ρ	n						
	4	8	16	32	64	128	10,000
-.900	.001	.0002	.0000	.0000	.0000	.0000	.0000
-.800	.004	.001	.001	.0003	.0002	.0002	.0002
-.700	.008	.004	.003	.002	.002	.002	.001
-.600	.012	.008	.006	.003	.005	.005	.005
-.500	.016	.013	.012	.009	.010	.011	.010
-.400	.022	.020	.018	.017	.017	.018	.017
-.300	.028	.027	.026	.025	.025	.026	.025
-.200	.035	.034	.034	.033	.033	.034	.034
-.100	.042	.042	.042	.041	.041	.043	.042
0.00	.050	.050	.050	.050	.050	.050	.050
.100	.058	.058	.058	.058	.058	.060	.059
.200	.067	.066	.067	.066	.066	.068	.067
.300	.077	.075	.075	.074	.074	.076	.075
.400	.087	.083	.082	.082	.082	.084	.082
.500	.097	.091	.090	.090	.089	.091	.090
.600	.109	.099	.098	.097	.096	.098	.097
.700	.121	.107	.105	.104	.104	.105	.104
.800	.134	.115	.112	.111	.110	.112	.111
.900	.147	.123	.119	.117	.117	.118	.117

Figure 4. Type I error rates of the as a function of the correlation among the elements of the couple, $n = 4$ and 10,000.

Table 1 shows that this finding is consistent across sample sizes and with the asymptotic findings. Although the type I error rate is substantially inflated for the asymptotic case, it is not as inflated in the case of positive correlation as it is for very small sample sizes. In the case of a negative correlation, the type I error is more deflated for the asymptotic case than it is for the small sample case. These mathematical results agree with Coren and Hakstian's (1990) assertion that the type I error rate should be inflated to approximately .140 with coupled data for a nominal alpha of .05.

Two points are noteworthy at this juncture. First, the case of negative correlation, although mathematically interesting, will probably not be highly likely in applied settings. That is, most of the cases of coupled data come from lateral measures (with the exception of the married couples example) where the correlation between the elements of couple are expected to be positive. This does not, however, preclude the possibility of negative correlation occurring. Second, although the type I error results are presented for the one-sided confidence interval they could easily generalize to the two-sided confidence interval and to hypothesis testing for more than one group. Results for these other cases can be easily derived in the same manner as the one-sided confidence intervals.

CHAPTER 4: MONTE CARLO INVESTIGATION

This chapter marks the end of the mathematical investigation into formulas to correct for nonindependence in inference, and a turn to the empirical (Monte Carlo) study of the analytical results. Interestingly, the derivations in Chapter 3 reflect how Student (W.S. Gossett) or R. A. Fisher might have approached the problem of coupled data. The formulas are generalizations of the methodology first formally presented by Student (1908; 1927) and later further developed by Fisher (1926) and others. Student broke new ground by mathematically deriving a crude form of the t distribution, after applying empirical sampling to 3000 measurements of finger lengths and heights to verify the sampling distribution (Irwin, 1978). Fisher (1926) verified Student's guess seventeen years later.

Like Student, we will now study the sampling distribution of our formulas via empirical methods. Student used real data from which he repeatedly sampled; however, we used simulated data. We examined the distributions by investigating the type I error rates of the formulas. This chapter defines the purpose, limitations and methodology of the Monte Carlo investigation.

4.1 Purpose

The essence of the Monte Carlo technique is to design an algorithm (or game) which models the problem. Any one replication (play of the game) will result in an outcome which will be determined by the relevant parameters in the algorithm. The Monte Carlo technique used in this investigation was designed to empirically estimate the true type I error rates and power functions. The purposes of the Monte Carlo investigation were to study the type I error rate, type II error rate and power of :

- (a) Student's t -test applied to coupled data,
- (b) the statistical tests presented in Chapter 3 with both population and sample values for the correlation,
- (c) the alternative proposed by Coren and Hakstian (1990),
- (d) the use of randomization tests with coupled data, and
- (e) the yet unexplored couplet referenced randomization test.

Where appropriate (i.e., where type I error rate is maintained at its nominal level) power comparisons were made among the results of Chapter 3 and the Coren and Hakstian solution, and between the results of Chapter 3 and the randomization tests.

The results of this investigation provide evidence to reassure researchers of the adequacy of using the formulas presented in Chapter 3 (i.e., their sampling distributions are, in fact close to those of the t - or F - distribution), particularly in the case of small samples and a sample estimate of ρ . Furthermore, the solutions commonly found in psychology were explored for their adequacy and how they compare to the formulas in Chapter 3. The relative performance of each of the alternatives should aid the researcher in choosing the best statistic for significance testing in the coupled data situation. Currently, little is known of how Coren and Hakstian's solution, randomization tests, and the results derived in Chapter 3 perform for the coupled data situation.

4.2 Assumptions and Limitations

1. The assumption of the data generation procedure is that the data are empirical realizations of a Gaussian or normal distribution. This is a commonly held assumption in data analysis within the field of psychology. We have chosen to restrict our attention to the

Gaussian distribution so that the results do not confound nonnormality with nonindependence.

2. In this investigation, only balanced completely randomized designs were considered.

3. The results of this study are restricted to the statistical solutions having their basis in the psychological literature. The other solutions presented in Chapter 2 will not be examined empirically.

4. The conclusions drawn from the study are limited to the extent that the Monte Carlo results are representative of what occurs in the research context. That is, because we are considering completely randomized designs, the results generalize to the context of coupled data where units are randomly assigned to treatment conditions. Furthermore, we only consider the case of positive ρ .

4.3 Methodology

Three studies were involved in this investigation. Study 1 investigated the behavior of the normal theory tests with coupled data. That is, we examined what happens to the type I and II error rates of the commonly found t and F tests with coupled data. Study 2 investigated the type I error rate and power values for the modified t and F formulas, presented in Chapter 3, and compared their power to Coren and Hakstian's (1990) proposal. Furthermore, we investigated the performance of the modified t and F formulas when sample estimates of the correlation were used instead of the known population values. Study 3 investigated the type I and II error properties of the randomization and the couplet referenced randomization test with coupled data.

Since there are three studies, we will first present the methodology that is common to each of the studies and then treat each study separately.

4.3.1 Computer simulation methodology

Random samples from the standard normal distribution were generated based on pseudorandom numbers. These normal deviates were generated using the method of Box and Muller (1958). The relation used was $X = \sqrt{-2\log_e X_1} (\cos 2\pi X_2)$, where X_1 and X_2 were pseudorandom uniform numbers on the interval $[0, 1]$. The uniform random number generator used in this investigation was the thoroughly tested prime-modulus multiplicative congruential generator described by Lewis and Orav (1989) and Lewis, Goodman, and Miller (1969). Furthermore, the adequacy of the random number generator was evaluated using tests by Lehman (1977), Morgan (1984), and Ripley (1987). The uniform generator met the accepted requirements of rectangularity, sequential independence, and lack of patterns in a sequence of numbers.

The type I error rate and power was calculated for the nominal α on data from independent populations of homogeneous variance. For studies 1 and 2 the number of replications were 5000. For study 3, involving the randomization tests, the number of replications was 2000. The smaller number of replications was chosen for the third study because of the excessive computational time required to perform the simulations. For example, the randomization test for two independent samples with a sample size of ten per group involves 184,756 statistical tests for each replication. We have chosen not to use the approximate randomization test because the complete randomization distribution allowed us to discount any possible effects of the approximation. See Zumbo and Zimmerman (in press) for a discussion of the possible biasing effects of approximate randomization tests.

For each replication, n_1 values were generated according to the specified distribution then n_2 values were generated from the same distribution. These sample values were then transformed to induce a variance-covariance pattern given in Matrix B of Figure 2. Correlations were induced by adding a multiple of one random variable to another

random variable. The multiplicative constant was chosen to obtain the desired correlation coefficient, and the variables were further transformed to have mean 0 and variance 1. More specifically, if Z , W_1 , and W_2 are independent random variables with mean

0 and variance 1, then $X = (W_1 + bZ) / \sqrt{1 + b^2}$ and

$Y = (W_2 + bZ) / \sqrt{1 + b^2}$, where $b = \sqrt{r/(1 - r)}$, are dependent random variables related by $\rho(X, Y) = r$, $E(X) = E(Y) = 0$, and $\sigma_x^2 = \sigma_y^2 = 1$.

The type I error rate was the proportion of replications in which the statistical test was significant. Since one of the objectives of this investigation was to study the performance of the tests under the alternative hypothesis of population differences, standardized differences between the means were created. This standardized difference is commonly referred to as delta, also denoted as δ (see, Hays, 1988, p. 304). If the type I error rate was maintained at its nominal level then the proportion of statistically significant tests is an empirical estimate of the power. If the type I error rate was not maintained at its nominal level then the proportion of significant differences represents $1 - \text{Pr}(\text{type II error})$, or equivalently $1 - \beta$.

4.3.2 Performance indices

There is no generally accepted, standard, quantitative definition of what constitutes robustness. Bradley (1978) suggests that a liberal definition of a robustness criterion is

$$|\hat{\alpha} - \alpha| \leq \frac{\alpha}{2}, \text{ where } \hat{\alpha} \text{ and } \alpha \text{ are the empirical and nominal type I error rates, respectively.}$$

Bradley's criterion is arbitrary and does not incorporate the possibility for power comparison. Rather than Bradley's approach, this author advocates a statistically-based methodology to compare the performance of two or more statistical tests.

Finally, we will consider the performance indices used in analysing the Monte Carlo results. The entire process of the Monte Carlo simulation can be considered an empirical estimate, $\hat{p}$, of a population parameter, p , where the obtained proportion is a sample point estimate. The replications can be considered as a binomial experiment.

Two types of questions were asked with the Monte Carlo studies: (1) is the empirical type I error rate within sampling error of its nominal value?, and (2) do two statistical tests have the same power? The first question was answered by using the confidence interval of a proportion, given by a normal approximation as

$$\hat{p} \pm z_{1-\alpha/2} (SE_{\hat{p}}), \text{ where } SE_{\hat{p}} = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{\# \text{ of replications}}}, \text{ and}$$

$SE_{\hat{p}}$ denotes the standard error of the proportion. If the nominal type I error rate was within the confidence interval then the empirical type I error rate was within sampling error of its nominal value. In examining the type I error rate a 95% confidence interval was used. It should be noted that the confidence interval methodology is simply a standard hypothesis test.

The second question was answered by computing the confidence interval for the difference between two proportions. Due to the fact that for any replication the same sample was used for both tests, the confidence interval was calculated for correlated proportions. The normal approximation to this confidence interval is given by

$$\hat{p}_{\text{diff}} \pm z_{1-\alpha/2} (SE_{\hat{p}}), \text{ where } SE_{\hat{p}} = \sqrt{\frac{1}{\# \text{ of replications}} [(\hat{p}_1 + \hat{p}_2) - (\hat{p}_1 - \hat{p}_2)^2]},$$

where $\hat{p}_{\text{diff}} = \hat{p}_1 - \hat{p}_2$. Then, if zero was within the confidence interval, the two power values were not different. It is important to note, however, that the method of confidence

intervals does not allow one to prove the equality of the two parameters. That is, the confidence interval method simply lends support for the equality.

4.3.2 Study 1

Because the purpose of this study was to investigate the performance of Student's *t*-test and the *F* test applied to coupled data, the following simulations were conducted:

- (a) Power functions for Student's *t*-test for independent samples were estimated for a delta of zero to four in increments of .50. A power function was obtained for each of the combinations of $n = 8, 32$, $\alpha = .01, .05, .10$, and $\rho = 0, .30, .60$, and $.90$. Furthermore, a large sample estimate of the power function were obtained from $n = 500$ with an $\alpha = .05$.
- (b) Next the *F* statistic was examined by computing empirical power functions for a delta of zero to four in increments of .50. The values of the correlation, ρ , considered were 0, .60, and .90. Furthermore, where k denotes the number of groups, the power functions were computed for $k=4, n=30, \alpha=.05$, $k=4, n=12, \alpha=.05$, and finally, $k=3, n=30, \alpha = .01$.

The type I error rates were analysed using the confidence interval for a proportion. Furthermore, in light of the fact that Coren and Hakstian (1990) clearly state that the type I error rate will increase above its nominal value when ρ is nonzero, we were not interested in statistically comparing the power functions for nonzero ρ . Note that it is incorrect to treat the values of $\delta \neq 0$ as power when the type I error rate is not maintained. However, these functions can be used to investigate the type II error rates. Given that $\beta = 1 - \text{Power}$, then when the power is large β is obviously small. We were not interested in statistically comparing the power functions when the type I error rate was not maintained. Rather, the functions were treated in an exploratory graphical manner looking for consistent trends in the empirical type I and II error rates.

4.3.3 Study 2

As in study one, this study examined both the t and F formulas. The results from the t formula were an analysis of the Zumbo and Zimmerman (1991) results with the present methodology. The simulations involved the Student's t -test, Formula (3.3.3), Formula (3.3.3) with a sample estimate of ρ^{15} , and Coren and Hakstian's suggestion of averaging across the ears. These tests are denoted t , t' , t'' , and t_{av} , respectively. The corresponding formulas for the F test are F , F' , F'' , and F_{av} , respectively.

The following simulations were conducted:

For the t test:

- a. The type I error rates of the four statistics for $n = 12, 20$, and for $\rho = 0$ to $.90$ in increments of $.10$.
- b. The power functions of the t' , t'' , and t_{av} formulas for a delta of zero to four in increments of $.5$ with $n = 12, \rho = .40, .60$, and $n = 16, \rho = .75$.

For the F test:

- a. The type I error rates of the four statistics for $n = 16, 32$, and for $\rho = 0$ to $.90$ in increments of $.10$.
- b. The power functions of the F' , F'' , and F_{av} formulas for a delta of zero to four in increments of $.5$ with $k = 4, n = 12, \rho = .60, .90, \alpha = .05$, and $k = 4, n = 30, \rho = .60, .90, \alpha = .05$, and finally $k = 3, n = 30, \rho = .60, .90, \alpha = .01$.

¹⁵ Given that we are dealing with coupled data the sample estimate is the Pearson product moment correlation coefficient.

The first simulations for the t and F statistics were interested in further establishing the relationship between the type I error rate and p . This relationship was examined by a graphical plot of the type I error rates for each value of p . The second question for the F test was analysed with confidence intervals of a proportion.

The last questions for the t and F statistics involve a statistical comparison of the power functions. We used a liberal statistical criterion wherein each of the 9 points of the power function as tested individually and furthermore apriori comparisons were planned between (i) t' versus t'' and t' versus t_{av} , and (ii) F' versus F'' , and F' versus F_{av} . Given the apriori comparisons, each confidence interval was computed with Dunn's simple suggestion to compute the confidence interval for each difference at 97.5% (Marascuilo & Serlin, 1988, pp. 371-372). This liberal statistical analysis of the power functions is a large step forward from the more heuristic descriptions of comparative power which are pervasive in behavioral and social science statistics (e.g., Zimmerman & Zumbo, in press a; Zumbo & Zimmerman, 1991).

4.3.4 Study 3

This last study was an investigation into the use of randomization tests and coupled referenced randomization tests with coupled data. This study is particularly important in light of Zumbo and Zimmerman's (1991; footnote 1) suggestion that the randomization test is a possible alternative with coupled data. The only test investigated was Student's t -test; however, the results are generalizable to the F test. Due to the excessive computational time¹⁶, the results of this study were limited to the small sample case.

The following simulations were conducted:

¹⁶ A randomization test for the two-group randomized design with 16 cases per group would require over 600 million t -tests per replication!

1. Type I error rates of the randomization test with $p = 0, .20, .40, .60, .80, .90$. Sample sizes of 4, 6, 8, and 10 were used. For the sample sizes of 4, 6, 8, and 10 $\alpha = .05$.
2. Type I error rates of the couplet referenced randomization test with $p = 0, .20, .40, .60, .80, .90$. Sample sizes of 4, 6, 8, and 10 were used. All tests were conducted at an $\alpha = .05$ except for $n=4$ where the natural limit was at $.1667$ ($1/6$)¹⁷.
3. The power functions of formula (3.3.3) and the couplet randomization test for a delta of zero to three in increments of 1.5. Sample sizes of 8, 10, 12, 16, $p = .70$ and $\alpha = .05$ were used.

The first two simulations were analysed by using the one sample confidence interval for the proportion. Because it involved only one comparison, the last simulation was analysed using a similar methodology as the comparative analyses in Study two except that the confidence intervals of the difference were computed at 95% not 97.5%, as in study two.

¹⁷ The natural significance level is computed by $1/(\text{number of permutations})$.

CHAPTER 5: RESULTS & DISCUSSION

5.1 Study 1

The results of the computer simulation are shown in Tables 2 and 3, and Figures 5

Table 2. Type I error rates for the t-test for various values of alpha.

α	N	ρ			
		0	.30	.60	.90
.01	8	.009	.019*	.042*	.055*
	32	.012	.016*	.031*	.046*
.05	8	.052	.079*	.112*	.128*
	32	.047	.073*	.096*	.130*
	500	.046	.075*	.102*	.144*
.10	8	.100	.131*	.171*	.193*
	32	.102	.139*	.164*	.184*

Note: A * indicates that the nominal alpha was not within the confidence interval.

Table 3. Type I error rates for the F-test for various values of alpha.

			ρ		
			0	.60	.90
$\alpha=.05$					
n = 12	k = 4		.052	.185*	.289*
n = 30	k = 4		.049	.180*	.268*
$\alpha=.01$					
n = 30	k = 3		.008	.075*	.093*

Note: A * indicates that the nominal alpha was not within the confidence interval.

through 14. Table 2 shows the type I error rates for the t-test for various values of the nominal type I error rate (denoted alpha or α) and correlation. Table 3 provides similar information for the F-test with the exception that k denotes the number of groups.

Given our definition of robustness of type I error rate in section 4.3.2, it is apparent from Tables 2 and 3 that a non-zero correlation (.30 - .90 for the t-test, and .60 - .90 for the F-test) results in tests that are not robust for type I error rate. Furthermore, it should be noted that as the correlation increases, the type I error rate increases. These findings are consistent with the mathematical results in Chapter 3; therefore they not only provide small sample results but also lend support for the validity of our simulation methodology.

Figure 5. Power functions of Student's t-test for various values of correlation ($n=8$, $\alpha=.01$).

Figures 5 through 14 are the empirical functions relating the effect size to the probability of rejecting the null hypothesis. Each figure is constructed such that the horizontal axis represents the effect size δ and the vertical axis represents the probability measure. If the probability at $\delta = 0$ is the nominal type I error rate, the function represents the power function. If the probability at $\delta = 0$ is not the nominal type I error then the function indicates the magnitude of the type II error rate (which is, in fact, the complement of the plotted value).

With regard to the effect of the correlation, it should be noted that a non-zero correlation results in an elevation of the functions over the value at $\rho = 0$. This finding was consistent for all of the sample sizes, nominal alphas, and statistical tests investigated. Furthermore, the elevation is more pronounced the larger the magnitude of the correlation. This indicates that an increase in ρ results in a decrease in type II error for all values of δ considered in this study. This is not so surprising given the well known inverse relationship of type I and type II error. Thus, Study 1 not only confirms the mathematical results of the type I error rate but also indicates that the type II error will be substantially deflated over its i.i.d. value if the standard statistical tests are used with coupled data.

Figure 6. Power functions of Student's t-test for various values of correlation ($n=8$, $\alpha=.05$).

Figure 7. Power functions of Student's t-test for various values of correlation ($n=8$, $\alpha=.10$).

Figure 8. Power functions of Student's t-test for various values of correlation ($n=32$, $\alpha=.01$).

Figure 9. Power functions of Student's t-test for various values of correlation ($n=32$, $\alpha=.05$).

Figure 10. Power functions of Student's t-test for various values of correlation ($n=32$, $\alpha=.10$).

Figure 11. Power functions of Student's t-test for various values of correlation ($n=500$, $\alpha=.05$).

Figure 12. Power functions for the F-test for various values of correlation ($k=4$, $n=30$, $\alpha=.05$).

Figure 13. Power functions for the F-test for various values of correlation ($k=3$, $n=30$, $\alpha=.01$).

Figure 14. Power functions for the F-test for various values of correlation ($k=4$, $n=12$, $\alpha=.05$).

5.2 Study 2

The purpose of this study was to (a) examine the formulas in Chapter 3 with both population and sample values of ρ , and (b) compare these formulas to Coren and Hakstian's solution.

It should be recalled that in the mathematical analysis, the type I error rate increased as a function of positive correlation. The first set of Monte Carlo simulations investigated whether this relationship still held when the Chapter 3 formulas (t' , t'' , F' , and F'' denote the t and F tests with population and sample values for ρ , respectively) and the Coren and Hakstian (t_{av} , F_{av}) solution were used with coupled data.

Figures 15 through 18 and Table 4 contain the results of this first investigation. The figures were constructed such that the horizontal axis is the magnitude of the correlation and the vertical axis is the probability measure. Figures 15 and 16 present the results of the t -tests while Figures 17 and 18 present the results of the F -tests. The results are consistent for both tests and for all of the sample sizes considered in this study. First, the inflation in type I error can be seen again in all of the figures. However, it is important to note that t' , t'' , F' , F'' , t_{av} , and F_{av} all have a relatively horizontal function. This shows that the formulas in Chapter 3 for both population and sample values of ρ , and the Coren and Hakstian solution all maintain the nominal type I error rate. This will be further examined in the following section using the Monte Carlo definition of robustness of type I error in section 4.3.2. That is, in comparing power functions we are also considering the type I error.

Figure 15. Type I error rates of the t-test as a function of the within couplet correlation ($n=12$).

Figure 16. Type I error rates of the t-test as a function of the within couplet correlation ($n=20$).

Figure 17. Type I error rate of the F-test as a function of the within couplet correlation ($k=3$, $n=32$).

Figure 18. Type I error rates of the F-test as a function of the within couplet correlation ($k=3$, $n=16$).

To this point, in this study, it has been shown that the type I error rate is maintained at its nominal level for: (a) the statistics derived in Chapter 3 using the population parameter ρ , (b) the statistics derived in Chapter 3 using the sample estimate of ρ , and (c) the Coren and Hakstian approach. The next issue, then, is a comparison of the power characteristics for these three statistics.

It should be noted that no absolute standard exists for a power comparison. Rather, in Chapter 4 I proposed a new methodology based on confidence intervals of planned paired comparisons. The comparisons are: (a) t' with t'' and F' with F'' , and (b) t' with t_{av} and F' with F_{av} . Tables 4 through 7 present the results of these comparisons.

Table 4. Type I error rates and power of the two independent samples t test, $\alpha = .05$.

n	ρ	δ	t'	t''	t_{av}
12	.40	0	.054	.052	.050
		.5	.078	.078	.075
		1.0	.163	.156	.151
		1.5	.303	.285	.278*
		2.0	.476	.447	.436*
		2.5	.653	.617*	.601*
		3.0	.806	.773	.762*
		3.5	.908	.882	.870
		4.0	.963	.950	.943
12	.60	0	.051	.053	.050
		.5	.082	.082	.073
		1.0	.185	.177	.158*
		1.5	.359	.344	.321*
		2.0	.564	.538	.509*
		2.5	.740	.716	.692*
		3.0	.879	.859	.839
		3.5	.951	.941	.930
		4.0	.984	.979	.975
16	.75	0	.058	.060	.052
		.5	.107	.110	.095
		1.0	.264	.257	.239*
		1.5	.492	.480	.457*
		2.0	.715	.709	.687
		2.5	.884	.875	.860
		3.0	.969	.964	.959
		3.5	.993	.993	.990
		4.0	.999	.999	.998

Note: * indicates that the confidence interval of the difference from t' did not contain zero.

Table 4 presents the results for the t -tests for sample sizes of twelve and sixteen and correlation values of .40, .60, and .75. Those values marked with an "*" are the instances where there was a significant difference from the value at t' . When comparing t' to t'' , there was only one significant difference found. This difference was found at a δ value of 2.5 with a sample size of 12 and a ρ of .40. In this case, the population-based t -test was

more powerful than the sample-based t -test. Overall, however, t' did not differ from t'' . When comparing t' to t_{av} , however, there were consistent differences in the range of $\delta = 1.0$ to 2.0 . That is, generally, t' was more powerful than t_{av} in the range of $\delta = 1.0$ to 2.0 . Furthermore, coupling this second finding with the one for the comparison of t' with t'' , one can tentatively conclude that for $\delta = 1.0$ to 2.0 , t'' will be more powerful than t_{av} . These findings are particularly important for applied researchers because many of the commonly found effect sizes are in this range (Cohen, 1988).

Next, F' , F'' , and F_{av} were compared in a manner similar to the t -test comparisons. Tables 5 through 7 present the power results for the F -tests conducted in Study 1. The findings are quite consistent with those of the t -tests. That is, there were no differences found when comparing F' and F'' . However, when differences were found they were between F' and F_{av} , and in all cases F' was favored. Again, the differences were in the ranges of commonly found effect sizes.

It is not surprising that the t' and F' methods are more powerful than the Coren and Hakstian methodology because the latter solution halves the effective sample size, thereby reducing the degrees of freedom and the probability of detecting a difference when one exists. Future studies should investigate further the role of sample size and the magnitude of the correlation on the difference between t' and t_{av} (also F' and F_{av}). That is, at what point does the difference between the Chapter 3 formulas and Coren and Hakstian's methodology become significant. The preliminary results in this study suggest that the difference will be most pronounced for small samples.

Table 5. Type I error rates and power of the F test for $k=4$, $n=12$, $\alpha=.05$.

ρ	δ	F'	F''	F_{av}
.60	0	.056	.056	.054
	.5	.081	.081	.076
	1.0	.174	.167	.157
	1.5	.344	.324	.308*
	2.0	.566	.538	.514*
	2.5	.777	.749	.729*
	3.0	.909	.891	.877
	3.5	.976	.965	.959
	4.0	.994	.992	.990
.90	0	.060	.060	.059
	.5	.098	.096	.081*
	1.0	.197	.195	.169*
	1.5	.402	.395	.348*
	2.0	.632	.625	.581*
	2.5	.838	.832	.799
	3.0	.952	.950	.931
	3.5	.99	.989	.983
	4.0	.997	.997	.996

Note: * indicates that the confidence interval for the difference from F' did not contain zero.

Table 6. Type I error rates and power of the F test for $k=4$, $n=30$, $\alpha=.05$.

ρ	δ	F'	F''	F_{av}
.60	0	.051	.051	.050
	.5	.119	.116	.114
	1.0	.380	.376	.369
	1.5	.719	.709	.704
	2.0	.946	.939	.936
	2.5	.994	.994	.994
	3.0	1	1	1
	3.5	1	1	1
	4.0	1	1	1
.90	0	.053	.053	.051
	.5	.150	.150	.140
	1.0	.460	.456	.440
	1.5	.816	.816	.804
	2.0	.979	.977	.975
	2.5	1	1	.999
	3.0	1	1	1
	3.5	1	1	1
	4.0	1	1	1

Note: There were no significant differences.

Table 7. Type I error rates and power of the F test for $k=3$, $n=30$, $\alpha=.01$.

ρ	δ	F	F''	F_{av}
.60	0	.013	.012	.011
	.5	.045	.045	.039
	1.0	.220	.214	.196*
	1.5	.557	.572	.551
	2.0	.891	.881	.870
	2.5	.987	.985	.982
	3.0	.999	.999	.999
	3.5	1	1	1
	4.0	1	1	1
.90	0	.01	.01	.008
	.5	.047	.046	.039
	1.0	.237	.235	.214*
	1.5	.617	.612	.581*
	2.0	.911	.907	.893
	2.5	.990	.989	.987
	3.0	1	1	.999
	3.5	1	1	1
	4.0	1	1	1

Note: * indicates that the confidence interval for the difference from F' did not contain zero.

5.3 Study 3

This final study investigated the statistical performance of the randomization test and the couplet referenced randomization test. This study is particularly important given that Zumbo and Zimmerman (1991) recommended the use of the randomization test for coupled data.

Table 8 presents the results of the investigation into the type I error of the randomization test and the couplet referenced randomization test with coupled data. First, focusing on the randomization tests, it is apparent that correlation values greater than .20 result in a substantial inflation of the type I error rate. Second, the couplet randomization test generally performs well for sample sizes of 8 and 10, with the noted exception of $n=8$ when $\rho = .80$. However, the couplet randomization test performs dismally for sample sizes of 4 and 6. This poor performance may be accounted for by the fact that the sample space for the referent distribution is quite small compared to the larger sample sizes. That is, for $n = 4$ there are 6 possible permutations of the couplets (i.e., the sampling distribution is made up of the permutation of 4 couplets). Furthermore, for $n=6$ the situation is not much better where only 20 elements make up the sample space. It appears, then, that as suggested by Harwell's (1991) findings, the randomization test should not be used with coupled data. However, the couplet referenced randomization test is useful for sample sizes of 8 and larger.

Given the above findings regarding the type I error rate of the couplet randomization test, a study of the power comparison between the couplet randomization test, denoted t_{couplet} , and t' was conducted. The findings are presented in Table 9. First, it should be noted that for a sample size of 8, the t' test is more powerful than the couplet randomization test. Second, this power advantage quickly diminishes to nonsignificance as

the sample size increases. Again, it is important to note that when differences occurred they were in favor of t' .

Table 8. The empirical type I error rate for the randomization and couplet referenced randomization test at various value of the within couple correlation, $\alpha=.05$.

ρ	$n=4$		$n=6$	
	Randomization	Couplet	Randomization	Couplet
0	.044	0*	.051	0*
.20	.056	0*	.060	0*
.40	.075*	0*	.088*	0*
.60	.107*	0*	.096*	0*
.80	.134*	0*	.126*	0*
.90	.148*	0*	.150*	0*
<hr/>				
ρ	$n=8$		$n=10$	
	Randomization	Couplet	Randomization	Couplet
0	.051	.053	.050	.049
.20	.062*	.046	.059	.044
.40	.088*	.042	.105*	.055
.60	.118*	.043	.130*	.046
.80	.150*	.040*	.133*	.058
.90	.146*	.046	.150*	.057

Note: * indicates that the nominal alpha was not within the confidence interval.

Table 9. Type I error rates and power of the t' and couplet-randomization tests, $\alpha=.05$.

n	ρ	δ	t'	t_{couplet}
8	.70	0	.058	.044
		1.5	.387	.310*
		3.0	.845	.739*
10	.70	0	.046	.042
		1.5	.428	.381*
		3.0	.900	.855
12	.70	0	.052	.047
		1.5	.517	.489
		3.0	.952	.940
16	.70	0	.047	.050
		1.5	.643	.600
		3.0	.982	.982

Note: * indicates that the confidence interval for the difference from t' did not contain zero.

5.4 Summary of the findings.

The findings of the Monte Carlo simulations can be summarized in four parts. First, the mathematical relationship seen in section 3.5.1 was confirmed. That is, the type I error rate was shown to increase with increasing correlation. This finding lends face validity to the Monte Carlo methodology. Furthermore, it was shown that the type II error rate decreases with increasing correlation.

of/de

**PM-1 3 1/2" x 4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT**

Second, the formulas derived in Chapter 3 were shown to maintain the nominal type I error rate when either the population or the sample estimate of p was used in the equations. Also, in regards to the comparative power it was shown that, generally, no power was lost or gained by using the sample-based formula. This last finding is particularly important for the application of the new formulas from Chapter 3.

Third, the results regarding Coren and Hakstian's (1990) recommendation to average (or transform the two scores by some other linear function of them) were quite interesting; however, perhaps in hindsight predictable from theory. That is, averaging across the two elements of the couple maintains the nominal type I error rate and is as powerful as the approach advocated in Chapter 3, except, for small sample sizes in the range of 1.0 to 2.0 standard units of effect size where the formulas from Chapter 3 are more powerful. In fact, the findings regarding type I error rate and statistical power were predicted in section 2.2.1.

Finally, the results regarding the randomization test indicated that the recommendation by Zumbo and Zimmerman (1991) would lead to an inflated number of false rejections of the null hypothesis, well over nominal alpha. Furthermore, it was shown that the couplet referenced randomization test performs adequately for a sample size of eight or larger per group. Regarding the comparative power of the couplet randomization test to the t' test, it was shown that for small sample sizes the t' test is more powerful than the couplet-randomization test. This finding is particularly interesting given the inordinate computational burden of randomization tests. That is, if the couplet referenced randomization test is less powerful than or as powerful as the t' test then perhaps it is not worth going through an enormous amount of computation to get a p -value. One can simply rely on the standard model-based approach of the t and F distributions.

CHAPTER 6: EXAMPLES

As I stated in Chapter 1, the primary purpose of this dissertation was to present a new set of formulas (Chapter 3) for the analysis of coupled data. Given that their distributional properties have now been shown to be good approximations to t or F distributions, this chapter will present some examples of how to use the approximate tests in Chapter 3. All tests conducted in this section are two-sided.

Example 1: Fictitious data.

This data can be found in Zumbo and Zimmerman (1991). The data, listed in Table 10, was simulated (with parameters from Coren & Hakstian, 1990) to represent two independent groups with 6 subjects per group. Each group, however, has twelve observations -- one from each ear of each subject. Zumbo and Zimmerman reported an approximate $t'(22) = -2.1726$, $p < .05$ with the population value of $\rho = .80$ in Equation (3.3-3). Furthermore, assuming an equal, but unknown, value of ρ for both groups they computed $t'(22) = -2.1116$, $p < .05$. The sample estimate of ρ was computed by the commonly used Pearson product moment correlation coefficient.

For the present purpose, the Zumbo and Zimmerman data will be analyzed in a more realistic context (i.e., ρ is neither known nor assumed to be equal for the two groups). The approximate t' statistic used in this context is Equation (3.3-5). The sum of squares for groups 1 and 2 are 9.278964 and 9.726794, respectively. The Pearson product moment correlation between the pairs of ears of groups 1 and 2 is 0.950 and

Table 10. Data for Example 1.

Subject	Group 1		Subject	Group 2	
	right	left		right	left
1	15.649	15.900	7	12.630	12.381
2	13.654	13.876	8	13.686	14.197
3	15.128	15.457	9	15.310	14.474
4	14.675	15.211	10	13.387	12.348
5	16.175	15.735	11	14.281	14.701
6	13.707	13.967	12	14.150	13.805

0.797, respectively. Finally, the mean for group 1 was 14.928 and the mean for group 2 was 13.779. The resulting value from Equation (3.3-5) is $t'(22) = -2.1234$, $p < .05$. Interestingly, for this example, the statistical decision regarding equality of means is the same whether p or a sample estimate was used. Furthermore, as Zumbo and Zimmerman note, applying Coren and Hakstian's clustering methodology results in a failure to reject the null hypothesis of equal means. It is important to note that the data was created with an effect size of 1.5 standard errors of separation in the population, a value that is reasonably found in research settings -- see Zumbo & Zimmerman (1991) for more details regarding the simulation methodology.

Example 2: Effect of sleep deprivation on cognitive performance.

This example comes from the Sleep Lab at Carleton University and was kindly supplied by Dr. Bob Hoffmann. The data comes from a large study examining the effect of sleep deprivation on the cognitive performance of human subjects. For the purpose of this section, I will focus on the self-paced letter recognition task (see Wimmer, 1990, for more details). The task required that the subject identify letters presented on a computer screen. The dependent variable was letter recognition latencies on two occasions of testing during the night. The two groups consisted of an experimental group ($n=12$) of sleep deprived subjects and a control group ($n=13$).

For the purpose of this example I will assume that the researcher was not interested in the difference between the two sessions. In fact, I will further assume that the cost of conducting the study was so great that the researcher wanted to effectively double the sample size without doubling the cost necessary to recruit and house more subjects. Therefore, each subject was tested twice during the night but the results for each session were treated as though they were for different subjects. This meant that there were 24 data points in the experimental group and 26 data points in the control group. The group means were 35.70 ($SS=3027.23$) for the experimental group and 29.60 ($SS=1960.23$) for the control group. The Pearson product moment correlation between the two testing occasions of the experimental group was 0.87 and for the control group was 0.68. It should be noted that the variances may not be equal for the two groups, however, recent research (Zimmerman & Zumbo, 1993) has shown that the t -test is robust to unequal variances given that the sample sizes are roughly equal. The resulting approximate test statistic from applying Equation (3.3-5) was $t'(48)=1.53$, $p > .05$. This indicates that the sleep deprivation did not have an effect on the self-paced letter recognition task. This example

shows that although the approximate tests were developed for lateral measures they can be used in any context wherein the research question has an implicit coupling of the data¹⁸.

¹⁸ Note that this problem would often be treated with a split-plot (mixed design) ANOVA. However, for this example, the researcher was not interested in the difference across the two sessions.

CHAPTER 7: CONCLUSIONS

The purpose of this dissertation was to investigate both novel and known solutions to overcoming nonindependence of coupled data in significance testing. The purpose of this chapter, then, is to make some concluding remarks and recommendations regarding the various statistical methods.

Nonindependence has been studied from a variety of statistical perspectives, such as sampling theory, hierarchical modeling, classical test theory, repeated measures designs, intra-class correlation, time series analysis, multivariate statistics, and now coupled data. The case of coupled data involves pairwise related random variables rather than the commonly assumed case of independent random variables. Experimentally, coupled data results from a researcher's lack of interest in the difference between the elements of the couple. Coren and Hakstian (1990) cite several papers, some of them quite recent, in which the researchers are not interested in the difference between the two ears despite the fact that data are recorded from both ears. It is perplexing to me why a researcher would record measures from both ears when he/she is not interested in the difference between the ears. Clearly, this violates the assumption of within-group independence.

One wonders, then, what motivates a researcher to gather coupled data. One could conjecture that subjects and the experimental procedures are so costly and time consuming that it is necessary to effectively double your sample size by obtaining two observations from each subject. It is interesting that doubling of the sample size can result in a substantial increase in statistical power over the i.i.d. case. An issue for further

examination then, would be the relative efficiency of properly analyzing coupled data versus using only one element of the couple in the data analysis.

It can be seen from the first three chapters that there are many ways to deal with coupled data. It is difficult, if not impossible, to derive a widely agreed upon standard by which to choose the "best" method to use with coupled data. The idea of a best method is difficult, primarily, because many of the alternative methods are special cases of each other. For example, in theory, standard i.i.d. statistical methods are approximations to permutation tests (Box, 1976). Therefore, the Coren and Hakstian methodology is an approximation to the couplet referenced randomization test. Given that these tests are special cases of each other, they are not really alternatives at all (Box, 1979).

Despite the theoretical equivalence of some tests, a commonly known statistical criterion to choose the best test can be established by stating that a best test will be one which maintains its level (i.e., nominal type I error rate), yet is statistically more powerful than its competitor. Although very useful in the asymptotic theory of mathematical statistics, this definition has been difficult to employ in applied settings because one needs to know ahead of time a meaningful difference between two or more power values. This methodology of comparative power has, however, been used with some success in the field of nonparametric statistics (e.g., Zimmerman & Zumbo, 1993). For coupled data then, I would buttress the findings for the comparative power with a consideration of the cumulative nature of research. That is, in choosing a best method to analyze coupled data one needs to also consider that the advocated method should allow researchers to correct for coupled data in past studies (e.g., in a meta-analysis). It would also be beneficial if the methodology could be easily implemented with existing statistical software.

Given all of the above criteria, I would recommend that researchers use the formulas derived in Chapter 3. These statistical tests maintain their level, are as powerful,

and in some instances are more powerful than their competitors. They also are easily used in a meta-analytic context. Finally, when one has equal sample sizes and equal correlations, equations (3.2-2), (3.3-3), or (3.4-2) can be used to correct the obtained statistics using any commonly-found statistical package, thus making them immediately accessible to most researchers.

REFERENCES

- Aitken, A. C. (1934). On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh, A*, 55, 42-47.
- Blair, R. C., Higgins, J. J., Topping, M. E. H., & Mortimer, A. L. (1983). An investigation of the robustness of the t-test to unit of analysis violations. *Educational and Psychological Measurement*, 43, 69-80.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York, N.Y.: McGraw-Hill.
- Bock, R. D., Wainer, H., Petersen, A., Thissen, D., Murray, J., & Roche, A. (1973). A parameterization for individual growth curves. *Human Biology*, 45, 63-80.
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791-799.
- Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, 74, 1-4.
- Box, G. E. P., & Muller, M. (1958). A note on the generation of normal deviates. *Annals of Mathematical Statistics*, 29, 610-611.
- Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Proceedings of the Cambridge Philosophical Society*, 30, 178-191.
- Cochran, W. G. (1947). Some consequences when the assumptions in the analysis of variance are not satisfied. *Biometrics*, 3, 22-38.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd Edition)*. New Jersey: Lawrence Erlbaum and Associates.
- Conover, W.J. (1980). *Practical nonparametric statistics (2nd ed.)*. New York: Wiley.
- Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, 35, 124-129.
- Coren, S., & Hakstian, A. R. (1990). Methodological implications of interaural correlation: Count heads not ears. *Perception & Psychophysics*, 48, 291-294.
- Darlington, R. B., & Carlson, P. M. (1987). *Behavioral Statistics: Logic and Methods*. New York, N.Y.: The Free Press.
- Divenyi, P. L., & Haupt, K. M. (1992). In defense of the right and left audiograms: A reply to Coren (1989) and Coren and Hakstian (1990). *Perception & Psychophysics*, 52, 107-110.
- Ederer, F. (1973). Shall we count number of eyes or number of subjects? *Archives of Ophthalmology*, 89, 1-2.
- Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, 66, 485-487.
- Edgington, E. S. (1980). *Randomization tests*. New York: Dekker.
- Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology*, 4, 437-442.
- Fisher, R. A. (1926). Applications of "Student's distribution". *Metron*, 5, 90 - 104.
- Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology (2nd Edition)*. Englewood Cliffs: Prentice-Hall.
- Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs: Prentice-Hall.
- Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

- Harwell, M. R. (1991). Using randomization tests when errors are unequally correlated. *Computational Statistics & Data Analysis*, 11, 75-85.
- Irwin, J. O. (1978). Gossett, William Sealy. In W. H. Kruskal and J. M. Tanur (Eds.), *International Encyclopedia of Statistics*, Vol. 1, pp. 409 - 413. London: Collier Macmillan.
- Keen, K. J. (1987). *Estimation of intraclass and interclass correlations*. Unpublished doctoral dissertation, University of Toronto.
- Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50, 946-967.
- Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, 99, 422-432.
- Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, B36, 1-37.
- Lana, R. E., & Lubin, A. (1963). The effect of correlation on the repeated measures design. *Educational and Psychological Measurement*, 23, 729-739.
- Lehman, R. S. (1977). *Computer simulation and modeling: an introduction*. Hillsdale, N.J.: Erlbaum.
- Lewis, P. A. W., Goodman, A. S., & Miller, J. M. (1969). A pseudorandom number generator for the System 360. *IBM Systems Journal*, 8, 136-146.
- Lewis, P. A. W., & Orav, E. J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers* (Vol. 1). Pacific Grove, CA: Wadsworth.
- Lind, J. C., & Zumbo, B. D. (in press). The continuity principle in Psychological research: An introduction to robust statistics. *Canadian Psychology*.
- Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.
- Maguire, T. O., & Rogers, W. T. (1989). Proposed solutions for nonrandomness in educational research. *Canadian Journal of Education*, 14, 170-181.

- Marascuilio, L. A., & Serlin, R. C. (1988). *Statistical methods for the behavioral and social sciences*. New York, N.Y.: W. H. Freeman and Company.
- Morgan, B. J. T. (1984). *Elements of simulation*. London: Chapman & Hall.
- Nathan, G. (1981). Notes on inference based on data from complex sample designs. *Survey Methodology*, 7, 109-129.
- Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied Linear Statistical Models (2nd ed.)*. Homewood, Illinois: Richard D. Irwin, Inc.
- Rao, J. N. K., & Thomas, D. R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.
- Rawlings, J. O. (1988). *Applied Regression Analysis: A research tool*. Pacific Grove, Calif.: Wadsworth & Brooks/Cole Statistics/Probability Series.
- Ripley, B. D. (1987). *Stochastic Simulation*. New York: John Wiley & Sons.
- Rosner, B. (1982). Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics*, 38, 105-114.
- Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, Illinois: The Dorsey Press.
- Scariano, S. M., Davenport, J. M. (1987). The effects of violations of the independence assumption in the One-Way ANOVA. *American Statistician*, 41, 123-129.
- Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
- Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.
- Searle, S. R. (1971). *Linear models*. New York, N.Y.: John Wiley and Sons.
- Searle, S. R. (1982). *Matrix algebra useful for statistics*. N.Y.: John Wiley and Sons.
- Student (W. S. Gossett). (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
- Student (W. S. Gossett). (1927). Errors of routine analysis. *Biometrika*, 19, 151 - 164.

- Thomas, D. R. (in press). Inference using complex data from surveys and experiments. *Canadian Psychology*.
- Thomas, D. R., & Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.
- Timm, N. H. (1975). *Multivariate analysis: With applications in education and psychology*. Belmont, California: Wadsworth Publishing Company.
- Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics*, 18, 88-96.
- Wiley, T. L., Oviatt, D. L., & Block, M. G. (1987). Acoustic-immittance measures in normal ears. *British Journal of Speech and Hearing Research*, 30, 161-170.
- Wimmer, F. (1990). *The effects of short-term, total sleep deprivation on cognitive performance in human subjects*. Unpublished M.A. thesis, Carleton University.
- Wu, C. F. J., Holt, D., & Holmes, D. J. (1988). The effect of two-stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-159.
- Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1992). Correction of the Student *t* statistic for nonindependence of sample observations. *Perceptual and Motor Skills*, 75, 1011-1020.
- Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In Gideon Keren & Charlie Lewis (Eds.), *A handbook for data analysis in the behavioral sciences, Volume 1: Methodological issues* (pp. 481-517). Hillsdale, NJ.: Lawrence Erlbaum.
- Zimmerman, D. W., & Zumbo, B. D. (in press a). Correction for Nonindependence of Sample Observations in ANOVA F-Tests. *Journal of Experimental Education*.
- Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (in press). Coefficient alpha as an estimator of test reliability under violation of two assumptions. *Educational and Psychological Measurement*.

- Zumbo, B.D. (in press). The lurking assumptions in using generalizability theory to monitor an individual's progress. In Dany Laveault, Bruno D. Zumbo, Marc E. Gessaroli, & Marvin Boss (Eds.), *Modern Theories of Measurement: Problems & Issues*. Ottawa, Ontario: Edumetrics Research Group, University of Ottawa .
- Zumbo, B. D., & Zimmerman, D. W. (in press). Postscript: Some closing comments on "Alternatives to classical statistics" (Normal Theory). *Canadian Psychology*.
- Zumbo, B. D., & Zimmerman, D. W. (1991). Further evidence for Coren and Hakstian's "Methodological implications of interaural correlation: Count heads not ears" and an alternative correction formula. *Perception and Psychophysics*, 50, 297-301.

END

1 2 - 1 2 - 9 3

FIN