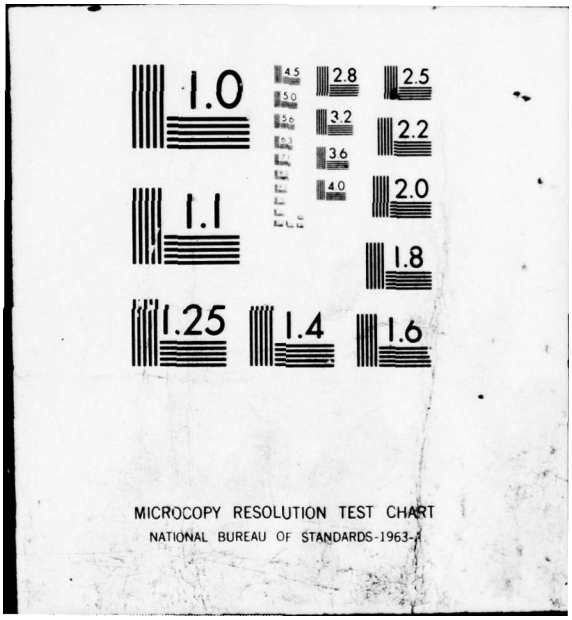
AD-A064 017 AIR FORCE ARMAMENT LAB EGLIN AFB FLA F/G 12/1
A STATISTICAL TOOL: ANALYSIS OF COVARIANCE. VOLUME I. A PRESENT--ETC(U)
APR 77 J C RICHARDSON

UNCLASSIFIED AFATL-TR-77-54-VOL-1 NL

1 OF 1
AD
A064 017

END
DATE
FILMED
4 -79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(14) AFATL-TR-77-54-VOL-1

(2)

A STATISTICAL TOOL: ANALYSIS OF COVARIANCE

VOLUME I. A PRESENTATION AND APPLICATION OF THE ANALYSIS OF COVARIANCE

LEVEL II

VULNERABILITY ASSESSMENTS BRANCH
ANALYSIS DIVISION

(12) 87p.

(11) APR 77

(9) FINAL REPORT FOR PERIOD JANUARY 1976-DECEMBER 1976

Approved for public release; distribution unlimited

(10) James C. Richardson

(16) 2549 (17) 04

AIR FORCE ARMAMENT LABORATORY
AIR FORCE SYSTEMS COMMAND • UNITED STATES AIR FORCE
EGLIN AIR FORCE BASE, FLORIDA

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFATL-TR-77-54, Volume I	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A STATISTICAL TOOL: ANALYSIS OF COVARIANCE - VOLUME I. A PRESENTATION AND APPLICATION OF THE ANALYSIS OF COVARIANCE		5. TYPE OF REPORT & PERIOD COVERED Final Report January - December 1976 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) James C. Richardson		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Vulnerability Assessments Branch (DLYV) Analysis Division Air Force Armament Laboratory		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element 62602F JON: 2549-04-07
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Armament Laboratory Armament Development and Test Center Eglin Air Force Base, Florida 32542		12. REPORT DATE April 1977 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Available in DDC

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Analysis of Covariance
Covariance Analysis
Missing Data Routings

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Volume I serves as an introduction to Volumes II and III for those who are not familiar with analysis of covariance. Volume I gives the purpose and uses of analysis of covariance, develops the theory for the univariate cases, expands the theory to the multivariate case, shows how unequal sample size affects the methodology, and how analysis of covariance is used as a tool for evaluating data containing missing observations on the response variable. Section V

→ next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ITEM 20 (CONCLUDED)

shows how analysis of covariance may be applied to nonparametric data. Examples depicting each situation are given.

Volume II incorporates all the situations presented in Volume I, except for Section V, and adds the condition of missing observations on the covariate and/or response variables. Volume II presents the theoretical development of the analysis of multivariate covariance in which missing values occur among both dependent and independent variables and presents an example.

Volume III contains the flow chart and program listing of the algorithm developed in Volume II. The theory developed in Volume II will serve for any categorized design, such as a randomized block, Latin square, etc., but the program is only suited for the randomized block model with additive block and treatment effects, or a general two factor additive effects model with no interaction, or the one way classification model.

PREFACE

This report consists of three volumes which present the theory and application of a valuable data reduction tool, the analysis of covariance. Volume I introduces the analysis of covariance as a general linear model (GLM) and then expands the model to incorporate the multivariate case, unequal sample size, and missing observations on the response variable. Volume I also covers the analysis of covariance for nonparametric data. This is Volume I.

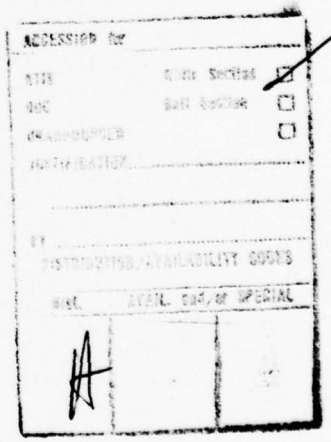
Volumes II and III were written by the Department of Statistics, Oklahoma State University, Stillwater, Oklahoma 74074, under Air Force Contract F08635-76-C-0154 with the Air Force Armament Laboratory, Armament Development and Test Center, Eglin Air Force Base, Florida 32542. The contract dealt with the development and programming of the methodology for evaluating multiple variable data with missing observations on dependent and independent variables by the analysis of covariance method. The methodology also covers case for unequal sample size. This work was begun in January 1976 and completed in December 1976.

This report has been reviewed by the Information Officer (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

J. R. MURRAY
Chief, Analysis Division



(The reverse of this page is blank)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SECTION I

MODEL, PURPOSE, AND USES

INTRODUCTION

This section introduces covariance analysis by explaining the model composition, by giving the purpose of the technique, by telling when it is applicable, and how it may be used. No details are presented, but general statements of results given in other parts of the report are presented.

In Section II the theory for covariance analysis in the univariate case with a single covariate is developed. Uses, such as adjusting treatment means, increasing the precision in randomized experiments, and obtaining insight into the nature of treatment effects, are explained. An example using the analysis of covariance in a completely randomized design with balanced data is given.

The theory for applying covariance analysis to a non-parametric situation is presented in Section III. Only one rank method is presented, but others are indicated. The data used in the example are real.

MODEL COMPOSITION

The covariance model consists of classification type variables, as found in an analysis of variance model, and a continuous type variable, as is usually found in regression models. Letting y_{ij} denote the j^{th} numbered observation in the i^{th} class, then in a covariance model, the response y_{ij} would be the result of a combination of features from the above conditions. For example, in a one-way classification with one covariate

$$y_{ij} = \mu_i + \beta (z_{ij} - \bar{z}_{..}) + \varepsilon_{ij} \qquad (1)$$

where μ_i represents the population mean of the i^{th} class when z_{ij} equals $\bar{z}_{..}$

β is a regression coefficient of y on z

z_{ij} is the covariate associated with the ij^{th} observation

$\bar{z}_{..}$ is the overall mean of the covariates and

ε_{ij} is the residual.

PURPOSE

Covariance analysis is primarily used in situations where one is interested in a response (dependent variable) which is influenced by one or more covariates which cannot be or have not been controlled by a randomization scheme. There may also be cases where the covariates have been controlled. The covariates usually reflect some characteristic which is related to, or influences, the response. This influence may affect the response directly or indirectly but does not necessarily have to produce a cause and effect situation. For example, in agronomy one may use the yield of grain per acre as a response and the number of plants per acre as the covariate. The covariate is also known as the independent variable or the concomitant variable.

PRINCIPAL USES

Covariance analysis has a variety of uses and its application will depend upon the investigator's objective.

(1) To adjust treatment means

Suppose the response contains contributions from the treatment effects, the covariate, and the error. To correct or adjust for the covariate, a quantity equal to the product of the estimated slope times the deviation of the mean of the covariate for a given treatment from the overall average of the covariate is subtracted from the average response of the treatment; i.e.,

$$\text{adj } \bar{y}_{i.} = \hat{\xi}_{i.} = \bar{y}_{i.} - \hat{\beta}\,(\bar{z}_{i.} - \bar{z}_{..}).$$

(2) To increase precision in randomized experiments

Covariance analysis converts the variance of the responses σ_y^2, to the variance about regression, $\sigma_{y.z}^2$. If $\sigma_{y.z}^2 \leq \sigma_y^2$, then covariance analysis is considered to have increased the precision and is an improvement over the analysis when covariance is not used. As long as the covariance model is linear, the covariance technique will result in the variance of a treatment mean, $V(\bar{y}_{i.})$, being changed from $\frac{\sigma_y^2}{n}$ to

$$\sigma_{y.z}^2 \left[\frac{1}{n} + \frac{(\bar{z}_{i.} - \bar{z}_{..})^2}{\sum_{ij} (z_{ij} - \bar{z}_{..})^2} \right]$$

for the univariate case.

(3) To remove the bias in observational studies

A researcher, conducting a survey, may be faced with taking a limited number of observations in a few locations. Also, these observations may not be randomized. Snedecor and Cochran (12) point out that these conditions would constitute an observational study. Suppose a researcher wished to study the relationship of obesity in workers by occupation and their physical activity. Since obesity may not be found in every worker, the researcher would have to take his observations wherever he can find a subject. Because of this, the researcher cannot predetermine a sampling scheme. Also, the response obesity would probably be measured as weight, a ratio scale measure, but the covariable, physical activity, would be measured on an ordinal scale. This may lead to problems of adjusting the means and in making inferences. Therefore, if another characteristic, such as age, is chosen as a substitute for physical activity, then a more sensitive comparison of obesity in workers may be made since age is measured on an interval scale.

(4) To provide additional information on the nature of treatment effects

Bancroft (1) points out that if treatment differences disappear after adjusting for the concomitant variable, then this may suggest that the unadjusted treatment differences are simply a reflection of the treatment effects on the concomitant variable. For this reason, treatments should not affect the concomitant variable.

(5) To analyze data when some observations are missing

Covariance analysis may be used as an alternative technique for analyzing data when some responses in an analysis of variance design are missing. The computations of the covariance technique are more involved than other missing data methods, but as Cochran (3) and Steel and Torrie (8) indicate, the technique yields unbiased sum of squares for estimating all classification effects. The technique also provides for exact F-tests to be made on the classification effects.

SECTION II

COVARIANCE ANALYSIS MODEL

INTRODUCTION

In this section, the theory will be developed for handling the covariance analysis model. The model coefficients, slopes, and means will be investigated, and a test statistic will be developed for testing hypotheses about these parameters. The assumptions underlying the model will be presented. Three of the principal uses (adjusting means, increasing precision, and obtaining information on treatment effects) will be discussed.

A method for determining if the analysis of covariance procedure offers advantages over the analysis when covariance is not used will be discussed.

MODEL

The model, as introduced in Section I, consists of t classes or treatments and n_i observations within each treatment. Then $i = 1, 2, \cdots, t$ and $j = 1, 2, \cdots, n_i$, where we assume $n_i \geq 2$, and for at least one treatment, $n_i \geq 3$. The way the model is subscripted indicates that each treatment may be estimated by a regression line of y on z. Therefore Equation (1) may be expressed as

$$y_{ij} = \mu_i + \beta_i (z_{ij} - \bar{z}_{..}) + \varepsilon_{ij} \qquad (1)$$

until it can be shown that one slope is common to all t regression lines. We will assume the error term ε_{ij} to have the following properties:

$$E(\varepsilon_{ij}) = 0 \text{ for all } i, j,$$

and

$$E(\varepsilon_{ij}\, \varepsilon_{i'j'}) = \sigma^2 \text{ when } i = i' \text{ and } j = j'$$

$$= 0 \text{ otherwise.}$$

By letting $\tau_i = \mu_i - \beta_i \bar{z}_{..}$, one will obtain an easier model with which to work:

$$y_{ij} = \tau_i + \beta_i z_{ij} + \varepsilon_{ij} . \qquad (2)$$

Assumptions for Analysis of Covariance

Cochran (3) lists two assumptions necessary to make covariance analysis valid:

(1) The design effect (blocks, treatments, etc.) and regression effect are additive. If for some reason they are not, one may still improve the precision, but

(a) The meaning of the adjusted treatment means may become questionable, and

(b) The true difference of treatment means will not be obtained.

(2) The residuals ε_{ij} are independent and normally distributed with zero means and equal variance. The normality assumption permits probability statements to be made about the statistics.

(3) Steel and Torrie (8) include one additional assumption. The covariate variables are measured without error.

Test for a Common Slope

Upon the completion of an experiment having a completely randomized design, one may display the test data as shown in Table 1.

TABLE 1. A RAW DATA SHEET FOR A COMPLETELY RANDOMIZED DESIGN EXPERIMENT

Treatments							
1		2		. . .	t		
y_{11}	z_{11}	y_{21}	z_{21}		y_{t1}	z_{t1}
y_{12}	z_{12}	y_{22}	z_{22}		y_{t2}	z_{t2}
⋮	⋮	⋮	⋮		⋮	⋮
y_{1n_1}	z_{1n_1}	y_{2n_2}	z_{2n_2}		y_{tn_t}	z_{tn_t}

It can be seen that by having $n_i \geq 2$, the data from the i^{th} treatment may be fitted to the model described by Equation (2).

We will now derive a test statistic for testing the following hypothesis:

H_0: all treatment slopes are equal ($\beta_1 = \beta_2 = \cdots = \beta_t = \beta$)

H_1: at least one slope is different from the rest.

Expressing Equation (2) in matrix notation, let

$$\underset{(n_i \times 1)}{\underset{\sim}{y}_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ y_{in_i} \end{bmatrix}, \qquad \underset{(n_\cdot \times 1)}{\underset{\sim}{y}} = \begin{bmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{y}_t \end{bmatrix},$$

$$\underset{(n_i \times 1)}{\underset{\sim}{z}_i} = \begin{bmatrix} z_{i1} \\ z_{i2} \\ \cdot \\ \cdot \\ \cdot \\ z_{in_i} \end{bmatrix}, \qquad \underset{(n_i \times 1)}{\underset{\sim}{\varepsilon}_i} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{in_i} \end{bmatrix},$$

$$\underset{(n_\cdot \times 2t)}{\Gamma} = \begin{bmatrix} \underset{\sim}{J}^{n_1}_1 & \underset{\sim}{0} & \cdots & \underset{\sim}{0} & \underset{\sim}{z}_1 & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{J}^{n_2}_1 & & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{z}_2 & & \underset{\sim}{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \underset{\sim}{0} & \underset{\sim}{0} & & \underset{\sim}{J}^{n_t}_1 & \underset{\sim}{0} & \underset{\sim}{0} & & \underset{\sim}{z}_t \end{bmatrix},$$

$$\underset{(t \times 1)}{\underset{\sim}{\tau}} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \cdot \\ \cdot \\ \cdot \\ \tau_t \end{bmatrix}, \qquad \underset{(t \times 1)}{\underset{\sim}{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_t \end{bmatrix},$$

$$\underset{\sim}{\eta}_{(2t \times 1)} = \begin{bmatrix} \underset{\sim}{\tau} \\ \\ \underset{\sim}{\beta} \end{bmatrix} \quad , \qquad \underset{\sim}{\varepsilon}_{(n_. \times 1)} = \begin{bmatrix} \underset{\sim 1}{\varepsilon} \\ \underset{\sim 2}{\varepsilon} \\ \\ \underset{\sim t}{\varepsilon} \end{bmatrix} .$$

$J^{n_i}_{\sim 1}$ is an $n_i \times 1$ vector of ones and $\underset{\sim}{0}$ is a vector of zeros, n_i in length.
We now have:

$$\underset{\sim}{y} = \Gamma \underset{\sim}{\eta} + \underset{\sim}{\varepsilon}$$

where $$E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$$

and $$E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}') = \sigma^2 I.$$

The normal equations are:

$$\Gamma' \Gamma \underset{\sim}{\hat{\eta}} = \Gamma' \underset{\sim}{y}$$

where

$$\Gamma' \Gamma_{(2t \times 2t)} = \begin{bmatrix} n_1 & 0 & \cdots & 0 & n_1\bar{z}_{1.} & 0 & \cdots & 0 \\ 0 & n_2 & & 0 & 0 & n_2\bar{z}_{2.} & & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & 0 & & n_t & 0 & 0 & & n_t\bar{z}_{t.} \\ n_1\bar{z}_{1.} & 0 & \cdots & 0 & \sum_j z_{1j}^2 & 0 & \cdots & 0 \\ 0 & n_2\bar{z}_{2.} & & 0 & 0 & \sum_j z_{2j}^2 & & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & 0 & & n_t\bar{z}_{t.} & 0 & 0 & & \sum_j z_{tj}^2 \end{bmatrix}$$

and

$$\underset{(2t \times 1)}{T' y} = \begin{bmatrix} n_1\bar{y}_{1.} \\ n_2\bar{y}_{2.} \\ \vdots \\ n_t\bar{y}_{t.} \\ \sum_j z_{1j}y_{1j} \\ \sum_j z_{2j}y_{2j} \\ \vdots \\ \sum_j z_{tj}y_{tj} \end{bmatrix}$$

The normal equations for the i^{th} treatment can be expressed as

$$n_i \hat{\tau}_i + n_i \bar{z}_{i.} \hat{\beta}_i = n_i \bar{y}_{i.} \tag{3}$$

$$n_i \bar{z}_{i.} \hat{\tau}_i + \sum_j z_{ij}^2 \hat{\beta}_i = \sum_j z_{ij} y_{ij} \quad . \tag{4}$$

Multiplying Equation (3) by $\bar{z}_{i.}$ and subtracting Equation (4) from it yields

$$\left(\sum_j z_{ij}^2 - n_i \bar{z}_{i.}^2\right) \hat{\beta}_i = \sum_j z_{ij} y_{ij} - n_i \bar{y}_{i.} \bar{z}_{i.} \quad .$$

Notice that $\left(\sum_j z_{ij}^2 - n_i \bar{z}_{i.}^2\right)$ is the corrected sum of squares for the covariate in the i^{th} treatment and that $\sum_j z_{ij}y_{ij} - n_i\bar{y}_{i.} \bar{z}_{i.}$ is the corrected cross product sum of the response and covariate in the i^{th} treatment. So $\hat{\beta}_i$ can be expressed as

$$\hat{\beta}_i = \frac{\sum_j (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})}{\sum_j (z_{ij} - \bar{z}_{i.})^2} \tag{5}$$

and from Equation (3), $\hat{\tau}_i$ is found to be

$$\hat{\tau}_i = \bar{y}_{i.} - \hat{\beta}_i \bar{z}_{i.} \quad .$$

Now calculating the sum of squares associated with the model containing each treatment mean and slope, one has

$$R(\tau_1, \cdots, \tau_t, \beta_1, \cdots, \beta_t) = \hat{\underset{\sim}{\eta}}' \Gamma' \underset{\sim}{y}$$

$$= \sum_i \hat{\tau}_i n_i \bar{y}_{i.} + \sum_i \hat{\beta}_i \sum_j z_{ij} y_{ij}$$

$$= \sum_i n_i \bar{y}_{i.}^2 + \sum_i \hat{\beta}_i \left[\sum_j (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})\right]$$

$$= \sum_i n_i \bar{y}_{i.}^2 + \sum_i \left[\frac{\left[\sum_j (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})\right]^2}{\sum_j (z_{ij} - \bar{z}_{i.})^2}\right]$$

For descriptive purposes, $R(\tau_1, \cdots, \tau_t, \beta_1, \cdots \beta_t)$ will be referred to in this subsection, as the reduction due to the full model. Subtracting the sum of squares of the reduction due to the full model from the total sum of squares in the model, one obtains the residual sum of squares for the model, or Residual (full):

$$\text{Residual (full)} = \underset{\sim}{y}'\underset{\sim}{y} - \hat{\underset{\sim}{\eta}}'\Gamma'\underset{\sim}{y}$$

$$= \sum_{ij} (y_{ij} - \bar{y}_{i.})^2 - \sum_i \left[\frac{\left[\sum_j (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})\right]^2}{\sum_j (z_{ij} - \bar{z}_{i.})^2}\right]$$

Express the residual sums of squares and cross products as

$$E_{yy}^{(i)} = \sum_j (y_{ij} - \bar{y}_{i.})^2$$

$$E_{zy}^{(i)} = \sum_j (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})$$

$$E_{zz}^{(i)} = \sum_j (z_{ij} - \bar{z}_{i.})^2$$

in order to simplify writing. The Residual (full) may now be written as

$$\text{Residual (full)} = \sum_i \left[E_{yy}^{(i)} - \frac{(E_{zy}^{(i)})^2}{E_{zz}^{(i)}} \right]$$

and this sum of squares has associated with it $(n_{.} - 2t)$ degrees of freedom since the rank of $\Gamma'\Gamma$ is $2t$.

The model describing the data may be simplified if a slope common in all treatments may be assumed. Consider now a reduced model incorporating a common slope and each treatment mean:

$$y_{ij} = \tau_i + \beta z_{ij} + \varepsilon_{ij} \quad . \tag{6}$$

In matrix notation, let $\underset{\sim}{y}_i$, $\underset{\sim}{y}$, $\underset{\sim}{\varepsilon}_i$, $\underset{\sim}{\tau}$, $\underset{\sim}{x}_i$, and $\underset{\sim}{\varepsilon}$ be defined as before.

Γ and $\underset{\sim}{\eta}$ for this model become:

$$\Gamma_{(n_{.} \times t+1)} = \begin{bmatrix} \underset{\sim}{J}^{n_1} & \underset{\sim}{0} & \cdots & \underset{\sim}{0} & \underset{\sim}{z}_1 \\ \underset{\sim}{0} & \underset{\sim}{J}^{n_2} & & \underset{\sim}{0} & \underset{\sim}{z}_2 \\ \vdots & & \ddots & & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{J}^{n_t} & \underset{\sim}{z}_t \end{bmatrix} ,$$

$$\underset{\sim}{\eta}_{(t+1 \times 1)} = \begin{bmatrix} \underset{\sim}{\tau} \\ \beta \end{bmatrix} .$$

In the reduced equation, $\underset{\sim}{y} = \Gamma \underset{\sim}{\eta} + \underset{\sim}{\varepsilon}$, one still assumes that $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ and $E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}') = \sigma^2 I$. The normal equations are

$$\Gamma'\Gamma \underset{\sim}{\hat{\eta}} = \Gamma' \underset{\sim}{y}$$

where

$$\Gamma'\Gamma_{(t+1 \times t+1)} = \begin{bmatrix} n_1 & 0 & \cdots & 0 & n_1\bar{z}_{1\cdot} \\ 0 & n_2 & & 0 & n_2\bar{z}_{2\cdot} \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & n_t & n_t\bar{z}_{t\cdot} \\ n_1\bar{z}_{1\cdot} & n_2\bar{z}_{2\cdot} & \cdots & n_t\bar{z}_{t\cdot} & \sum_{ij} z_{ij}^2 \end{bmatrix},$$

and

$$\Gamma'\underset{\sim}{y}_{(t+1 \times 1)} = \begin{bmatrix} n_1\bar{y}_{1\cdot} \\ n_2\bar{y}_{2\cdot} \\ \vdots \\ n_t\bar{y}_{t\cdot} \\ \sum_{ij} z_{ij}y_{ij} \end{bmatrix}$$

In solving for $\hat{\beta}$, one may multiply the:

1^{st} row by $\bar{z}_{1\cdot}$ and subtract from the last row

2^{nd} row by $\bar{z}_{2\cdot}$ and subtract from the last row

$\vdots$

t^{th} row by $\bar{z}_{t\cdot}$ and subtract from the last row.

This leaves all but the last term in the last row with zeros. The equation associated with the last row then becomes

$$\left(\sum_{ij} z_{ij}^2 - \sum_i n_i \bar{z}_{i.}^2 \right) \hat{\beta} = \sum_{ij} z_{ij} y_{ij} - \sum_i n_i \bar{z}_{i.} \bar{y}_{i.}$$

or $$\sum_i \left[\sum_j (z_{ij} - \bar{z}_{i.})^2 \right] \hat{\beta} = \sum_i \left[\sum_j (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.}) \right] .$$

Using the same notation as before, we get

$$\sum_i E_{zz}^{(i)} \hat{\beta} = \sum_i E_{zy}^{(i)} .$$

By defining

$$E_{wv}^{(\cdot)} = \sum_i E_{wv}^{(i)} ,$$

then

$$\hat{\beta} = \frac{E_{zy}^{(\cdot)}}{E_{zz}^{(\cdot)}} .$$

The numerator is the pooled (summed) sum of cross products in each treatment, and the denominator is the pooled sum of squares of the z_{ij}'s in each treatment. Solving for $\hat{\tau}_i$, one obtains

$$\hat{\tau}_i = \bar{y}_{i.} - \hat{\beta} \bar{z}_{i.}$$

where $\hat{\beta}$ is an estimate of the common slope.

We now need to find the sum of squares accounted for by the reduced model. It will become a component in the test statistic for a common slope.

$$R(\tau_1, \cdots, \tau_t, \beta) = \hat{\underset{\sim}{\eta}}' \, \Gamma' \, \underset{\sim}{y}$$

$$= \sum_i n_i \bar{y}_{i.}^2 + \frac{(E_{zy}^{(\cdot)})^2}{E_{zz}^{(\cdot)}} .$$

The Residual (reduced) becomes:

$$\text{Residual (reduced)} = \underset{\sim}{y}'\underset{\sim}{y} - \hat{\underset{\sim}{\eta}}'\Gamma'\underset{\sim}{y}$$

$$= \sum_{ij} y_{ij}^2 - \sum_i n_i \, y_{i\cdot}^2 - \frac{(E_{zy}^{(\cdot)})^2}{E_{zz}^{(\cdot)}}$$

$$= E_{yy}^{(\cdot)} - \frac{(E_{zy}^{(\cdot)})^2}{E_{zz}^{(\cdot)}} \quad .$$

The Residual (reduced) has associated with it $n_\cdot - (t + 1)$ degrees of freedom (d.f.) since the rank of $\Gamma'\Gamma$ is $(t + 1)$.

One can derive the likelihood ratio test, but an equivalent test statistic is given by U_1,

$$U_1 = \frac{\text{Residual (Reduced) - Residual (Full)}}{\text{Residual (Full)}}$$

$$\times \frac{\text{d.f. Residual (full)}}{[\text{d.f. Residual (reduced) - d.f. Residual (full)}]}$$

and if we assume $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2 I_n)$, then U_1 has an F-distribution under the null hypothesis. Therefore, U_1, becomes

$$U_1 = \frac{\left[E_{yy}^{(\cdot)} - \frac{(E_{zy}^{(\cdot)})^2}{E_{zz}^{(\cdot)}}\right] - \sum_i \left[E_{yy}^{(i)} - \frac{(E_{zy}^{(i)})^2}{E_{zz}^{(i)}}\right]}{\sum_i \left[E_{yy}^{(i)} - \frac{(E_{zy}^{(i)})^2}{E_{zz}^{(i)}}\right]} \cdot \frac{n_\cdot - 2t}{t - 1}$$

and $U_1 \sim F(t - 1, n_\cdot - 2t)$ when H_0 is true. $U_1 \sim F(t - 1, n_\cdot - 2t, \lambda)$

when H_0 is not true, where $\lambda = \frac{1}{2\sigma^2_{y \cdot z}} \sum_i n_i(\mu_i - \beta_i \bar{z}_{i.})^2$.

Test for a Common Mean

After one has tested for a common slope, one may then wish to test for a common treatment mean, that is

H_0: $\tau_1 = \cdots = \tau_n = \tau$

H_1: At least one treatment mean is different from the other treatment means.

In testing for a common mean, one must consider the test in terms of what has already transpired; i.e., the results of the previous "Test for a Common Slope," must be considered. Therefore, two situations should be considered:

(a) Case 1, where H_0 was rejected in the test for a common slope. The model to be considered under the hypothesis for a common mean is:

$$y_{ij} = \tau + \beta_i z_{ij} + \varepsilon_{ij}$$

versus

$$y_{ij} = \tau_i + \beta_i z_{ij} + \varepsilon_{ij} \quad .$$

The test for a common mean (intercept) will depend upon the covariate location. This situation will not be pursued.

(b) Case 2, where H_0 was not rejected in the test for a common slope. The model to be considered under the hypothesis for a common mean is:

$$y_{ij} = \tau + \beta z_{ij} + \varepsilon_{ij} \tag{7}$$

versus

$$y_{ij} = \tau_i + \beta z_{ij} + \varepsilon_{ij} \quad . \tag{8}$$

It may be noted that Equation (8) is the same as the reduced model under the hypothesis of a common slope. It now becomes the full model for testing a common intercept and therefore the Residual (full) is

$$E_{yy}^{(\cdot)} - \frac{(E_{zy}^{(\cdot)})^2}{E_{zz}^{(\cdot)}} \quad .$$

We now need to develop the Residual (reduced) for Equation (7). First, defining the components of the matrix model describing the reduced model, $\underset{\sim}{y}_i$, $\underset{\sim}{y}$, $\underset{\sim}{z}_i$, $\underset{\sim}{\varepsilon}_i$, and $\underset{\sim}{\varepsilon}$ remain as before. The other components are defined as:

$$\Gamma_{(n. \times 2)} = \begin{bmatrix} J_1^{n_1} & \underset{\sim}{z}_1 \\ \vdots & \vdots \\ J_1^{n_t} & \underset{\sim}{z}_t \end{bmatrix} \quad , \text{ and } \underset{\sim}{\eta}_{(2 \times 1)} = \begin{bmatrix} \tau \\ \beta \end{bmatrix} \quad .$$

The matrix model is

$$\underset{\sim}{y} = \Gamma \underset{\sim}{\eta} + \underset{\sim}{\varepsilon}$$

where $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ and $E(\underset{\sim\sim}{\varepsilon\varepsilon'}) = \sigma^2 I$.

The normal equations are

$$\Gamma'\Gamma \hat{\underset{\sim}{\eta}} = \Gamma' \underset{\sim}{y}$$

where

$$\Gamma'\Gamma_{(2 \times 2)} = \begin{bmatrix} n_. & n_. \bar{z}_{..} \\ n_. \bar{z}_{..} & \sum_{ij} z_{ij}^2 \end{bmatrix} \quad , \text{ and } \Gamma' y_{(2 \times 1)} = \begin{bmatrix} n_. \bar{y}_{..} \\ \sum_{ij} z_{ij} y_{ij} \end{bmatrix} \quad .$$

The normal equations may be expressed as

$$n_. \hat{\tau} + n_. \bar{z}_{..} \hat{\beta} = n_. \bar{y}_{..}$$

and

$$n_. \bar{z}_{..} \hat{\tau} + \sum_{ij} z_{ij}^2 \hat{\beta} = \sum_{ij} z_{ij} y_{ij} \quad .$$

So $\hat{\beta}$ and $\hat{\tau}$ are estimated by

$$\hat{\beta} = \frac{\sum_{ij} (z_{ij} - \bar{z}_{..})(y_{ij} - \bar{y}_{..})}{\sum_{ij} (z_{ij} - \bar{z}_{..})^2}$$

and $\hat{\tau} = \bar{y}_{..} - \hat{\beta}\ \bar{z}_{..}$.

The sum of squares associated with $\hat{\tau}$ and $\hat{\beta}$ may now be found to be:

$$R(\tau, \beta) = \hat{\underset{\sim}{\eta}}' \Gamma' \underset{\sim}{y}$$

$$= \hat{\tau}\, n_{.}\, \bar{y}_{..} + \hat{\beta} \sum_{ij} z_{ij}\, y_{ij}$$

and the residual sum of squares for the reduced model becomes

$$\text{Residual (reduced)} = \underset{\sim}{y}'\underset{\sim}{y} - \hat{\underset{\sim}{\eta}}'\, \Gamma'\, \underset{\sim}{y}$$

$$= \sum_{ij} y_{ij}^2 - \hat{\tau}\, n_{.}\, \bar{y}_{..} - \hat{\beta} \sum_{ij} z_{ij}\, y_{ij}$$

$$= \sum_{ij} (y_{ij} - \bar{y}_{..})^2 - \hat{\beta} \left[\sum_{ij} (z_{ij} - \bar{z}_{..})(y_{ij} - \bar{y}_{..}) \right].$$

Let the following notation stand for the respective sum of squares and cross products:

$$T_{zy} = \sum_{ij} (\bar{z}_{i.} - \bar{z}_{..})(\bar{y}_{i.} - \bar{y}_{..})$$

$$T_{zz} = \sum_{ij} (\bar{z}_{i.} - \bar{z}_{..})^2$$

$$T_{yy} = \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2 \quad .$$

Now consider the following identity:

$$(y_{ij} - \bar{y}_{..}) \equiv (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}) \quad .$$

Squaring and summing over all observations, one obtains

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 \equiv \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ij} (y_{ij} - \bar{y}_{i.})^2 \quad .$$

This result is shown in Appendix A. Using the notation previously given, the identity may be expressed as

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 \equiv T_{yy} + E_{yy}^{(\cdot)} \quad .$$

Likewise, it can be shown that the following identities hold:

$$\sum_{ij} (z_{ij} - \bar{z}_{..})^2 \equiv \sum_{ij} (\bar{z}_{i.} - \bar{z}_{..})^2 + \sum_{ij} (z_{ij} - \bar{z}_{i.})^2$$

$$\equiv T_{zz} + E_{zz}^{(\cdot)}$$

and that

$$\sum_{ij} (z_{ij} - \bar{z}_{..})(y_{ij} - \bar{y}_{..}) \equiv \sum_{ij} (\bar{z}_{i.} - \bar{z}_{..})(\bar{y}_{i.} - \bar{y}_{..})$$

$$+ \sum_{i} \left[\sum_{j} (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.}) \right]$$

$$\equiv T_{zy} + E_{zy}^{(\cdot)}$$

Thus,

$$\text{Residual (reduced)} = (T_{yy} + E_{yy}^{(\cdot)}) - \frac{(T_{zy} + E_{zy}^{(\cdot)})^2}{(T_{zz} + E_{zz}^{(\cdot)})}$$

with (n. - 2) degrees of freedom since the rank of the $\Gamma'\Gamma$ matrix is 2. The test statistic for a common mean, say U_2, is given by

$$U_2 = \frac{n. - t + 1}{t - 1} \cdot \frac{\left[(T_{yy} + E_{yy}^{(\cdot)}) - \frac{(T_{zy} + E_{zy}^{(\cdot)})^2}{(T_{zz} + E_{zz}^{(\cdot)})} \right] - \left[E_{yy}^{(\cdot)} - \frac{(E_{zy}^{(\cdot)})^2}{E_{zz}^{(\cdot)}} \right]}{\left[E_{yy}^{(\cdot)} - \frac{(E_{zy}^{(\cdot)})^2}{E_{zz}^{(\cdot)}} \right]} .$$

When normality is assumed, $U_2 \sim F\ (t - 1, n. - t + 1)$ when H_0 is true. $U_2 \sim F\ (t - 1, n. - t + 1, \lambda)$ when H_0 is not true with

$$\lambda = \frac{1}{2\sigma^2_{y.z}} \sum_i n_i\ (\mu_i - \beta\ \bar{z}_{i.})^2 \quad .$$

The results of an analysis of covariance for a completely randomized design are shown in Table 2, where D_1SS and n. - 2t are the terms used in computing the numerator of U_1, and D_2SS and n. - t + 1 are the terms used in computing the numerator of U_2. $(T_{wv} + E_{wv})$ represents the sum of products for the treatment and error terms for the indicated subscripts. Draper and Smith (4) point out that $s^2_{y.z}$ is an estimate of the variance about the regression line in each treatment when a common slope is assumed. An estimate of the error mean square will be $s^2_{y.z}$.

Adjustment of Treatment Means

The formula for adjusting treatment means was presented in Section I as being $\hat{\zeta}_{i.} = \bar{y}_{i.} - \hat{\beta}\ (\bar{z}_{i.} - \bar{z}_{..})$. It is assumed that a common slope was obtained for all treatments. Steel and Torrie (8) state, "Adjusted treatment means are estimates of what the treatment means would be if all $\bar{z}_{i.}$'s were at $\bar{z}_{..}$." The idea is presented graphically in Figure 1.

Suppose the results of two treatments are plotted. Let one treatment response be represented by +'s with response and concomitant means given by $(\bar{y}_{1.}, \bar{z}_{1.})$, respectively, and the other treatment represented by o's having response and concomitant means $(\bar{y}_{2.}, \bar{z}_{2.})$, respectively. Let $\bar{z}_{..}$ be the overall concomitant variable mean, $\hat{\zeta}_{1.}$ be the adjusted mean for

TABLE 2. ANALYSIS OF COVARIANCE TABLE FOR A COMPLETELY RANDOMIZED DESIGN

Source of Variation	Sum of Products				Regression		
	d.f.	z·z	z·y	y·y	d.f.	S. S.	$\hat{\beta}$
Total	$n_{\cdot} - 1$	S_{zz}	S_{zy}	S_{yy}	1	$(S_{zy})^2/S_{zz}$	
Treatment	$t - 1$	T_{zz}	T_{zy}	T_{yy}	1	$(T_{zy})^2/T_{zz}$	
Error	$n_{\cdot} - t$	$E_{zz}^{(\cdot)}$	$E_{zy}^{(\cdot)}$	$E_{yy}^{(\cdot)}$	1	$(E_{zy}^{(\cdot)})^2/E_{zz}^{(\cdot)}$	$E_{zy}^{(\cdot)}/E_{zz}^{(\cdot)}$
GP 1	$n_1 - 1$	$E_{zz}^{(1)}$	$E_{zy}^{(1)}$	$E_{yy}^{(1)}$	1	$(E_{zy}^{(1)})^2/E_{zz}^{(1)}$	$E_{zy}^{(1)}/E_{zz}^{(1)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
GP t	$n_t - 1$	$E_{zz}^{(t)}$	$E_{zy}^{(t)}$	$E_{yy}^{(t)}$	1	$(E_{zy}^{(t)})^2/E_{zz}^{(t)}$	$E_{zy}^{(t)}/E_{zz}^{(t)}$
Subtotal $(E^{(i)})$		$\Sigma_i E_{zz}^{(i)}$	$\Sigma_i E_{zy}^{(i)}$	$\Sigma_i E_{yy}^{(i)}$			
$(T_{wv} + E_{wv}^{(\cdot)})$		$(T_{zz}+E_{zz}^{(\cdot)})$	$(T_{zy}+E_{zy}^{(\cdot)})$	$(T_{yy}+E_{yy}^{(\cdot)})$			

TABLE 2. ANALYSIS OF COVARIANCE TABLE FOR A COMPLETELY RANDOMIZED DESIGN (CONCLUDED)

Source of Variation	ADJUSTED SUM OF PRODUCTS			F
	d.f.	S.S.	M.S.	
Total	$n_{\cdot} - 2$			
Treatment	$t - 2$	$T_{yy} - (T_{zy})^2/T_{zz}$		
Error	$n_{\cdot}-t-1$	$E_{yy}^{(\cdot)} - (E_{zy}^{(\cdot)})^2/E_{zz}^{(\cdot)}$	$s_{y\cdot z}^2$	
GP 1	$n_1 - 2$	$E_{yy}^{(1)} - (E_{zy}^{(1)})^2/E_{zz}^{(1)}$		
⋮	⋮	⋮		
GP t	$n_t - 2$	$E_{yy}^{(t)} - (E_{zy}^{(t)})^2/E_{zz}^{(t)}$		
Subtotal $(E^{(i)})$	$n_{\cdot} - 2t$	$\Sigma E_{yy}^{(i)} - (E_{zz}^{(i)})^2/E_{zz}^{(i)}$	$MS\Sigma E^{(i)}$	
(Difference for Testing Slopes Equal)	$t - 1$	D_1SS		For U_1
$(T_{wv} + E_{wv}^{(\cdot)})$	$n_{\cdot} - 2$	$(T_{yy}+E_{yy}^{(\cdot)}) - \dfrac{(T_{zy}+E_{zy}^{(\cdot)})^2}{(T_{zz}+E_{zz}^{(\cdot)})}$		
(Difference for Testing Means Equal)	$t - 1$	D_2SS		For U_2

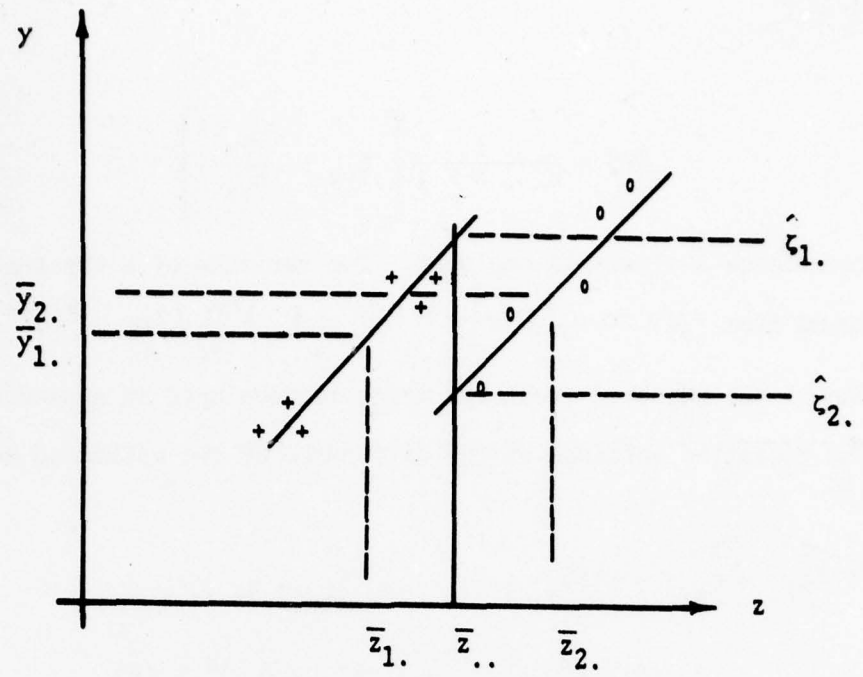


Figure 1. Adjustment of Treatment Means by Covariance Analysis

Treatment 1, and $\hat{\zeta}_{2.}$ be the adjusted mean for Treatment 2. Then $\hat{\zeta}_{i.}$ is the estimate of what the treatment would be if all $\bar{z}_{i.}$'s were at $\bar{z}_{..}$.

When considering an estimate of the difference between two adjusted treatment means, one would have

$$(\hat{\zeta}_{i.} - \hat{\zeta}_{k.}) = (\bar{y}_{i.} - \bar{y}_{k.}) - \hat{\beta}(\bar{z}_{i.} - \bar{z}_{k.}) .$$

Increase of Precision for Randomized Experiments

Dot notation will no longer be used for sum of products associated with the common slope model. It can be shown, by applying the results in Table 2 that the estimated variance of the responses without covariance is given by

$$EMS = \frac{1}{n. - t} E_{yy};$$

and becomes

$$EMS = \frac{1}{n_{\cdot} - t - 1}\left[E_{yy} - \frac{(E_{zy})^2}{E_{zz}}\right]$$

when covariance analysis is employed. The variance of a treatment mean is changed from σ_y^2/n to $\sigma_{y\cdot z}^2\left[\frac{1}{n} + (\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2/\sum_{ij}(z_{ij} - \bar{z}_{\cdot\cdot})^2\right]$. The variance of the adjusted treatment means is developed in Appendix B.

The estimated variance of the difference of two estimated adjusted means is

$$\widehat{V(\hat{\zeta}_{i\cdot} - \hat{\zeta}_{k\cdot})} = V[\bar{y}_{i\cdot} - \bar{y}_{k\cdot} - \hat{\beta}(\bar{z}_{i\cdot} - \bar{z}_{k\cdot})]$$

$$= V[\bar{y}_{i\cdot} - \bar{y}_{k\cdot}] + (\bar{z}_{i\cdot} - \bar{z}_{k\cdot})^2 V(\hat{\beta})$$

$$- 2\,\mathrm{cov}[(\bar{y}_{i\cdot} - \bar{y}_{k\cdot}), (\bar{z}_{i\cdot} - \bar{z}_{k\cdot})\hat{\beta}] \quad .$$

It has been shown in Appendix B that $\mathrm{cov}(\bar{y}_{i\cdot}, \hat{\beta}) = 0$. Likewise, $\mathrm{cov}(\bar{y}_{k\cdot}, \hat{\beta}) = 0$. $V(\hat{\beta})$ is also developed in Appendix B.

The above expression then reduces to

$$\widehat{V(\hat{\zeta}_{i\cdot} - \hat{\zeta}_{k\cdot})} = s_{y\cdot z}^2/n_i + s_{y\cdot z}^2/n_k + (\bar{z}_{i\cdot} - \bar{z}_{k\cdot})^2 s_{y\cdot z}^2/\Sigma(z_{ij} - \bar{z}_{i})^2$$

$$= s_{y\cdot z}^2\left[\frac{1}{n_i} + \frac{1}{n_k} + (\bar{z}_{i\cdot} - \bar{z}_{k\cdot})^2/E_{zz}\right]$$

where $s_{y\cdot z}^2$ estimates $\sigma_{y\cdot z}^2$. A disadvantage of the above form is that $\widehat{V(\hat{\zeta}_{i\cdot} - \hat{\zeta}_{k\cdot})}$ is different for every pair of treatments being compared. One may then like to have an average value for the variance.

Since the average would be over t treatments taken two at a time, the average value for the difference between two adjusted means is

$$\frac{1}{t\,(t-1)} \sum_{\substack{i\,k \\ i\neq k}} V\,\widehat{(\hat{\zeta}_{i.} - \hat{\zeta}_{k.})} = \frac{s_{y\cdot z}^2}{t\,(t-1)} \left[\sum_{\substack{i\,k \\ i\neq k}} \left(\frac{1}{n_i} + \frac{1}{n_k} + \frac{(\bar{z}_{i.} - \bar{z}_{k.})^2}{E_{zz}} \right) \right]$$

$$= \frac{s_{y\cdot z}^2}{t\,(t-1)} \left[(t-1) \sum_i \frac{1}{n_i} + (t-1) \sum_k \frac{1}{n_k} + \frac{1}{E_{zz}} \sum_{\substack{i\,k \\ i\neq k}} (\bar{z}_{i.} - \bar{z}_{k.})^2 \right].$$

One may now apply the identity expressing the sum of squares of deviations about the mean in terms of the sum of squares of all differences; that is,

$$\sum_{i=1}^{n} (z_i - \bar{z})^2 \equiv \frac{1}{2n} \sum_{\substack{i\,k \\ i\neq k}} (z_i - z_k)^2 \equiv \frac{1}{2n} \sum_{ik} (z_i - z_k)^2 \quad .$$

This identity is proved in Appendix C. We now have

$$\frac{1}{t\,(t-1)} \sum_{\substack{i\,k \\ i\neq k}} V\,\widehat{(\hat{\zeta}_{i.} - \hat{\zeta}_{k.})} = \frac{2s_{y\cdot z}^2}{t} \sum_{i=1}^{t} \frac{1}{n_i}$$

$$+ \frac{2s_{y\cdot z}^2}{(t-1)\,E_{zz}} \sum_{i=1}^{t} (\bar{z}_{i.} - \bar{z}_{..})^2 \qquad (9)$$

where $\bar{z}_{..}$ is the unweighted mean of the treatment means,

$$\bar{z}_{..} = \frac{1}{t} \sum_{i=1}^{t} \bar{z}_{i.} \quad .$$

The harmonic mean, n_H, is defined as

$$n_H = \frac{t}{\sum_{i=1}^{t} \frac{1}{n_i}} \quad .$$

Using this, Equation (9) reduces to

$$\frac{1}{t\,(t-1)} \sum_{\substack{i\ k \\ i \neq k}} \widehat{V\,(\hat{\zeta}_{i.} - \hat{\zeta}_{k.})}$$

$$= 2\, s^2_{y \cdot z} \left[\frac{1}{n_H} + \frac{1}{(t-1)\, E_{zz}} \sum_{i=1}^{t} (\bar{z}_{i.} - \bar{z}_{..})^2 \right] \quad .$$

If the design is balanced, then $n_i = r$ for each treatment. This implies that $n_H = r$. Recall that

$$T_{zz} = \sum_{ij} (\bar{z}_{i.} - \bar{z}_{..})^2 = r \sum_{i} (\bar{z}_{i.} - \bar{z}_{..})^2 \quad .$$

So employing these conditions we have

$$\frac{1}{t\,(t-1)} \sum_{\substack{i\ k \\ i \neq k}} \widehat{V\,(\hat{\zeta}_{i.} - \hat{\zeta}_{k.})} = \frac{2 s^2_{y \cdot z}}{r} \left[1 + \frac{T_{zz}}{(t-1)\, E_{zz}} \right]$$

as a final result for the case when $n_i = r$.

Efficiency

Snedecor and Cochran (12) state that a method for determining whether analysis of covariance is more beneficial than the analysis

without covariance is to calculate the efficiency. The efficiency is defined to be

$$\text{Efficiency} = \frac{s_y^2}{s_{y\cdot z}^2 \left[1 + \frac{T_{zz}}{(t-1)E_{zz}}\right]} .$$

The denominator is defined by Snedecor and Cochran as being "the effective error mean square per observation when computing the error variance for any comparison among the treatment means." The larger the value of the ratio, the more efficient is analysis of covariance.

EXAMPLE OF A COMPLETELY RANDOMIZED DESIGN

A tool manufacturer markets three kit sizes, each consisting of seven bits. The amount of alloy added for hardness varies in each bit because of bit design and of kit size: small, medium, and large. The manufacturer is interested in finding if the life expectancy of the kits as a whole are the same. Each bit was mounted and subjected to material of like density for equivalent lengths of time. It was decided that the quantity of alloy (z) added for hardness would influence the amount of wear (y). The test results are presented in Table 3.

In testing for a common slope, one would want to determine if the model

$$y_{ij} = \tau_i + \beta z_{ij} + e_{ij}$$

can predict the same results as

$$y_{ij} = \tau_i + \beta_i z_{ij} + e_{ij} .$$

Therefore the hypothesis will be

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = \beta$$

$$H_1: \quad \beta_1 \neq \beta_j \text{ for at least one i and j.}$$

Calculations for the residual sum of products will be shown for the small kit only.

TABLE 3. DATA TABLE FOR EXAMPLE 1

Kit Size					
Small		Medium		Large	
Alloy z Milligrams	Wear y Millimeters	Alloy z Milligrams	Wear y Millimeters	Alloy z Milligrams	Wear y Millimeters
15	33	28	24	40	16
16	31	31	22	43	14
19	31	34	23	48	13
22	30	38	19	50	11
24	29	40	20	53	11
25	27	43	17	55	9
32	26	46	18	58	9
TOTALS 153	207	260	143	347	83
GRAND TOTAL				760	433
CROSS PRODUCT TOTAL	4443		5217		4016
SUM OF SQUARES 3551	6157	9910	2963	17451	1025

$$E_{zz}^{(1)} = \sum_j (z_{ij} - \bar{z}_{i.})^2$$

$$= \sum_j z_{ij}^2 - (\Sigma z_{ij})^2/n_i$$

$$= 3551 - (153)^2/7$$

$$= 206.857143$$

$$E_{yy}^{(1)} = \sum_j (y_{ij} - \bar{y}_{i.})^2$$

$$= \sum_j y_{ij}^2 - (\Sigma y_{ij})^2 n_i$$

$$= 6157 - (207)^2/7$$

$$= 35.714286$$

$$E_{zy}^{(1)} = \sum_j (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})$$

$$= \sum_j z_{ij}y_{ij} - (\sum_j z_{ij})(\sum_j y_{ij})/n_i$$

$$= 4443 - (207)(153)/7$$

$$= -81.428571$$

Now solving for an estimate β_1, the slope for the small kit,

$$\hat{\beta}_1 = \frac{E_{zy}^{(1)}}{E_{zz}^{(1)}}$$

$$= -81.428571/206.857143$$

$$= -.39$$

and for the adjusted sum of squares

$$\text{ADJ S.S.} = E_{yy}^{(1)} - \frac{(E_{zy}^{(1)})^2}{E_{zz}^{(1)}}$$

$$= 35.714286 - (-81.428571)^2/206.857143$$

$$= 3.660222 .$$

The residual sum of products are presented in Table 4. The test statistic shows a value of 0.0421. Comparing this to a tabulated F value, we have

$$F(2, 15, \alpha = .10) = 2.70 .$$

One would not reject H_0 at this level. Therefore, accept the model that estimates the three sets of data by a common regression slope but possibly having a different intercept.

The adjusted and unadjusted means are presented in Table 5. Comparing the unadjusted means for the kits, one may conclude that the average amount of wear between the large and small kit is significant, but looking at the adjusted means for the same kit sizes, the difference in the amount of wear has been greatly reduced. The adjusted means are estimates of what the average kit wear would be if compared on the basis of each kit having the same amount of alloy.

The interest now is to determine if the average life expectancy of the three kits is the same. The hypothesis is

H_0: $\tau_1 = \tau_2 = \tau_3 = \tau$

H : $\tau_i \neq \tau_j$ for at least one i and j.

In calculating the sum of products, only the cross products will be shown.

TABLE 4. THE RESIDUAL ANALYSIS OF COVARIANCE TABLE FOR EXAMPLE 1.

Source	d.f.	Sum of Products Z·Z	Z·Y	Y·Y	$\hat{\beta}$	ADJ. d.f.	ADJ. Products	$s_{y\cdot z}^2$
Small	6	206.857143	-81.428571	35.714286	-.3936	5	3.660222	.7320
Medium	6	252.857143	-94.428571	41.714286	-.3734	5	6.450283	1.2901
Large	6	249.71429	-98.428571	40.8571429	-.3942	5	2.06007	.4120
			$\sum_i [E_{yy}^{(i)} - (E_{zy}^{(i)})^2/E_{zz}^{(i)}]$			15	12.170575	
$E^{(\cdot)}$	18	709.428576	-274.285713	118.2857149	-.3866	17	12.238884	.7199

$$U_1 = \frac{\text{Residual (reduced)} - \text{Residual (full)}}{\text{Residual (full)}} \cdot \frac{\text{d.f. Residual (full)}}{\text{d.f. Residual (reduced)} - \text{d.f. Residual (full)}}$$

$$= \frac{12.238884 - 12.170575}{12.170575} \cdot \frac{15}{2} = 0.0421$$

TABLE 5. MEAN - VARIANCE TABLE

Table of Unadjusted/Adjusted Means

	Small	Medium	Large
Unadjusted	29.6 mm	20.4 mm	11.9 mm
Adjusted	24.0 mm	20.8 mm	17.0 mm

Unadjusted Mean = $\bar{y}_{i.}$

$$\hat{\zeta}_{i.} = \bar{y}_{i.} - \hat{\beta}\,(\bar{z}_{i.} - \bar{z}_{..})$$

ESTIMATED VARIANCES FOR THE ADJUSTED TREATMENT MEAN WITH COMMON $\hat{\beta}$

Small	Medium	Large
0.3243 mm^2	0.1168 mm^2	0.2976 mm^2

$$\widehat{V(\hat{\zeta}_{i.})} = s^2_{y \cdot z}\left[\frac{1}{n} + (\bar{z}_{i.} - \bar{z}_{..})^2 / \Sigma\,(z_{ij} - \bar{z}_{..})^2\right]$$

$$S_{zy} = \sum_{ij} (z_{ij} - \bar{z}_{..})(y_{ij} - \bar{y}_{..})$$

$$= \sum_{ij} z_{ij} y_{ij} - \frac{1}{n_{..}} (\sum_{ij} z_{ij})(\sum_{ij} y_{ij})$$

$$= 13676 - \frac{1}{21}(760)\ (433)$$

$$= -1994.47619 .$$

$$T_{zy} = \sum (\bar{z}_{i.} - \bar{z}_{..})(\bar{y}_{i.} - \bar{y}_{..})$$

$$= \frac{1}{n_i} \sum (z_{i.})(y_{i.}) - \frac{1}{n_{..}} (\sum_{ij} z_{ij})(\sum_{ij} y_{ij})$$

$$= \frac{1}{7}\ (97652) - \frac{1}{21}\ (760)(433)$$

$$= -1720.19048 .$$

The results are tabulated in the analysis of covariance table, Table 6. The test statistic gives a value 25.848 and the tabulated F value for two and 17 degrees of freedom (d.f.) for $\alpha = .10$ is 2.64. The null hypothesis is rejected, and the manufacturer concludes that the expected life for at least one kit is different from the rest.

An increase in the precision of the responses can be seen by comparing the estimated variance without covariance, 6.75, to the estimate of the variance about regression, 0.7199 (see page 2). The estimated variance associated with each adjusted treatment mean is presented in Table 3.

If one wished to consider the difference between two adjusted means, such as $\hat{\zeta}_1$ and $\hat{\zeta}_3$, then the difference is estimated by:

$$(\hat{\zeta}_1 - \hat{\zeta}_3) = 38.10 - 31.19 = 6.91,$$

while an estimate of the average variance would be

TABLE 6. THE ANALYSIS OF COVARIANCE FOR EXAMPLE 1

Source	d.f.	SUM OF PRODUCTS z·z	z·y	y·y	β	ADJ d.f.	ADJUSTED PRODUCTS	MEAN SQUARE
Total	20	3407.2381	-1994.47619	1216.952381		19	49.456845	
Treatment	2	2697.80953	-1720.19048	1098.666671		1	1.830421	
Error	18	709.428576	-274.285713	118.2857149	-.39	17	12.238884	.72

Treatment + error is the same as total; therefore,

$$U_2 = \frac{49.456845 - 12.238884}{12.238884} \cdot \frac{17}{2} = 25.848$$

$$\frac{1}{t\,(t-1)} \sum_{\substack{i\ k \\ i \neq k}} V\,(\hat{\tau}_i - \hat{\tau}_k) = \frac{2\,s_{y\cdot z}^2}{r} \left[1 + \frac{T_{zz}}{(t-1)\,E_{zz}}\right]$$

$$= \frac{2\,(.7199)}{7} \left[1 + \frac{2697.80953}{(2)(709.428576)}\right]$$

$$= 0.5968 \quad .$$

One may now like to see how efficient covariance analysis was:

$$E = \frac{s_y^2}{s_{y\cdot z}^2} \left[1 + \frac{T_{zz}}{(t-1)\,E_{zz}}\right]$$

$$= \frac{6.57}{2.09}$$

$$= 3.15 \quad .$$

Analysis of covariance with 10 replicates per treatment will give estimates of treatment difference which are just as precise as 32 replicates per treatment without covariance analysis.

SECTION III

MULTIVARIATE COVARIANCE ANALYSIS MODEL

INTRODUCTION

In this section, the analysis of covariance will be extended from the univariate case to the multivariate case. The multivariate case to be considered is one where there is a single response and more than one concomitant variable. Complete development of the theory will not be presented; only enough will be presented to tie in with what was presented in Section II. An example displaying the multiple covariance technique in a randomized block design with unequal sample sizes will be presented, and a test statistic for the hypothesis of no interaction will be derived.

The case of many responses and many concomitant variables will not be considered. Morrison (8) gives a brief account of this case. Hazel (6) presents an analysis of covariance for multivariate data with unequal subclass sizes. The data is presented in a regression type of analysis of variance table with no indication of adjustments for the concomitant variables. The Statistical Analysis System (SAS) general linear model routine will present the data in the same format. The regression routine is used instead of the analysis of variance routine because of the computational procedures required to deal with unequal sample sizes.

THEORY

It was shown in Section II that the analysis of covariance model can be written in matrix form as

$$\underset{\sim}{y} = \Gamma \underset{\sim}{\eta} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{y}$ is a nx1 vector of responses, Γ is an nxp matrix containing the classification pattern and values of the covariates, and $\underset{\sim}{\eta}$ is a px1 vector of unknown constants. When dealing with multivariate data, it may be helpful to partition Γ and $\underset{\sim}{\eta}$, thereby separating design and covariate information:

$$\underset{\sim}{y} = [X : Z] \begin{bmatrix} \underset{\sim}{\tau} \\ \\ \underset{\sim}{\beta} \end{bmatrix} + \underset{\sim}{\varepsilon}$$

$$\underset{\sim}{y} = X\underset{\sim}{\tau} + Z\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}.$$

Searle's (11) approach is more compatible with the example to be presented and, therefore, will be followed.

Normal Equations. In solving the normal equations for the best estimates of $\underset{\sim}{\tau}$ and $\underset{\sim}{\beta}$, $\underset{\sim}{a}$ and $\underset{\sim}{b}$ will serve as trial estimators for τ and $\hat{\beta}$, respectively.

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix} \begin{bmatrix} \underset{\sim}{a} \\ \underset{\sim}{b} \end{bmatrix} = \begin{bmatrix} X'\underset{\sim}{y} \\ Z'\underset{\sim}{y} \end{bmatrix}$$

In the situation with more than one observation under each set of conditions, i.e., the design conditions, X'X will not be of full rank but more than likely Z'Z will be of full rank. Let $(X'X)^-$ be a generalized inverse of X'X. Using the first equation of the normal equations, a solution for $\underset{\sim}{a}$ in terms of $\underset{\sim}{b}$ may be obtained.

$$\underset{\sim}{a} = (X'X)^- \; [X'\underset{\sim}{y} - X'Z \; \underset{\sim}{b}]$$

$$= (X'X)^- \; X'\underset{\sim}{y} - (X'X) - X'Z \; \underset{\sim}{b} \; .$$

Now using the second equation of the normal equations in solving for $\underset{\sim}{b}$ after substituting for $\underset{\sim}{a}$:

$$Z'X \; [(X'X)^- \; X'\underset{\sim}{y} - (X'X)^- \; X'Z \; \underset{\sim}{b}] + Z'Z \; \underset{\sim}{b} = Z'\underset{\sim}{y}$$

$$\underset{\sim}{b} = \{Z'[I - X \; (X'X)^- \; X'] \; Z\}^- \; Z' \; [I - X \; (X'X)^- \; X'] \; \underset{\sim}{y} \; .$$

Let $H = I - X \; (X'X)^- \; X'$, then

$$\underset{\sim}{b} = (Z'HZ)^- \; Z'H\underset{\sim}{y}$$

Even though $(X'X)^-$ is not unique, it appears in $\underset{\sim}{b}$ only in the form $X' \; (X'X)^- \; X$ which is unique for any generalized inverse of X'X. Searle (11) states that H is both symmetric and idempotent. This ensures Z'HZ and HZ to have the same rank, and based on the properties of X and Z given in the partitioned equation, will guarantee that HZ has full column rank and hence Z'HZ is non-singular. Therefore, $\underset{\sim}{b}$ is a unique solution and is the b.l.u.e. of β. Following standard notation, let $\hat{\beta}$ represent $\underset{\sim}{b}$; $\underset{\sim}{\hat{\beta}} = (Z'HZ)^{-1} \; Z'H\underset{\sim}{y}$.

Solving for the Covariate Coefficients. Working with $\hat{\beta}$ in the following matrix form

$$\hat{\underset{\sim}{\beta}} = \{Z' [I - X'(X'X)^{-}X] Z\}^{-} Z' [I - X'(X'X)^{-} X'] \underset{\sim}{y}$$

may lead to difficult calculations, so a better computational method will now be sought. Considering the following part of the above relation,

$$I - X'(X' X)^{-}X',$$

it can be seen that the identity matrix can be identified with a total amount of variation and the term

$$X'(X'X)^{-}X'$$

can be identified with some other amount of variation. The two parts together form a difference or residual which is idempotent. Looking at the pre- and post-multipliers of the residual, one is able to see that $\hat{\beta}$ consists of the inverse sum of squares of the covariate values and the sum of products of the covariate and response. Let R be the matrix of residuals for the covariate terms, then

$$\hat{\underset{\sim}{\beta}} = (R' R)^{-1} R' \underset{\sim}{y} \quad .$$

If one, so to speak, takes a step backward and expresses the relationship as

$$R' R \hat{\underset{\sim}{\beta}} = R' \underset{\sim}{y},$$

then the $\hat{\beta}_i$ values may be easily found. The above matrix may now be expressed in equations as

$$E_{Z_1Z_1} \hat{\beta}_1 + E_{Z_1Z_2} \hat{\beta}_2 + \cdots + E_{Z_1Z_K} \hat{\beta}_K = E_{Z_1y}$$

$$E_{Z_2Z_1} \hat{\beta}_1 + E_{Z_2Z_2} \hat{\beta}_2 + \cdots + E_{Z_2Z_K} \hat{\beta}_K = E_{Z_2y}$$

$$\vdots$$

$$E_{Z_KZ_1} \hat{\beta}_1 + E_{Z_KZ_2} \hat{\beta}_2 + \cdots + E_{Z_KZ_K} \hat{\beta}_K = E_{Z_Ky} \ ,$$

where

$E_{Z_i Z_j}$ are error or residual entries in a covariate analysis table. The solution vector of the normal equations in a covariate model then becomes

$$\begin{bmatrix} \underset{\sim}{a} \\ \underset{\sim}{\hat{\beta}} \end{bmatrix} = \begin{bmatrix} (X'X)^{-} X'y - (X'X)^{-} X'Z \underset{\sim}{\hat{\beta}} \\ (Z'HZ)^{-} ZH\underset{\sim}{y} \end{bmatrix}$$

Analysis of Covariance Table. The format for the analysis of covariance table is basically the same as presented in Section II. Modifications will be necessary due to the hypothesis to be tested, having unequal sample sizes, and additional independent variables.

In covariance analysis, interest usually centers on making inferences about aspects of the classification part of the model. For the case under consideration, interest is on whether interaction is important in the model. Considering a randomized block design with two blocks, four treatments, and two covariates, a test will be made to see if the full model

$$y = \rho_i + \tau_j + (\rho\tau)_{ij} + \beta_i z_i + \beta_k z_k + \varepsilon_i,$$

where $E(\varepsilon) = 0$,

$V(\varepsilon) = \sigma^2$,

can be predicted by the model

$$y = \rho_i + \tau_j + \beta_i z_i + \beta_k z_k + \varepsilon_i.$$

The hypothesis to be tested is

H_0: $(\rho\tau)_{ij} = 0$ for all ij

H_1: $(\rho\tau)_{ij} \neq 0$ for at least one ij combination.

The analysis of covariance table is summarized in Table 7. Table 8 presents the equations necessary for obtaining Table 7. The equations are expressed in terms of the dummy variables w and v. B_{ww} represents

TABLE 7. ANALYSIS OF COVARIANCE TABLE FOR A RANDOMIZED BLOCK DESIGN WITH UNEQUAL SAMPLE SIZES

Source of Variation	d.f.	Sum of Products						$\hat{\beta}_1$	$\hat{\beta}_2$	Adjusted Sum of Squares $\Sigma y^2 - \beta_1 \Sigma z_1 y - \beta_2 \Sigma z_2 y$	Adjusted d.f.	Adjusted Mean Square
		$z_1 \cdot z_1$	$z_1 z_2$	$z_2 z_2$	$z_1 y$	$z_2 y$	$y y$					
Total	$n_{...}-1$	$S_{z_1z_1}$	$S_{z_1z_2}$	$S_{z_2z_2}$	S_{z_1y}	S_{z_2y}	S_{yy}					
Due to full model	$n_{i..}n_{.j.}-1$	$H_{z_1z_1}$	$H_{z_1z_2}$	$H_{z_2z_2}$	H_{z_1y}	H_{z_2y}	H_{yy}					
Due to reduced model	$n_{i..}+n_{.j.}-1$	$P_{z_1z_1}$	$P_{z_1z_2}$	$P_{z_2z_2}$	P_{z_1y}	P_{z_2y}	P_{yy}					
Rows adjusted	$n_{i..}-1$	$B_{z_1z_1}$	$B_{z_1z_2}$	$B_{z_2z_2}$	B_{z_1y}	B_{z_2y}	B_{yy}					
Columns adjusted	$n_{.j.}-1$	$T_{z_1z_1}$	$T_{z_1z_2}$	$T_{z_2z_2}$	T_{z_1y}	T_{z_2y}	T_{yy}					
Difference (interaction)	$n_{i..}n_{.j.}-n_{i..}-n_{.j.}$	$D_{z_1z_1}$	$D_{z_1z_2}$	$D_{z_2z_2}$	D_{z_1y}	D_{z_2y}	D_{yy}					
Residual	$n_{...}-n_{i..}n_{.j.}$	$E_{z_1z_1}$	$E_{z_1z_2}$	$E_{z_2z_2}$	E_{z_1y}	E_{z_2y}	E_{yy}			ESS	$n_{...}-n_{i..}n_{.j.}-2$	EMS
Difference + residual	$n_{...}-n_{i..}-n_{.j.}$									(D+R)SS	$n_{...}-n_{i..}-n_{.j.}-2$	(D+R)MS
Difference*											$n_{i..}n_{.j.}-n_{i..}-n_{.j.}$	DMS

$U = \frac{DMS}{EMS}$, $U \sim F\,(n_{i..}\,n_{.j.} - n_{i..} - n_{.j.}\,,\ n_{...} - n_{i..}\,n_{.j.})$

DIFFERENCE* is Residual (Reduced) - Residual (Full)

TABLE 8. EQUATIONS FOR TABLE 7

Let w_{ijk} represent any variable where $i = 1, \cdots, r$

$j = 1, \cdots, t$

$k = 1, \cdots, n_i$

$$S_{ww} = \sum_{ijk} (w_{ijk})^2 - CF$$

$$CF = n...\bar{w}.^2.$$

$$H_{ww} = \sum_{ij} \frac{1}{n_{ij}} [w_{ij}.]^2 - CF$$

$$P_{ww} = \sum_{j} \frac{1}{n} \cdot j \cdot [w \cdot j \cdot]^2 + \sum_{i} q_i \phi_i - CF$$

$$B_{ww} = \sum_{i} \phi_i \, q_i \quad \text{(obtained from Doolittle table)}$$

$$T_{ww} = \sum_{j} \phi_j \, q_j$$

$$D_{ww} = H_{ww} - P_{ww}$$

$$E_{ww} = S_{ww} - H_{ww}$$

Let V_{ijk} be a variable different from w_{ijk},

$$S_{wv} = \sum_{ijk} w_{ijk} \, v_{ijk} - CCF$$

$$CCF = n... \, \bar{w}... \, \bar{v}...$$

$$H_{wv} = \sum_{ij} w_{ij}. \, v_{ij}. - CCF$$

$$P_{wv} = \sum_{j} w_{.j}. \, v_{.j}. + \sum_{i} \phi_i \, q_i$$

$$B_{wv} = \sum_{i} \phi_i q_i \quad \text{(obtained from Doolittle table)}$$

$$T_{wv} = \sum_{j} \phi_j \, q_j$$

$$D_{wv} = H_{wv} - P_{wv}$$

$$E_{wv} = S_{wv} - H_{wv}$$

the row sum of squares after adjusting for column effects, and T_{ww} represents the column sum of squares after adjusting for row effects. Either B_{ww} or T_{ww} is obtained by the Doolittle method (13). Table 9 presents the required format for employing the Doolittle method for determining B_{ww} and B_{wv}. Table 12 presents the results of employing the Doolittle method to an example. B_{ww} is calculated by

$$B_{ww} = \sum_i \phi_i q_i$$

where ϕ_j and q_j are obtained from Table 12. T_{ww} can be obtained from the following relationship:

$$\sum_j \frac{w_{\cdot j \cdot}^2}{n_{\cdot j \cdot}} + B_{ww} = \sum_i \frac{w_{i \cdot \cdot}^2}{n_{i \cdot \cdot}} + T_{ww} \quad . \tag{10}$$

and T_{wv} can be obtained from the following relationship:

$$\sum_j w_{\cdot j \cdot} v_{\cdot j \cdot} + B_{wv} = \sum_i w_{i \cdot \cdot} v_{i \cdot \cdot} + T_{wv} \quad , \tag{11}$$

where $B_{wv} = \sum_{ii'} \phi_i q_i$.

EXAMPLE OF A RANDOMIZED BLOCK DESIGN WITH UNEQUAL SAMPLE SIZES

A researcher, working for a well-known organization, wanted to determine some penetration properties of projectiles with various nose shapes against armor plating. He decided on four nose shapes and two types of armor plating. After securing the four types of projectiles, it was noticed that the weight of the projectiles varied by shape. His original idea was to eliminate the influence of projectile weight by having all shapes contain the same mass. Further, he knew that equal amounts of propellant will not necessarily give the same velocity to like projectiles. Not wanting the influence of the two variables, weight (Z_1) and velocity (Z_2), in his results, the data was reduced using the analysis of covariance method.

The data is presented in Table 10. Totals for the raw data are presented in Table 11. The experimental unit is the projectile mass and is subjected to four shapes (treatments). The response variable is the weight of the projectile after penetrating the armor plating. By using the equations given in Table 8 and the values given in Table 11, one is then able to construct the analysis of covariance table (Table 13). The Doolittle values, B_{ww}, are obtained from Table 12,

TABLE 9. DOOLITTLE TABLE - COLUMNS SWEPT OUT

						Y	Z_1	Z_2
$n_{.1.}$	0	0	0	$n_{11.}$	$n_{21.}$	$C_{.1.}$	$C_{.1.}$	$C_{.1.}$
	$n_{.2.}$	0	0	$n_{12.}$	$n_{22.}$	$C_{.2.}$	$C_{.2.}$	$C_{.2.}$
		$n_{.3.}$	0	$n_{13.}$	$n_{23.}$	$C_{.3.}$	$C_{.3.}$	$C_{.3.}$
			$n_{.4.}$	$n_{14.}$	$n_{24.}$	$C_{.4.}$	$C_{.4.}$	$C_{.4.}$
				$n_{1..}$	0	$R_{1..}$	$R_{1..}$	$R_{1..}$
					$n_{2..}$	$R_{2..}$	$R_{2..}$	$R_{2..}$

Where $C_{.j.}$ stands for a column (treatment) total

$R_{i..}$ stands for a row (block) total

for the indicated variable.

TABLE 10. RAW DATA TABLE FOR EXAMPLE 2

OBS	Metal	Shape	Z_1	Z_2	Y
1	A	C	113.8	677	113.7
2	A	C	113.2	589	113.0
3	A	C	114.0	556	114.0
4	A	C	114.1	880	113.4
5	A	C	112.9	331	112.9
6	A	C	113.7	319	113.7
7	A	C	113.2	236	113.2
8	A	C	112.8	458	112.3
9	A	C	112.8	405	112.7
10	A	C	113.5	589	113.5
11	A	C	113.9	570	113.9
12	A	C	114.1	557	114.1
13	A	C	114.1	529	114.0
14	A	C	113.6	512	113.2
15	A	S	117.5	965	116.8
16	A	S	116.8	993	116.1
17	A	S	118.5	959	118.0
18	A	S	117.4	853	117.4
19	A	S	116.7	786	116.5
20	A	S	117.7	704	117.4
21	A	S	118.3	626	118.2
22	A	S	118.0	604	117.9
23	A	S	117.7	564	117.7
24	A	S	117.2	431	117.1
25	A	S	118.0	371	117.9
26	A	S	118.4	316	118.3
27	A	T	111.2	372	110.5
28	A	T	111.0	365	110.9
29	A	T	110.7	278	110.7
30	A	T	109.7	414	109.7
31	A	T	109.2	499	109.1
32	A	T	112.7	565	112.7
33	A	T	114.9	924	114.8
34	A	T	112.9	857	112.6
35	A	C	111.1	514	111.1
36	A	C	111.1	410	111.1
37	A	C	111.5	368	111.4
38	A	C	111.3	356	111.3
39	A	C	110.9	306	110.9
40	A	C	110.8	845	110.9
41	A	C	110.4	903	110.1
42	A	C	110.9	905	109.6
43	A	C	110.5	872	108.6
44	A	C	111.9	731	111.9
45	A	C	110.5	700	110.3

TABLE 10. RAW DATA TABLE FOR EXAMPLE 2 (CONTINUED)

OBS	Metal	Shape	Z_1	Z_2	Y
46	A	C	110.3	703	110.3
47	A	C	109.3	593	109.3
48	A	C	107.8	582	107.7
49	S	C	113.7	780	82.9
50	S	C	114.0	822	83.7
51	S	C	114.0	845	85.5
52	S	C	113.5	881	87.0
53	S	C	113.8	777	72.5
54	S	C	113.3	870	97.2
55	S	C	112.2	895	96.1
56	S	C	112.3	918	97.9
57	S	C	112.3	938	96.0
58	S	C	112.2	962	98.1
59	S	C	112.3	1016	96.9
60	S	C	112.4	1030	99.0
61	S	C	112.6	1091	94.0
62	S	C	112.1	1104	93.5
63	S	S	117.8	871	85.6
64	S	S	116.7	925	111.9
65	S	S	117.4	926	91.3
66	S	S	116.9	957	94.5
67	S	S	117.0	980	94.9
68	S	S	117.0	1002	93.4
69	S	S	117.5	870	82.6
70	S	S	117.1	871	84.9
71	S	S	117.7	833	83.1
72	S	S	118.3	802	78.2
73	S	S	117.7	783	73.0
74	S	T	113.5	943	92.5
75	S	T	113.3	888	78.3
76	S	T	113.0	896	77.1
77	S	T	114.2	859	68.7
78	S	C	114.4	955	90.4
79	S	C	113.8	862	86.9
80	S	C	113.9	956	92.3
81	S	C	114.1	871	84.6
82	S	C	111.3	854	66.1
83	S	C	111.2	880	70.9
84	S	C	110.9	916	80.0
85	S	C	111.0	944	84.9
86	S	C	112.3	991	94.4
87	S	C	114.8	825	64.3
88	S	C	110.4	849	92.0

TABLE 10. RAW DATA TABLE FOR EXAMPLE 2 (CONCLUDED)

OBS	Metal	Shape	Z_1	Z_2	Y
89	S	C	114.7	926	106.3
90	S	C	113.7	934	105.7
91	S	C	112.7	1047	107.7
92	S	C	112.3	1127	108.1
93	S	C	113.0	1143	104.1
94	S	C	114.3	1112	110.0
95	S	C	103.8	982	96.3

TABLE 11. TABLE OF DATA TOTALS

TABLE FOR n

METAL	SHAPE C	S	T	O	
A	14	12	8	14	48
S	14	11	4	18	47
	28	23	12	32	95

TABLE FOR Z_1

METAL	SHAPE C	S	T	O	
A	1589.7	1412.2	892.3	1548.3	5442.5
S	1580.7	1291.1	454	2022.6	5348.4
	3170.4	2703.3	1346.3	3570.9	10790.9

TABLE FOR Z_2

METAL	SHAPE C	S	T	O	
A	7208	8172	4274	8788	28442
S	12929	9820	3586	17174	43509
	20137	17992	7860	25962	71951

TABLE FOR y

METAL	SHAPE C	S	T	O	
A	1587.6	1409.3	891	1544.5	5432.4
S	1280.3	973.4	316.6	1645	4215.3
	2867.9	2382.7	1207.6	3189.5	9647.7

$\sum_{ijk} Z_1 Z_2 = 8,178,973.4$

$\sum_{ijk} Z_1 y = 1,095,882.6$

$\sum_{ijk} Z_2 y = 7,160,083.4$

TABLE 12. DOOLITTLE TABLE FOR EXAMPLE 2

						y	z_1	z_2	
28	0	0	0	14	14	2867.9	3170.4	20137	
	23	0	0	12	11	2382.7	2703.3	17992	
		12	0	8	4	1207.6	1346.3	7860	
			32	14	18	3189.5	3570.9	25962	
				48	0	5432.4	5442.5	28442	
					47	4215.3	5348.4	43509	
28	0	0	0	14	14	2867.9	3170.4 •	20137	
1	0	0	0	0.5	0.5	102.425	113.2285714	719.1785714	
	23	0	0	12	11	2382.7	2703.3	17992	
	1	0	0	0.5217391304	0.4782608696	103.5956522	117.5347826	782.2608696	
		12	0	8	4	1207.6	1346.3	7860	
		1	0	0.666666667	0.33333333	100.6333333	112.1916666	655	
			32	14	18	3189.5	3570.9	25962	
			1	0.4375	0.5625	99.671875	111.590625	811.3125	
				23.2807971	-23.2807971	554.8292573	-12.91947458	-7612.005434	
					23.2807971	-554.8292573	12.91947460	7612.005434	
				23.7807971	-22.7807971	554.8292573	-12.91947458	-7612.005434	
					23.7807971	-554.8292573	12.91947460	7612.005434	
				23.7807971	-22.7807971	554.8292573	-12.91947458	-7612.005434	(q_2)
				1	-.9579492649	23.33097814	-0.5432734036	-320.0904243	(ϕ_2)
					1.957949264	-23.33097813	0.5432734032	320.0904241	(q_1)
					1	-11.91602794	0.2774706236	163.4824916	(ϕ_1)
						13222.72186	7.169549333	2488859.228	

TABLE 13. ANALYSIS OF COVARIANCE TABLE FOR EXAMPLE 2

Source of Variation	d.f.	Sum of Products					
		Z_1^2	Z_1Z_2	Z_2^2	Z_1Y	Z_2Y	Y_2
Total	94	719.49726	6,172.91684	5,189,060.35789	15.59021	-146,881.47053	19,800.29537
Due to Pull	7	551.44646	5,150.18556	2,889,053.82687	29.56663	-179,640.4512	14,015.415761
Due to Reduced	5	520.1211393	5,135.59043	2,762,818.169	4.39419992	-183,290.1917	13,463.36281
Metal Adj	1	7.169549333	4,224.21579	2,488,859.228	-307.8976401	-181,409.7388	13,222.72186
Shape Adj	3	516.1228193	1,888.997440	126,621.132	233.2639299	2,549.46289	362.52978
Difference	2	31.3253207	14.59513	126,235.65787	25.17243008	3,649.7405	552.052951
Residual	87	168.05081	1,022.73128	2,300,006.53102	-13.97642	32,758.98067	5,784.97961

	d.f.	$\hat{\beta}_1$	$\hat{\beta}_2$	$\Sigma y^2 - \hat{\beta}_1\Sigma z_1 y - \hat{\beta}_2\Sigma z_2 y$	Adjusted d.f.	Adjusted Mean Squares
Residual	87	-0.17031	0.01432	5,313.49068272	85	62.5116550908
Difference + Residual	89			5,817.56646632	87	66.8685800726
Difference*					2	4.3569249818

$$U = \frac{4.3569249818}{62.5116550908} = 0.0697$$

$F(2,85)$ for $(\alpha = .05) = 3.105$

and T_{ww} and T_{wv} values are obtained by employing equations (10) and (11). It is much easier to show how B_{wv} is calculated than to try to explain. Refer to Table 12 and the Z_1 and Z_2 columns.

$$B_{Z_1Z_2} = \sum_{ii'} \phi_i \, q_{i'}$$

$$= (.2774706236)\ (320.0904241)$$

$$+ (-.5432734036)(-7612.005434)$$

$$= 4,224.21579$$

After obtaining the sum of the products, one may now solve for the concomitant coefficients by:

$$E_{z_1z_1} \hat{\beta}_1 + E_{z_1z_2} \hat{\beta}_2 = E_{z_1y}$$

$$168.05081\hat{\beta}_1 + 1022.73128\hat{\beta}_2 = -13.97642$$

and

$$E_{z_1z_2} \hat{\beta}_1 + E_{z_2z_2} \hat{\beta}_2 = E_{z_2y}$$

$$1022.73128\hat{\beta}_1 + 2,300,006.53102\hat{\beta}_2 = 32,758.98067$$

thus obtaining

$$\hat{\beta}_1 = -.17031 \qquad \text{and} \qquad \hat{\beta}_2 = .01432$$

The comparison of the test statistic U to a tabulated F (2, 85) at the 95 percent level indicates that the Null Hypothesis is not rejected. The interaction term need not be considered in the building of a predictive model. Table 14 contains the means for any comparisons that one may want to make.

If one wished to pursue the problem further, a test for treatments and blocks may be made and corresponding adjusted means may be calculated.

Table 14 contains the unadjusted and adjusted means for the response variable.

TABLE 14. TABLE OF MEANS

UNADJUSTED $\bar{Y}$

	C	S	T	O	
A	113.4	117.4417	111.375	110.3214	113.175
S	91.45	88.4909	79.15	91.3889	89.6872
	102.425	103.5957	100.6333	99.6719	101.5547

ADJUSTED $\bar{Y}$

	C	S	T	O	
A	116.8661	119.2327	114.2206	111.6679	115.5007
S	88.9553	87.1974	77.1429	88.3639	87.3121
	102.9107	103.9115	101.8614	98.5594	101.5547

SECTION IV

COVARIANCE ANALYSIS AS A TECHNIQUE FOR ANALYZING INCOMPLETE DATA

INTRODUCTION

In Section I, it was stated that one of the principal uses of covariance analysis was to analyze data when some responses are missing. The covariance analysis technique is an alternative method of predicting missing values to that described by Snedecor and Cochran (12) under missing data. Both techniques apply to data containing missing responses that are to be analyzed by the analysis of variance method. The covariance missing data technique presented here does not apply to predicting missing values for analysis of covariance data, responses or covariates.

M. S. Bartlett introduced the concept of using covariance analysis on missing data. The reason why an alternative method was sought was because no general algorithm exists for dealing with missing values. Special formulae exist for each randomization scheme, and adjusting for the bias becomes tedious.

This section is based on an article by Coons (4) in which the author presents a general method to the problem of missing data and also demonstrates the case with which exact tests of significance may be obtained. The tests are exact when the errors are assumed to be independent and normally distributed.

PROPERTIES FOR JUSTIFYING THE COMPUTATIONAL PROCEDURES

The following properties are quoted from Coons' article and are attributed to various individuals. The article indicated that Property 1 is attributed to Fisher, Property 2 is implicitly assumed by several authors, Property 3 to Bartlett, and Properties 4, 5, and 6 to Kempthorne.

1. If an analysis of variance is made with symbols $\beta_1, \beta_2, \cdots \beta_q$ in the place of missing observations, then the best linear unbiased estimates of the missing observations are the quantities $\hat{\beta}_1, \beta_2, \cdots, \hat{\beta}_q$ which minimize the error sum of squares.

2. Given that, with full data $(y_1, y_2, \cdots, y_n)$, the best linear unbiased estimate of some linear function of the parameters is $v_1y_1 + v_2y_2 + \cdots + v_ny_n$, then the best estimate of that function with missing data is obtained by replacing the missing y's with the missing value estimates.

3. Let the data be observed data where obtained and zero where missing. Introduce a concomitant variable $X_m (m = 1 \cdots q)$ corresponding to the mth missing observation; let X_m take the value -v for the mth missing observation and zero for all others, missing or not. If the error partial regression coefficients obtained from an analysis of covariance are denoted by $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_q$, then $v\hat{\beta}_1, v\hat{\beta}_2, \cdots, v\hat{\beta}_q$ are the best linear unbiased estimates of the missing observations.

4. Estimates of functions of data with missing observations, and variances and covariances of these estimates may be obtained by the routine application of formulae for adjusted means in the analysis of covariance; i.e., by regarding the zero yields supplied in the analysis of covariance procedure as having variances of σ^2. The above statement applies to functions of the augmented data; the variance of a missing observation per se is given by statement 5 following.

5. Denote the error sum of squares of X_i by E_{ii} and the error sum of products of X_i and X_j by E_{ij}. Then the variance of the i^{th} missing value estimate is $(v^2u_{ii} - 1)\sigma^2$, and the covariance of the i^{th} and j^{th} missing value estimates is $v^2u_{ij}\sigma^2$, when

$$\begin{bmatrix} E_{11} & E_{12} & \cdots & E_{1q} \\ E_{21} & & & \\ \vdots & & & \\ E_{q1} & & & E_{qq} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1q} \\ u_{21} & & & \\ \vdots & & & \\ u_{q1} & & & u_{qq} \end{bmatrix} = \begin{bmatrix} 1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}.$$

6. The sum of squares for treatments obtained by analyzing the data augmented by the missing value estimate is always greater than or equal to the exact sum of squares for treatment.

COVARIANCE TECHNIQUE APPLIED TO ONE MISSING OBSERVATION

The covariance technique will be discussed as the following problem is being worked.

Due to the world hunger problem, it has become important to try to recover farm land in countries where herbicides were used during recent military actions. Neutralizers were added to the soil samples collected from various regions. Grain crops were then planted in the treated soils to determine how much of the toxin in herbicides would be passed on to humans and animals. It was decided to randomize the experiment in a 4x4 Latin square and take two observations per condition. The results of treating one herbicide is given in Table 15.

The experimental unit is a pot containing a plant. Applying the covariance technique, the covariate, z, would take the value zero for all responses, y, not missing and -n with the missing response. There are 32 observations including the missing value, so n = 32. Other authors have suggested that any convenient value may be assigned as the covariate to the missing response because z and y are unrelated. Using -n simplifies calculations for any line entry in the analysis of covariance table for the covariate sum of squares is simply n x (degrees of freedom). The missing response takes the value zero, as stated in property 3, and the non-missing responses retain their values. Table 16 shows how the technique is applied.

With a single degree of freedom, the line entry for each of the cross product sum of squares is

$$X_1 - X_2$$

where X_1 is the total of Y observations for the effect level which does not contain the missing observation, and X_2 is the total of Y observations for the effect level which contains the missing observation. For line entries containing more than one degree of freedom,

$$\Sigma zy = \sum_i (x_{i_1} - X_{i_2}).$$

(See Table 16.) The calculations of Σy^2 are as usual and will not be shown. β is estimated by $\hat{\beta}$. An estimate of the missing value is given by Property 3 to be $n\hat{\beta}$. It is not necessary to estimate the missing value since a complete analysis of the data may be performed with the value remaining unknown. The covariance technique enables one to make exact tests readily with only minor supplementary computations.

An approximate test of significance may be obtained by computing the biased sum of squares which is equivalent to an analysis of $Y - \hat{\beta}Z$. Property 6 states that the approximate sum of squares is greater than, or equal to, the exact sum of squares. Therefore, any approximate mean square which is not significant may be eliminated from consideration and thereby shorten the calculations. The approximate sum of squares may be computed as

$$\Sigma y^2 - 2\hat{\beta}\,\Sigma zy + \hat{\beta}^2\,\Sigma z^2 .$$

TABLE 15. DATA TABLE FOR EXAMPLE 3

SOIL	PLANTS			
	P1	P2	P3	P4
S1	A 91 M	D 105 100	C 52 61	B 12 9
S2	C 73 65	B 2 7	D 112 110	A 93 91
S3	D 102 111	A 89 91	B 3 7	C 54 59
S4	B 10 8	C 52 77	A 92 90	D 103 108

SOIL - S1 - Sand

- S2 - Sand + Herbicide

- S3 - Clay

- S4 - Clay + Herbicide

PLANT - P1 - Wheat

- P2 - Rice

- P3 - Grass

- P4 - Barley

NEUTRALIZERS - A

- B

- C

- D - Nothing

The response is the average amount of herbicide toxin found in the grains of each plant, measured in count per million.

TABLE 16. COMPUTATIONAL TABLE FOR EXAMPLE 3

Soils	Plants								ΣS_i
	P1		P2		P3		P4		
	y	z	y	z	y	z	y	z	
S1	A 91 0	0 -32	D 105 100	0 0	C 52 61	0 0	B 12 9	0 0	430
S2	C 73 65	0 0	B 2 7	0 0	D 112 110	0 0	A 93 91	0 0	553
S3	D 102 111	0 0	A 89 91	0 0	B 3 7	0 0	C 54 59	0 0	516
	B 10 8	0 0	C 52 77	0 0	A 92 90	0 0	D 103 108	0 0	540
ΣP_i	460		523		527		529		2039

$\Sigma A = 637$
$\Sigma B = 58$

$\Sigma C = 493$
$\Sigma D = 851$

Each Σz^2 line entry = n X (degrees of freedom)

Total $\Sigma z^2 = (32)\ (31) = 992$

Σzy line entry = $\sum_i (Y_i - Y_2)$

Soil $\Sigma zy = (553 - 430) + (516 - 430) + (540 - 430) = 319$

$\hat{\beta} = E_{zy}/E_{zz} = 2030/704 = 2.88$

Missing Value Estimated = $n\hat{\beta} = 92$

So for soil,

$$\text{Soil Approx SS} = 1148.0937 + (2.88)^2 (96) - 2 (2.88) (319)$$

$$= 106.9161$$

The approximate mean squares are obtained by dividing the approximate sum of squares by the appropriate degrees of freedom. All of the above calculations are summarized in the analysis of covariance table (Table 17). The adjusted sum of squares is obtained in the usual way. Exact test of significance may now be made on the variations of interest. Estimates of treatment means must be adjusted to the value zero of the covariate variable instead of to the covariate average; i.e.,

$$\text{ADJ } \bar{Y} = \bar{Y} - \hat{\beta}\, \bar{Z}$$

where $\bar{Z}$ is the average of the number of responses making up $\bar{Y}$. Since Treatment A contained the missing value,

$$\text{ADJ } \bar{Y}_A = 79.63 - (2.88)\ (-4) = 91.15.$$

The variance is given by

$$V\ (\text{ADJ } \bar{Y}_A) = \sigma^2/n + (\bar{Z})^2\sigma^2/E_{xx}$$

where σ^2 is estimated by $s^2_{y \cdot x}$. Therefore,

$$V\ (\text{ADJ } \bar{Y}_A) = 32.79\ [1/8 + (-4)^2/704]$$

$$= 4.84.$$

COVARIANCE TECHNIQUE APPLIED TO MORE THAN ONE MISSING OBSERVATION

The application of the technique will be discussed as the following problem is being worked.

A research laboratory received four new growth chambers. Before putting them into use, it was decided to conduct a trial experiment to determine the variations within and among each chamber. Since all chambers were large, it was decided to divide each into three horizontal positions and two vertical positions to determine if location had any effect on plant growth. Six pots containing similar seed, soil, and

TABLE 17. ANALYSIS OF COVARIANCE TABLE FOR EXAMPLE 3

Source	d.f.	Sum of Products			Approximate Mean Square	$\hat{\beta}$	d.f.	Adjusted Sum of Squares	Adjusted Mean Square
		z^2	zy	y^2					
Total	31	992	2039	50230.4687					
Soil	3	96	319	1148.0937	35.64				
Plants	3	96	199	414.8437	21.62				
Treatments	3	96	-509	42125.3437	15284.48				
Error	22	704	2030	6542.1876		2.88	21	688.6365	32.79
Treatment and error	25	800	1521	48667.5313			24	45,775.73	
Difference							3	45,087.09355	

nutrients were randomly placed within each chamber. The experiment was replicated twice for two months at a time. The response, plant height in centimeters, was to be analyzed using a split unit analysis in strips.

Table 18 contains the raw data, augmented covariates, and the totals necessary for computations. As before, the number of y observations, including those missing, is equal to n. The value zero is assigned to each missing y observation and to each covariate where the y observation is not missing. For covariate values associated with missing observations, the value of -n is assigned. When more than one observation is missing, a multiple covariance analysis is needed. There will be one covariate for each missing value.

The computations for Σz_1^2 and Σz_2^2 are the same as before. The one column entry Σz_i^2 will suffice for Σz_1^2 and Σz_2^2. Two situations may occur in computing $\Sigma z_1 z_2$:

1. When Z_1 and Z_2 occur in the same level, the results are the same as for Σz_i^2

$$\Sigma z_1 z_2 = n \times \text{(degrees of freedom)}.$$

2. When z_i and z_j occur in different levels,

(i) for no interactinn levels,

$$\Sigma z_i z_j = -nr,$$

where r depends upon the hierarchy classification. With no hierarchy, r = 1.

(ii) for levels in which there is interaction, the main effects and lower order interactions must be subtracted from

$$\Sigma z_i z_j = -nr.$$

The author was unable to obtain Coons' results when following his computational methods, so the usual method for obtaining sum of squares was employed. Table 19 contains an example of the computations for the cross products needed in building the analysis of covariance table (Table 20). The line entries for z y cross product sum of squares is obtained as before,

$$\Sigma zy = \sum_i (x_{i_1} - x_{i_2}).$$

TABLE 18. TABLE OF TOTALS FOR EXAMPLE 4

REP 1

	Chamber 1				Chamber 2				Chamber 3				Chamber 4				Totals		
Levels	V_1	V_2	Z_1	Z_2	V_1	V_2	Z_1	Z_2	V_1	V_2	Z_1	Z_2	V_1	V_2	Z_1	Z_2	H	Z_1	Z_2
H_1	23	21	0	0	20	18	0	0	21	19	0	0	20	25	0	0	167	0	0
H_2	19	17	0	0	16	14	0	0	17	15	0	0	21	19	0	0	138	0	0
H_3	8	6	0	0	5	3	0	0	6	4	0	0	10	8	0	0	50	0	0
Totals	50	44	0	0	41	35	0	0	44	38	0	0	51	52	0	0	355	0	0

REP 2

	Chamber 1				Chamber 2				Chamber 3				Chamber 4				Totals		
Levels	V_1	V_2	Z_1	Z_2	V_1	V_2	Z_1	Z_2	V_1	V_2	Z_1	Z_2	V_1	V_2	Z_1	Z_2	H	Z_1	Z_2
H_1	24	21	0	0	21	19	0	0	21	17	0	0	25	23	0	0	171	0	0
H_2	18	16	0	0	15	14	0	0	16	M	-48	0	19	16	0	0	114	-48	0
H_3	10	7	0	0	8	7	0	0	6	2	0	0	M	13	0	-48	53	0	-48
Totals	52	44	0	0	44	40	0	0	43	19	-48	0	44	52	0	-48	338	-48	-48

REP X CHAMBER

	CHM 1	Z_1	Z_2	CHM 2	Z_1	Z_2	CHM 3	Z_1	Z_2	CHM 4	Z_1	Z_2	Totals		
	6														
REP 1	94	0	0	76	0	0	82	0	0	103	0	0	355	0	0
REP 2	96	0	0	84	0	0	62	-48	0	96	0	-48	338	-48	-48
TOTALS	190	0	0	160	0	0	144	-48	0	199	0	-48	693	-48	-48

FACTOR	V	Z_1	Z_2	V	Z_1	Z_2	TOTALS		
	4								
H_1	175	0	0	163	0	0	338	0	0
H_2	141	0	0	111	-48	0	252	-48	0
H_3	53	0	-48	50	0	0	103	0	-48
TOTALS	369	0	-48	324	-48	0	693	-48	-48

TABLE 18. TABLE OF TOTALS FOR EXAMPLE 4 (CONCLUDED)

MAIN A UNIT X CHAMBER X REP

REP	MAIN UNIT	CHM 1	Z_1	Z_2	CHM 2	Z_1	Z_2	CHM 3	Z_1	Z_2	CHM 4	Z_1	Z_2
		2											
	H_1	44	0	0	38	0	0	40	0	0	45	0	0
1	H_2	36	0	0	30	0	0	32	0	0	40	0	0
	H_3	14	0	0	8	0	0	10	0	0	18	0	0
	H_1	45	0	0	40	0	0	38	0	0	48	0	0
2	H_2	34	0	0	29	0	0	16	-48	0	35	0	0
	H_3	17	0	0	15	0	0	8	0	0	13	0	-48
		190	0	0	160	0	0	144	-48	0	199	0	-48

MAIN B UNIT X CHAMBER X REP

REP	MAIN UNIT	CHM 1	Z_1	Z_2	CHM 2	Z_1	Z_2	CHM 3	Z_1	Z_2	CHM 4	Z_1	Z_2
		3											
	V_1	50	0	0	41	0	0	44	0	0	51	0	0
1	V_2	44	0	0	35	0	0	38	0	0	52	0	0
	V_1	52	0	0	44	0	0	43	0	0	44	0	-48
2	V_2	44	0	0	40	0	0	19	-48	0	52	0	0

TABLE 19. COMPUTATIONAL TABLE FOR EXAMPLE 4

The Z_1Z_2 cross product sum of squares is obtained by using the appropriate cross product table. Using the main A unit X chamber X rep table, the main A unit analysis is obtained:

$$\Sigma Z_1 Z_2 = \frac{1}{2}[(0)(0) + \cdots + (-48)(0) + \cdots + (0)(-48)] - \frac{1}{48}(-48)(-48)$$

$$= -48.$$

The z_1y cross product sum of squares for main B unit is

$$\Sigma z_1 y = \sum_i (x_{i_1} - x_{i_2})$$

$$= (50 - 19) + (44 - 19) + \cdots + (52 - 19) - \text{Reps} - \text{Chambers}$$

$$= 255.$$

Estimates of the missing values are as follows:

Missing	Main A Unit	Main B Unit	Subunit AB
z_1	13.44	21.12	7.2
z_2	4.8	12	18.24

TABLE 20. ANALYSIS OF COVARIANCE TABLE FOR EXAMPLE 4

Source	d.f.	Sum of Products					$\hat{\beta}_1$	$\hat{\beta}_2$	Approximate S. S.	Adjusted d.f.	Adjusted Sum of Squares	Adjusted Mean Square
		z_i^2	z_1z_2	z_1y	z_2y	y^2						
Total	47	2256	-48	693	693	2331.8125						
Main Unit A Analysis	23	1104	-48	309	381	2048.3125						
Reps	1	48	48	17	17	6.0208						
Chambers	3	144	-48	117	-103	164.5625			132.3721			
Horizontal	2	96	-48	-63	384	1767.125			1734.0914			
Error A	18	864	0	238	89	110.6042	.28	.10		17	35.0642	2.0626
Main Unit B Analysis	15	528	-48	255	75	175.2297						
Vertical	1	48	-48	45	-45	42.1875			37.3803			
Error B	14	480	0	210	120	133.0422	.44	.25		13	10.6422	0.81863
Subunit AB Analysis	9	624	48	129	237	108.2703						
HV	2	96	-48	47	36	23.625			-2.9225			
Error C	7	538	0	82	201	84.6453	.15	.38		6	-4.0347	-0.67245

For the y^2 sum of squares, one follows the same procedures as in an analysis of variance table. Note that, for the split unit in strips with two main units, an adjustment is made in calculating the main unit sum of squares. Looking at the main A unit X chamber X reps table and the main B unit X chamber X reps table, the entries' chambers and reps are included in both main unit calculations. Since accounting for them once, they must be removed from the interaction unit. In this example, chamber and reps sum of squares were subtracted out in the $z_1 z_2$ column.

Values for $\hat{\beta}$ are obtained by solving the appropriate set of equations as explained in Section III. Missing values are estimated by

$$y_m = n\hat{\beta}_E = n \times (\hat{\beta} \text{ associated with the missing observation for the particular level}).$$

Obtaining the approximate sum of squares may again help reduce computations. For multivariate data, the approximate sum of squares is computed by

$$[\underset{\sim}{Y} - z\underset{\sim}{\hat{\beta}}]' \; [\underset{\sim}{Y} - z\underset{\sim}{\hat{\beta}}]$$

$$\underset{\sim}{Y}' \underset{\sim}{Y} - \underset{\sim}{Y} z\underset{\sim}{\hat{\beta}} - \underset{\sim}{\hat{\beta}}' z' \underset{\sim}{Y} + \underset{\sim}{\hat{\beta}} z' z \underset{\sim}{\hat{\beta}} ,$$

and for this example,

$$\Sigma y^2 - 2\hat{\beta}_1 \Sigma z_1 y - 2\hat{\beta}_2 \Sigma z_2 y + \hat{\beta}_1^2 \Sigma z_1^2 + \hat{\beta}_2^2 \Sigma z_2^2 .$$

The adjusted sum of squares for the error terms is obtained by,

$$\Sigma y^2 - \hat{\beta}_1 \Sigma z_1 y - \hat{\beta}_2 \Sigma z_2 y$$

where the $\hat{\beta}_i$ and $\Sigma z_i y$ correspond to the appropriate level. When testing line entries, follow the normal covariance analysis procedure and use the appropriate error term.

SECTION V

COVARIANCE ANALYSIS FOR NON-PARAMETRIC DATA

INTRODUCTION

Bross (2) put forth a non-parametric procedure for handling data by means of covariance analysis. The procedure, the Covariable Adjusted Sign Test (COVAST), is designed for detecting differences between two treatments having binary responses with a single covariate. The assumptions are:

(a) The covariable and response have a monotone relationship.

(b) The observations are independent.

(c) The measurement scale of the covariate is at least ordinal.

In practice, subjects are divided into two subsets such that the individuals in each set possess covariate values which are representative of the covariate range. Treatments are applied to subjects in each subset, and it is expected that the portion of subjects responding to a treatment is 0.5. Ury (9) recognized that the expected portion in each subset may not be 0.5 and expanded the work of Bross to include these cases.

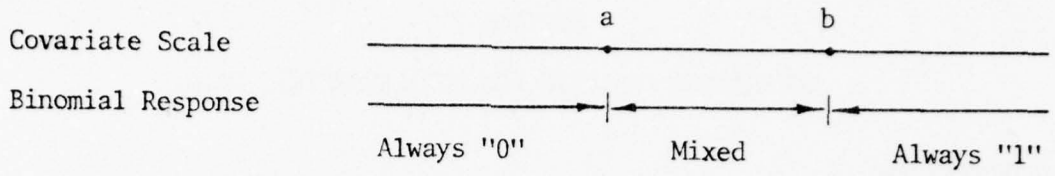
Quade (6) develops a procedure called "Rank Analysis of Covariance" designed for handling treatment differences in responses measured on at least an ordinal scale and having one or more covariates. The procedure compares to a completely randomized analysis of covariance. He also discusses other methods developed along this line. Puri and Sen (5) develop a theoretical approach to the completely randomized case.

The procedures of Bross and Ury will be presented in this section along with an example using real data. The other procedures will not be included in this report.

THE COVAST TEST

Rationale

Suppose one is faced with a situation where the result of an event is a binomial response. Let this event be associated with a variable, measured on an ordinal scale or better, which will have a changing influence on the response of the event. Consider the following illustration:



At point a and below on the covariate measurement scale, the response is always the same. At point b and above, the response is always the same but different from the response at a. For the interval (a,b), the responses are mixed. For example, babies of a certain weight (covariate) may live or die (event) when afflicted with a certain disease. Another example may be combustion or non-combustion (event) at a given temperature (covariate).

One may then be interested in determining if there is a statistically significant difference between two treatments under the situation being considered. A treatment may be a drug cure to the disease or an ignitor for stimulating combustion. To see how the covariate is taken into account for comparison tests, one needs to assume the following:

1. That one treatment is better than the other.

2. That the chance for an improvement increases either as the covariate increases or as it decreases.

The words "better" and "an improvement" may be understood in terms of ordering the observed values of the covariate from values less that a to greater than b where the response at the a end of the scale represents an unfavorable response. Suppose two ignitors, H and M, are being compared to determine whether M is significantly better than H for starting fires. If the outcomes are the same for both treatments regardless of temperature, no evidence is provided for a clear-cut superiority. If one treatment started fires at high temperatures and the other did not start a fire at low temperatures, the results might be attributed to the initial conditions rather than to the treatments.

However, if one treatment starts fires at low temperatures and the other treatment does not start fires at high temperatures, then this would be evidence (but not conclusive) for an advantage to the treatment which does start fires. One can compare the performance of the two treatments by making pairwise comparisons. The comparisons would be made on the basis of the following:

1. One of the ignitors starts a fire, and

2. The fire was started at a lower temperature.

In order to show a definite advantage for M, it must be shown that M's ability to start fires is greater than that expected from sampling

variation alone. This is accomplished by counting the number of instances where M starts fires at a lower temperature and H does not start fires at higher temperatures. Let these situations be designated "non-inversions" (NI). The opposite situation would be to count the number of instances where H started fires at lower temperatures, and M does not start fires at higher temperatures. Let these situations be designated "inversions" (I). NI and I may now be compared, and if the value of NI is found to be greater than its expectation, then one would have direct evidence of an advantage for M.

Hypothesis

Let $i = 0$ if H is used; $j = 0$ if no fire

$= 1$ if M is used; $= 1$ if fire

and let N_{ij} be the number of observations in the i^{th} series having the j^{th} response. Let I_{kg} be the number of inversions where fires started in the k^{th} series are compared to no-fires in the g^{th} series:

$k = 0$ if H is used; $g = 0$ if H is used

1 if M is used; $= 1$ if M is used.

The covariate complicates the hypothesis statement because of the fact that it determines the ordering which affects the inversions. As a result of this complication, we must test a compound hypothesis. First, consider the hypothesis by parts and then as combined

$H_{0\cdot}$: the two treatments are equally effective

$H_{\cdot 0}$: the covariable is irrelevant

H_{00} : neither the treatments nor the covariate are relevant to the event.

The respective alternative hypothesis may be stated as follows:

$H_{1\cdot}$: the two treatments are not equally effective

$H_{\cdot 1}$: the covariable is relevant

H_{11} : the treatment or the covariate is relevant.

The above compound hypotheses, H_{00} and H_{11}, would be used for a two-tailed test. The following compound hypotheses are used for a one-tailed test.

H_{00} : Treatment 1 is equivalent to treatment 2, and the covariable is not important.

H'_{11} : Treatment 1 is less (greater) than treatment 2, or the covariable is important or both.

Less (greater) may be interpreted as being better or an improvement. The way the test statistic is taken will determine if the hypothesis is for an upper or lower tailed test.

Test Statistic

Under $H_{0\cdot}$ and H_{00}, with the statement of no treatment difference, one would expect the portion of events occurring for the k^{th} and g^{th} series to be the same. Let r be the proportion of events favoring the k^{th} series and (1-r) be the proportion of events favoring the g^{th} series. Therefore, their expected proportions would be

$$E\ (r) = E\ (1\text{-}r) = 0.5$$

which implies that we expect $I_{10} = I_{01}$. The alternative hypothesis, H_{11}, is supported when $I_{10} \neq I_{01}$, and the alternative hypothesis, H'_{11}, is supported when $I_{10} > I_{01}$ or $I_{10} < I_{01}$ depending upon the upper or lower one-tailed test.

Bross states that Mann and Whitney (1947) proved that, given the observed values N_{00}, N_{01}, N_{10}, *and* N_{11} *along with* H_{00}, I_{10}, *and* I_{01} *have* the following expected values (E) and variances (V):

$$E\ (I_{10}) = N_{11}\ N_{00}/2$$

$$V\ (I_{10}) = N_{11}\ N_{00}\ (N_{11} + N_{00} + 1)/12$$

$$E\ (I_{01}) = N_{10}\ N_{01}/2$$

$$V\ (I_{01}) = N_{10}\ N_{01}\ (N_{10} + N_{01} + 1)/12$$

where N_{ij} is the number of observations in the i^{th} series having the j^{th} response.

I_{10} and I_{01} involve two distinct sets of data and are therefore conditionally independent provided the original observations were independent. So,

$$E\ (I_{10} - I_{01}) = (N_{11}\ N_{00} - N_{10}\ N_{01})/2$$

$$V\ (I_{10} - I_{01}) = [N_{11}\ N_{00}\ (N_{11} + N_{00} + 1) + N_{10}\ N_{01}\ (N_{10} + N_{01} + 1)]/12.$$

A relationship between COVAST and the chi-square test for independence is suggested because the expected value of $I_{10} - I_{01}$ is one-half the numerator of the short cut form of the chi-square test.

Each N_{ij} is, in reality, a random variable having a binomial distribution with mean $N_{i.}\pi$ and variance $N_{i.}\pi\ (1 - \pi)$ where π is the probability of a fire, and the marginal totals $N_{1.} = N_{11} + N_{10}$, and $N_{0.} = N_{00} + N_{01}$ are fixed. Thus, the expected value and variance of N_{ij} under H_{00} is

$$E^* (N_{ij}) = \pi N_{i.} \text{ if a success occurs}$$

$$= (1-\pi) N_{i.} \text{ if a failure occurs}$$

$$V^* (N_{ij}) = E^* (N_{ij} - \pi N_{i.})^2 = N_{i.}\ \pi\ (1 - \pi).$$

Substituting these into the above expectation:

$$E^* [E\ (I_{10} - I_{01})] = [E^* (N_{11})\ E^* (N_{00}) - E^* (N_{10})\ E^* (N_{01})]/2$$

$$= [\pi N_{1.}\ (1-\pi) N_{0.} - (1-\pi) N_{1.}\ \pi N_{0.}]/2$$

$$= 0$$

$$12E^* [V\ (I_{10} - I_{01})] = E^* (N_{00})\ E^* (N_{11})^2 + E^* (N_{11})\ E^* (N_{00})^2$$

$$+ E^* (N_{11})\ E^* (N_{00}) + E^* (N_{01})\ E^* (N_{10})^2$$

$$+ E^* (N_{10})\ E^* (N_{01})^2 + E^* (N_{10})\ E^* (N_{01})$$

$$= (1 - \pi) N_{0.}\ E^* (N_{11})^2 + \pi N_{1.}\ E^* (N_{00})^2$$

$$+ \pi N_{0.}\ E^* (N_{10})^2 + (1 - \pi) N_{1.}\ E^* (N_{01})^2$$

$$+ 2\pi\ (1 - \pi) N_{1.}\ N_{0.}$$

Based on this value of $E^* [V\ (I_{10} - I_{01})]$, Bross argues that $V\ (I_{10} - I_{01})$ may be estimated by

$$\frac{(I_{10} + I_{01})(N_{..} + 4)}{12}\ .$$

Hence, the statistic

$$\text{COVAST} = \frac{12\ (I_{10} - I_{01})^2}{(I_{10} + I_{01})(N_{..} + 4)}$$

$$= \frac{12\ \text{UST}}{N_{..} + 4}$$

has approximately a chi-square distribution with one degree of freedom. COVAST is then a variation of the Uncorrected Sign Test (UST) and in this form becomes a test statistic for a two-tailed test for non-parametric covariance analysis.

Ury (9) proposes a method of testing a one-sided hypothesis for Bross's COVAST. Ury defines an r value as being "the proportion of comparisons potentially favoring the treatment" considered to be an improvement; i.e.,

$$r = \frac{T_4}{N_{1.}\ N_{0.}}$$

where T_4 is the total of the entries of column 4 in a table such as Table 21. After ranking the treatments, a count is made to see how many times the new treatment ranks below the standard treatment. This count is made for each subject given the new treatment. The expectations of I_{10} and I_{01} under H_{00}, when r is considered, becomes:

$$E\ (I_{10}) = rN_{11}\ N_{00}$$

$$E\ (I_{01}) = (1-r)\ N_{10}\ N_{01}\ .$$

For a given r, r_0, the following conditional expectations hold:

$$E^*\ E\ (I_{10} \mid r_0) = \pi\ (1 - \pi)\ r_0\ N_{1.}\ N_{0.}$$

$$E^*\ E\ (I_{01} \mid r_0) = \pi\ (1 - \pi)\ (1 - r_0)\ N_{1.}\ N_{0.}$$

$$E^*\ E\ (I_{10} + I_{01} \mid r_0) = \pi\ (1 - \pi)\ N_{1.}\ N_{0.}$$

$$E^*\ E\ (I_{10} - I_{01} \mid r_0) = \pi\ (1 - \pi)\ (2r_0 - 1)\ N_{1.}\ N_{0.}\ .$$

TABLE 21. TABULATION DATA FOR EXAMPLE 5

(1) TEMP	(2) IGNITOR	(3) RESULTS	(4) r	F / N	(5) M / H	(6) H / M	(7) M / M	(8) H / H
20.2	M	N	11					
21.4	H	H				9		3
22.0	M	N	10					
22.0	H	F				8		3
23.0	H	N						
25.0	H	F				8		2
26.0	M	N	7					
26.0	M	F	7		2		7	
26.0	M	F	7		2		7	
26.8	M	N	7					
27.2	H	F				6		2
28.0	M	F	6		2		6	
28.8	M	N	6					
29.0	M	N	6					
30.4	H	F				4		2
32.0	M	F	5		2		4	
32.6	M	N	5					
33.0	M	N	5					
33.5	M	N	5					
34.0	M	F	5		2		1	
34.0	H	N						
35.0	M	N	4					
35.0	H	F				0		1
37.0	M	F	3		1		0	
37.0	H	N						
37.0	H	F				0		0

TABLE 21. TABULATION DATA FOR EXAMPLE 5 (CONCLUDED)

(1) TEMP	(2) IGNITOR	(3) RESULTS	(4) r	F / N	(5) M / H	(6) H / M	(7) M / M	(8) H / H
37.0	H	F				0		0
40.0	M	F	0				0	
47.0	M	F	0				0	
		TOTAL	99		11	35	25	13

N_{ij} TABLE

	NO FIRE	FIRE		
H	3	8	11	$N_{0.}$
M	10	8	18	$N_{1.}$
	13	16	29	
	$N_{.0}$	$N_{.1}$	$N_{..}$	

$$r = \text{Col (4)} \Big/ N_{1.}\, N_{0.} = 99 \Big/ (11)(18) = \frac{99}{198} = .50$$

When $r_0 = 0.5$, $E^* E (I_{10} - I_{01}) = 0$ which agrees with Bross. Ury suggests using the square root of COVAST for the one-sided test,

$$C = \left[\frac{12 (I_{10} - I_{01})^2}{(N_{..} + 4)(I_{10} + I_{01})} \right]^{\frac{1}{2}}$$

$$= (I_{10} - I_{01}) \left[\frac{12}{(N_{..} + 4)(I_{10} + I_{01})} \right]^{\frac{1}{2}}$$

Decision Rule

As with the two-tailed Sign Test, the COVAST test statistic for the two-tailed alternative would be compared with the tabulated chi-square with one degree of freedom. Since most chi-square tables are based on a two-tailed distribution, COVAST may be compared directly at the appropriate α level.

Two conditions must be considered for a one-tailed test. If the alternate hypothesis is:

H_{11}: The new treatment mean is less than the standard treatment mean or the covariable is important or both,

then one would expect T_5, the total of column 5 from Table 21, to be less than T_6, the total of column 6; i.e., expect $(I_{10} - I_{01}) < 0$. If it is and if $C < -z_\alpha$, then reject H_{00} where z_α is from the standard normal distribution. If T_5 is greater than T_6, then do not reject H_{00}.

If the alternate hypothesis is:

H_{11}: The new treatment mean is greater than the standard treatment mean or the covariable is important or both,

then one would expect $T_5 > T_6$; i.e., expect $(I_{10} - I_{01}) > 0$. If it is and if $C > z_\alpha$, then reject H_{00}. If $T_5 < T_6$, do not reject H_{00}.

EXAMPLE

An Air Force officer developed a new incendiary material for a standard round and claimed that his was better than those presently in stock. An independent Air Force test group was given the task of conducting a comparison test. Due to a time limitation, it was decided to test the new incendiary against one which was readily available. The test plan called for shooting both incendiary rounds against fuel cells instrumented to give inside temperature readings in degrees centigrade. The fuel cells contained a common fuel and the decision as to a fire or no fire was determined by the project officer. Questionable situations were resolved by using a time history plot of the temperature. Ties in the data occurring while ranking the observations were eliminated by using the time of day a shot occurred. Table 21 presents the data ordered by temperature. The ignitors are represented by an H for the standard and an M for the new material. The results of each shot was a fire (F) or a no fire (N). Columns 5 through 8, respectively, represent the number of times a fire was started by material M at low temperature and material H started no fire at higher temperatures; the number of times material H started a fire at low temperatures and material M started no fire at higher temperatures; the number of times material M started a fire at low temperatures and material M started no fire at higher temperatures; and the number of times material H started a fire at low temperatures and material H started no fire at higher temperatures. The column total for $\frac{M}{H}$ is I_{10}, and the column total for $\frac{H}{M}$ is I_{01}.

In testing the one-sided hypothesis with H_{11}: H is a better incendiary than M or that temperature has no affect upon the results or both, one would use Ury's C. First check to see if column 6 > column 5. It is; therefore, C is calculated and found to be

$$C = \left[\frac{12\,(11 - 35)^2}{(29 + 4)(11 + 35)} \right]^{\frac{1}{2}}$$

$$= 2.1338$$

Comparing this to the standard normal distribution, the observed significance level, $\hat{\alpha}$, is found to be 0.017. This was determined to be both statistically and practically significant, so H_{00} was rejected at the 98.3 percent confidence level.

APPENDIX A

IDENTITY: DEVIATION OF OBSERVATIONS FROM THE MEAN

The identity for deviation of observations from the mean and the identity cross product deviation from the means will be developed in this appendix. They are used in the development of the test statistic U_2.

Deviation of observations from the mean:

$$(y_{ij} - \bar{y}_{..}) \equiv (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}) \quad .$$

Squaring both sides and summing over i and j, one obtains

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 \equiv \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ij} (y_{ij} - \bar{y}_{i.})^2$$

$$+ 2 \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.}) \quad .$$

Working only with the cross product term, we have

$$2 \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.}) = 2 \sum_{i} (\bar{y}_{i.} - \bar{y}_{..})[\sum_{j} (y_{ij} - \bar{y}_{i.})]$$

$$= 0 \quad .$$

Therefore, the cross product term sums to zero and the identity may be expressed as

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 \equiv \sum_{i} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i} \left[\sum_{j} (y_{ij} - \bar{y}_{i.})^2\right]$$

$$\equiv T_{yy} + E_{yy}^{(\cdot)}$$

Cross product deviation of observations from the mean

$$(z_{ij} - \bar{z}_{..})(y_{ij} - \bar{y}_{..}) \equiv [(z_{ij} - \bar{z}_{i.}) + (\bar{z}_{i.} - \bar{z}_{..})] [(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})]$$

$$\equiv (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.}) + (z_{ij} - \bar{z}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) + (\bar{z}_{i.} - \bar{z}_{..})(y_{ij} - \bar{y}_{i.}) + (\bar{z}_{i.} - \bar{z}_{..})(\bar{y}_{i.} - \bar{y}_{..}) \quad .$$

First sum over j then over i.

$$\sum_{ij} (z_{ij} - \bar{z}_{..})(y_{ij} - \bar{y}_{..}) \equiv \sum_{i} \left[\sum_{j} (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..}) \sum_{j} (z_{ij} - \bar{z}_{i.}) + (\bar{z}_{i.} - \bar{z}_{..}) \sum_{j} (y_{ij} - \bar{y}_{i.}) + \sum_{j} (\bar{z}_{i.} - \bar{z}_{..})(\bar{y}_{i.} - \bar{y}_{..})\right]$$

$$\equiv \sum_{ij} (\bar{z}_{i.} - \bar{z}_{..})(\bar{y}_{i.} - \bar{y}_{..})$$

$$+ \sum_{ij} (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})$$

$$\sum_{ij} (z_{ij} - \bar{z}_{..})(y_{ij} - \bar{y}_{..}) \equiv T_{zy} + E_{zy}^{(\cdot)} \quad .$$

APPENDIX B

VARIANCE OF THE ADJUSTED TREATMENT MEAN

The variance of the estimated adjusted treatment mean is developed in this appendix. It is used in the discussion in Section V, "Decision Rule."

$$V(\hat{\zeta}_{i.}) = V[\bar{y}_{i.} - \hat{\beta}(\bar{z}_{i.} - \bar{z}_{..})]$$

$$= V(\bar{y}_{i.}) - (\bar{z}_{i.} - \bar{z}_{..})^2 V(\hat{\beta}) - 2 \operatorname{cov}[\bar{y}_{i.}, \hat{\beta}(\bar{z}_{i.} - \bar{z}_{..})]$$

Consider the above equation term by term:

$$\operatorname{cov}[\bar{y}_{i.}, \hat{\beta}(\bar{z}_{i.} - \bar{z}_{..})] = (\bar{z}_{i.} - \bar{z}_{..}) \operatorname{cov}(\bar{y}_{i.}, \beta)$$

But

$$\hat{\beta} = \sum_{ij} (z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.}) / \sum_{ij} (z_{ij} - \bar{z}_{i.})^2 \quad .$$

Let

$$K = (\bar{z}_{i.} - \bar{z}_{..}) \sum_{ij} (z_{ij} - \bar{z}_{i.}) / \sum_{ij} (z_{ij} - \bar{z}_{i.})^2 \quad ,$$

So we have

$$\operatorname{cov}[\bar{y}_{i.}, \hat{\beta}(\bar{z}_{i.} - \bar{z}_{..})] = K \operatorname{cov}[\bar{y}_{i.}, (y_{ij} - \bar{y}_{i.})]$$

$$= K[\operatorname{cov}(\bar{y}_{i.}, y_{ij}) - \operatorname{cov}(\bar{y}_{i.}, \bar{y}_{i.})]$$

$$= K[\operatorname{cov}(\frac{1}{n} \sum_{ij} y_{ij}, y_{ij}) - V(\bar{y}_{i.})]$$

$$\text{cov}\,[\bar{y}_{i.}, \hat{\beta}\,(\bar{z}_{i.} - \bar{z}_{..})] = K\,[\frac{1}{n}\,\sigma_{y\cdot z}^2 - \frac{1}{n}\,\sigma_{y\cdot z}^2]$$

$$= 0\ .$$

Now consider the term:

$$V\,(\hat{\beta}) = V\,[\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})(y_{ij} - \bar{y}_{i.})/\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})^2]$$

$$= [1/\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})^2]^2\,[V\{\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})(y_{ij})\}$$

$$- V\{\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})\,\bar{y}_{i.}\}]\ .$$

$\bar{y}_{i.}$ is constant with respect to j, and $\Sigma_{j}\,(z_{ij} - \bar{z}_{i.}) = 0$;

therefore $V\left[\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})\,\bar{y}_{i.}\right] = 0$.

This then leaves

$$V\,(\hat{\beta}) = [1/\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})^2]^2\,[\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})]^2\,\sigma_{y\cdot z}^2\ .$$

The term $[\Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})]^2$ in the numerator will divide out with one of the terms in the denominator if they are corrected to the proper i^{th} treatment. Continuing, we have

$$V\,(\hat{\beta}) = \sigma_{y\cdot z}^2 / \Sigma_{ij}\,(z_{ij} - \bar{z}_{i.})^2$$

$$= \sigma_{y\cdot z}^2 / E_{zz}\ .$$

The variance of the adjusted treatment means now becomes

$$V(\hat{\zeta}_{i.}) = \sigma_{y.z}^2/n_i + (\bar{z}_{i.} - \bar{z}_{..})^2 \sigma_{y.z}^2/E_{zz}$$

$$= \sigma_{y.z}^2 \left[\frac{1}{n_i} + (\bar{z}_{i.} - \bar{z}_{..})^2/E_{zz}\right]$$

where $\sigma_{y.z}^2$ is estimated by $s_{y.z}^2$.

APPENDIX C

IDENTITY: SUM OF SQUARES OF ALL DIFFERENCES

The identity, the sum of squares of all differences which is identical to 2n times the sum of squares about the mean, will be developed in this appendix.

$$\sum_{\substack{ik \\ i \neq k}} (z_i - z_k)^2 = \sum_i \sum_k (z_i - z_k)^2$$

$$= \sum_i \sum_k (z_i^2 - 2z_i z_k + z_k^2)$$

$$= \sum_i \sum_k z_i^2 - 2\sum_i z_i \sum_k z_k + \sum_i \sum_k z_k^2$$

$$= \sum_i n z_i^2 - 2n \bar{z} n \bar{z} + \sum_k n z_k^2$$

$$= 2n \left(\sum_i z_i^2 - n \bar{z}^2\right)$$

$$= 2n \sum_i (z_i - \bar{z})^2$$

(The reverse of this page is blank)

REFERENCES

1. Bancroft, T.A., Topics Intermediate Statistical Methods, Iowa State University Press, 1968, pp. 75-83.

2. Bross, Irwin D.J., "Taking a Covariable into Account," Journal of the American Statistical Association, September 1964, pp. 725-736.

3. Cochran, W.G., "Analysis of Covariance: Its Nature and Uses," Biometrics, September 1957, pp. 261-280.

4. Coons, I., "The Analysis of Covariance as a Missing Plot Technique," Biometrics, September 1957.

5. Draper, N.R., and Smith, H., Applied Regression Analysis, Wiley and Sons, Inc., 1966., pp. 1-43.

6. Hazel, L.N., "The Covariance Analysis of Multiple Classification Tables With Unequal Subclass Numbers," Biometrics, Volume 2, No. 2, April 1946, pp. 21-25.

7. Katti, S.K., "Multiple Covariance Analysis," Biometrics, Volume 21 No. 4, December 1965, pp. 957-974.

8. Morrison, D.F., Multivariate Statistical Methods, McGraw-Hill, 1967, pp. 180-182.

9. Puri, Madan L. and Sen, Pranab K., "Analysis of Covariance Based on General Rank Scores," The Annals of Mathematical Statistics, 1969, Volume 40, No. 2, pp. 610-618.

10. Quade, Dana, "Rank Analysis of Covariance," Journal of the American Statistical Association, Volume 62, September - December 1967, pp. 1187-1200.

11. Searle, S.R., Linear Models, Wiley and Sons, Inc., 1971, pp. 340-361.

12. Snedecor, G.W. and Cochran, W.G., Statistical Methods, 6th Ed., 1967, pp. 419-446.

13. Steel, R.G.D. and Torrie, J.H., Principles and Procedures of Statistics, McGraw-Hill, 1960, pp. 305-330.

14. Ury, Hans K., "A Note on Taking a Covariable Into Account," Journal of the American Statistical Association, Volume 61, March - June 1966, pp. 490-495.

(The reverse of this page is blank)

INITIAL DISTRIBUTION

HQ USAF/SAMI	1
USAFE/DOQ	1
PACAF/DOOFQ	1
TAC/DRA	1
ASD/ENFEA	1
AUL/LSE- 71-249	1
SAC/NRI (STINFO LIB)	1
NWC/CODE 318	1
NWC/CODE 317	1
OO-ALC/MMWMP	2
AFIS/INTA	1
DDC	2
AFATL/DLODL	2
AFATL/DL	1
AFATL/DLY	1
ADTC/XRS	1
AFATL/DLYV	20
AFATL/DLYW	10
USA ENG WatWay Ex Sta/VMS	1
BAL RESRCH LAB/AMXBR-VL	1
AMSAA/DRXSY-J	2
USAMSAA/DRXSY-S	1
ARRADCOM/DRDAR-LCU-TM	1
AFOSR/NM	1
ARMY RESEARCH OFFICE/NC	1
OKLAHOMA ST UNIV/Dept of Stat	20
TAC/INAT	1
USA TRADOC SYS ANN ACT	1
ASD/XRP	1
COMIPAC/I-232	1

(The reverse of this page is blank)