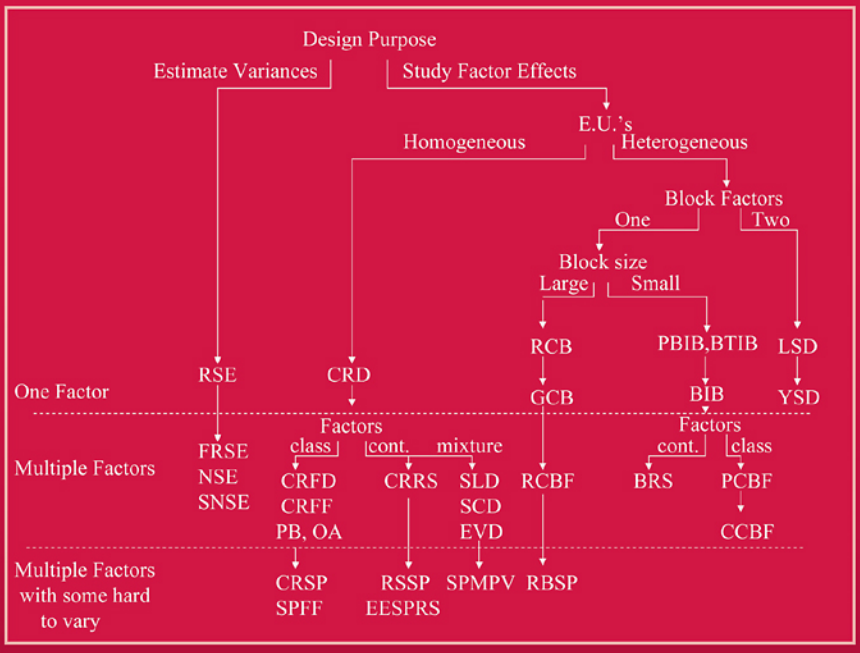
Texts in Statistical Science

Design and Analysis of Experiments with SAS



John Lawson

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Design and Analysis of Experiments with SAS

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
P. J. Smith

The Analysis of Time Series — An Introduction, Sixth Edition
C. Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West and J. Harrison

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Applied Statistics — Handbook of GENSTAT Analysis
E.J. Snell and H. Simpson

Applied Statistics — Principles and Examples
D.R. Cox and E.J. Snell

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Bayesian Data Analysis, Second Edition
A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Beyond ANOVA — Basics of Applied Statistics
R.G. Miller, Jr.

Computer-Aided Multivariate Analysis, Fourth Edition
A.A. Afifi and V.A. Clark

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Design and Analysis of Experiment with SAS
J. Lawson

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis, Second Edition
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

Extending the Linear Model with R — Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Generalized Additive Models: An Introduction with R
S. Wood

Interpreting Data — A First Course in Statistics
A.J.B. Anderson

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Introduction to Probability with R
K. Baclawski

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Linear Models with R
J.J. Faraway

Logistic Regression Models
J.M. Hilbe

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Pólya Urn Models
H. Mahmoud

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
P.S.R.S. Rao

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Stochastic Processes: An Introduction, Second Edition

P.W. Jones and P. Smith

Survival Analysis Using S — Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Texts in Statistical Science

Design and Analysis of Experiments with SAS

John Lawson

Brigham Young University

Provo, Utah, U.S.A.

CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the

Taylor & Francis Group an **informa** business

A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-6060-7 (Hardback)

Library of Congress Cataloging-in-Publication Data

Lawson, John, 1947-
Design and analysis of experiments with SAS / John Lawson.
p. cm. -- (Chapman & Hall/CRC texts in statistical science series)
Includes bibliographical references and index.
ISBN 978-1-4200-6060-7 (hard back : alk. paper)
1. Experimental design. 2. SAS (Computer file) I. Title. II. Series.

QA279.L36 2010
519.5'7--dc22 2009052546

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

After studying experimental design a researcher or statistician should be able to: (1) choose an experimental design that is appropriate for the research problem at hand; (2) construct the design (including performing proper randomization and determining the required number of replicates); (3) execute the plan to collect the data (or advise a colleague to do it); (4) determine the model appropriate for the data; (5) fit the model to the data; and (6) interpret the data and present the results in a meaningful way to answer the research question. The purpose of this book is to focus on connecting the objectives of research to the type of experimental design required, describing the actual process of creating the design and collecting the data, showing how to perform the proper analysis of the data, and illustrating the interpretation of results. Exposition on the mechanics of computation is minimized by relying on a statistical software package.

With the availability of modern statistical computing packages, the analysis of data has become much easier and is well covered in statistical methods books. There is no longer a need to show all the computational formulas that were necessary before the advent of modern computing, in a book on the design and analysis of experiments. However, there is a need for careful explanation of how to get the proper analysis from a computer package. The default analysis performed by most statistical software assumes the data have come from a completely randomized design. In practice, this is often a false assumption. This book emphasizes the connection between the experimental units, and the way treatments are randomized to experimental units, and the proper error term for an analysis of the data.

The SAS system for statistical analysis is used throughout the book to illustrate both construction of experimental designs and analysis of data. This software was chosen to be illustrated because it has extensive capabilities in both creating designs and analyzing data, the command language has been stable for over thirty years, and it is widely used in industry. SAS version 9.2 has been used in the text, and all the SAS code for examples in the book is available at http://lawson.mooo.com. The ods graphics used in the book require version 9.2 or later. In earlier versions of SAS similar graphs can be created with the legacy SAS/GRAPH routines and the code to do this is also available on the Web site. Examples of SAS data step programming and IML are presented, and procedures from SAS Stat, SAS QC, and SAS OR are illustrated.

With fewer pages devoted to computational formulas, I have attempted to

spend more time discussing the following: (1) how the objectives of a research project lead to the choice of an appropriate design, (2) practical aspects of creating a design, or list of experiments to be performed, (3) practical aspects of performing experiments, and (4) interpretation of the results of a computer analysis of the data. Items (1)-(3) can best be taught by giving many examples of experiments and exercises that actually require readers to perform their own experiments.

This book attempts to give uniform coverage to experimental designs and design concepts that are most commonly used in practice, rather than focusing on specialized areas. The selection of topics is based on my own experience working in the pharmaceutical industry, and in research and development (R&D) and manufacturing in agricultural and industrial chemicals, and machinery industries. At the end of each chapter a diagram is presented to help identify where the various designs should be used. Examples in the book come from a variety of application areas. Emphasis is placed on how the sample size, the assignment of experimental units to combinations of treatment factor levels (error control), and the selection of treatment factor combinations (treatment design) will affect the resulting variance and bias of estimates and the validity of conclusions.

Intended audience This book was written for first-year graduate students in statistics or advanced undergraduates who intend to work in an area where they will use experimental designs. To be fully understood, a student using this book should have had previous courses in calculus, introductory statistics, basic statistical theory, applied linear models such as Kutner *et al.* (2004) and Faraway (2004), and some familiarity with SAS. Matrix notation for analysis of linear models is used throughout the book, and students should be familiar with matrix operations at least to the degree illustrated in chapter 5 of Kutner *et al.* (2004).

However, for students from applied sciences or engineering who do not have all these prerequisites, there is still much to be gained from this book. There are many examples of SAS code to create and analyze experiments as well plentiful examples of (1) diagnosing the experimental environment to choose the correct design, and (2) interpreting and presenting results of analysis. One with a basic understanding of SAS should be able to follow these examples and modify them to complete the exercises in the book and solve problems in their own research, without needing to understand the detailed theoretical justification for each procedure.

For instructors This book can be used for a one-semester or two-quarter course in experimental design. There is too much material for a one-semester course, unless the students have had all the prerequisites mentioned above. The first four chapters in the book cover the classical ideas in experimental design, and should be covered in any course for students without a prior background in designed experiments. Later chapters start with basics, but proceed to the latest research published on particular topics, and they include code to implement all of these ideas. An instructor can pick and choose from

these remaining topics, although if there is time to cover the whole book, I would recommend presenting the topics in order.

Some instructors who do not intend to cover the entire book might consider covering factorial experiments in Chapter 3, fractional factorials in Chapter 6, and response surface methods in Chapter 9, following the pattern established by the DuPont Strategies of Experimentation Short Courses that were developed in the 1970s. I chose the ordering of chapters in the book so that variance component designs in Chapter 5, would be presented before describing split plot experiments that are so commonplace in practice. I did this because I feel it is important to understand random factors before studying designs where there is more than one error term.

Acknowledgments This book is the culmination of many years of thought prompted by consulting and teaching. I would be remiss if I did not thank Melvin Carter, my advisor at Brigham Young University (BYU) who introduced me to the computer analysis of experimental data over forty years ago, and whose enthusiasm about the subject of designed experiments inspired my lifelong interest in this area. I would also like to thank John Erjavec, my boss and mentor at FMC Corp., for introducing me to the ideas of Box, Hunter and Hunter long before their original book *Statistics for Experimenters* was published. I also thank the many consulting clients over the years who have challenged me with interesting problems, and the many students who have asked me to explain things more clearly. Special thanks to my former students Willis Jensen at Gore and Michael Joner at Procter and Gamble for their careful review and comments on my manuscript, and Seyed Mottaghinejad for providing a solutions manual and finding many typos and unclear points in the text. Finally, I thank my wife Francesca for her never-ending support and encouragement during the writing of this book.

John Lawson
Department of Statistics
Brigham Young University

CHAPTER 1

Introduction

1.1 Statistics and Data Collection

Statistics is defined as the science of collecting, analyzing and drawing conclusions from data. Data is usually collected through sampling surveys, observational studies, or experiments.

Sampling surveys are normally used when the purpose of data collection is to estimate some property of a finite population without conducting a complete census of every item in the population. For example, if there were interest in finding the proportion of registered voters in a particular precinct that favor a proposal, this proportion could be estimated by polling a random sample of voters rather than questioning every registered voter in the precinct.

Observational studies and experiments, on the other hand, are normally used to determine the relationship between two or more measured quantities in a conceptual population. A conceptual population, unlike a finite population, may only exist in our minds. For example, if there were interest in the relationship between future greenhouse gas emissions and future average global temperature, the population, unlike registered voters in a precinct, cannot be sampled from because it does not yet exist.

To paraphrase the late W. Edwards Deming, the value of statistical methods is to make predictions which can form the basis for action. In order to make accurate future predictions of what will happen when the environment is controlled, cause and effect relationships must be assumed. For example, to predict future average global temperature given that greenhouse gas emissions will be controlled at a certain level, we must assume that the relationship between greenhouse gas emissions and global temperature is cause and effect. Herein lies the main difference in observational studies and experiments. In an observational study, data is observed in its natural environment, but in an experiment the environment is controlled. In observational studies it cannot be proven that the relationships detected are cause and effect. Correlations may be found between two observed variables because they are both affected by changes in a third variable that was not observed or recorded, and any future predictions made based on the relationships found in an observational study must assume the same interrelationships among variables that existed in the past will exist in the future. In an experiment, on the other hand, some variables are purposely changed while others are held constant. In that way the effect that is caused by the change in the purposely varied variable can

be directly observed, and predictions can be made about the result of future changes to the purposely varied variable.

1.2 Beginnings of Statistically Planned Experiments

There are many purposes for experimentation. Some examples include: determining the cause for variation in measured responses observed in the past; finding conditions that give rise to the maximum or minimum response; comparing the response between different settings of controllable variables; and obtaining a mathematical model to predict future response values.

Presently planned experiments are used in many different fields of application such as engineering design, quality improvement, industrial research and manufacturing, basic research in physical and biological science, research in social sciences, psychology, business management and marketing research, and many more. However, the roots of modern experimental design methods stem from R. A. Fisher's work in agricultural experimentation at the Rothamsted Experimental Station near Harpenden, England.

Fisher was a gifted mathematician whose first paper as an undergraduate at Cambridge University introduced the theory of likelihood. He was later offered a position at University College, but turned it down to join the staff at Rothamsted in 1919. There, inspired by daily contact with agricultural research, he not only contributed to experimental studies in areas such as crop yields, field trials, and genetics, but also developed theoretical statistics at an astonishing rate. He also came up with the ideas for planning and analysis of experiments that have been used as the basis for valid inference and prediction in various fields of application to this day. Fisher (1926) first published his ideas on planning experiments in his paper "The arrangement of field experiments"; nine years later he published the first edition of his book *The Design of Experiments*, Fisher (1935).

The challenges that Fisher faced were the large amount of variation in agricultural and biological experiments that often confused the results, and the fact that experiments were time consuming and costly to carry out. This motivated him to find experimental techniques that could:

- eliminate as much of the natural variation as possible
- prevent unremoved variation from confusing or biasing the effects being tested
- detect cause and effect with the minimal amount of experimental effort necessary.

1.3 Definitions and Preliminaries

Before initiating an extended discussion of experimental designs and the planning of experiments, I will begin by defining the terms that will be used frequently.

- *Experiment* (also called a *Run*) is an action where the experimenter changes at least one of the variables being studied and then observes the effect of his or her actions(s). Note the passive collection of observational data is not experimentation.
- *Experimental Unit* is the item under study upon which something is changed. This could be raw materials, human subjects, or just a point in time.
- *Sub-sample, sub-unit or observational unit* When the experimental unit is split, after the action has been taken upon it, this is called a sub-sample or sub-unit. Sometimes it is only possible to measure a characteristic separately for each sub-unit; for that reason they are often called observational units. Measurements on sub-samples, or sub-units of the same experimental unit are usually correlated and should be averaged before analysis of data rather than being treated as independent outcomes. When sub-units can be considered independent and there is interest in determining the variance in sub-sample measurements, while not confusing the F-tests on the treatment factors, the mixed model described in Section 5.8 should be used instead of simply averaging the sub-samples.
- *Independent Variable* (*Factor* or *Treatment Factor*) is one of the variables under study that is being controlled at or near some target value, or *level*, during any given experiment. The level is being changed in some systematic way from run to run in order to determine what effect it has on the response(s).
- *Background Variable* (also called *Lurking variable*) is a variable that the experimenter is unaware of or cannot control, and which could have an effect on the outcome of the experiment. In a well-planned experimental design, the effect of these lurking variables should balance out so as to not alter the conclusion of a study.
- *Dependent Variable* (or the *Response* denoted by Y) is the characteristic of the experimental unit that is measured after each experiment or run. The magnitude of the response depends upon the settings of the independent variables or factors and lurking variables.
- *Effect* is the change in the response that is caused by a change in a factor or independent variable. After the runs in an experimental design are conducted, the effect can be estimated by calculating it from the observed response data. This estimate is called the *calculated effect*. Before the experiments are ever conducted, the researcher may know how large the effect should be to have practical importance. This is called a *practical effect* or the *size of a practical effect*.
- *Replicate* runs are two or more experiments conducted with the same settings of the factors or independent variables, but using different experimental units. The measured dependent variable may differ in replicate runs due to changes in lurking variables and inherent differences in experimental units.

- *Duplicates* refer to duplicate measurements of the same experimental unit from one run or experiment. The measured dependent variable may vary among duplicates due to measurement error, but in the analysis of data these duplicate measurements should be averaged and not treated as separate responses.

- *Experimental Design* is a collection of experiments or runs that is planned in advance of the actual execution. The particular runs selected in an experimental design will depend upon the purpose of the design.

- *Confounded Factors* arise when each change an experimenter makes for one factor, between runs, is coupled with an identical change to another factor. In this situation it is impossible to determine which factor causes any observed changes in the response or dependent variable.

- *Biased Factor* results when an experimenter makes changes to an independent variable at the precise time when changes in background or lurking variables occur. When a factor is biased it is impossible to determine if the resulting changes to the response were caused by changes in the factor or by changes in other background or lurking variables.

- *Experimental Error* is the difference between the observed response for a particular experiment and the long run average of all experiments conducted at the same settings of the independent variables or factors. The fact that it is called "error" should not lead one to assume that it is a mistake or blunder. Experimental errors are not all equal to zero because background or lurking variables cause them to change from run to run. Experimental errors can be broadly classified into two types: bias error and random error. Bias error tends to remain constant or change in a consistent pattern over the runs in an experimental design, while random error changes from one experiment to another in an unpredictable manner and average to be zero. The variance of random experimental errors can be obtained by including replicate runs in an experimental design.

With these definitions in mind, the difference between observational studies and experiments can be explained more clearly. In an observational study, variables (both independent and dependent) are observed without any attempt to change or control the value of the independent factors. Therefore any observed changes in the response, or dependent variable, cannot necessarily be attributed to observed changes in the independent variables because background or lurking variables might be the cause. In an experiment, however, the independent variables are purposely varied and the runs are conducted in a way to balance out the effect of any background variables that change. In this way the average change in the response can be attributed to the changes made in the independent variables.

1.4 Purposes of Experimental Design

The use of experimental designs is a prescription for successful application of the scientific method. The scientific method consists of iterative application of the following steps: (1) observing of the state of nature, (2) conjecturing or hypothesizing the mechanism for what has been observed, then (3) collecting data, and (4) analyzing the data to confirm or reject the conjecture. Statistical experimental designs provide a plan for collecting data in a way that they can be analyzed statistically to corroborate the conjecture in question. When an experimental design is used, the conjecture must be stated clearly and a list of experiments proposed in advance to provide the data to test the hypothesis. This is an organized approach which helps to avoid false starts and incomplete answers to research questions.

Another advantage to using the experimental design approach is ability to avoid confounding factor effects. When the research hypothesis is not clearly stated and a plan is not constructed to investigate it, researchers tend toward a trial and error approach wherein many variables are simultaneously changed in an attempt to achieve some goal. When this is the approach, the goal may sometimes be achieved, but it cannot be repeated because it is not known what changes actually caused the improvement.

One of Fisher's early contributions to the planning of experiments was popularizing a technique called randomization which helps to avoid confusion or biases due to changes in background or lurking variables. As an example of what we mean by bias is "The Biggest Health Experiment Ever," Meier (1972), wherein a trial of a polio vaccine was tested on over 1.8 million children. An initial plan was proposed to offer vaccination to all children in the second grade in participating schools, and to follow the polio experience of first through third graders. The first and third grade group would serve as a "control" group. This plan was rejected, however, because doctors would have been aware that the vaccine was only offered to second graders. There are vagaries in the diagnosis of the majority of polio cases, and the polio symptoms of fever and weakness are common to many other illnesses. A doctor's diagnosis could be unduly influenced by his knowledge of whether or not a patient had been vaccinated. In this plan the factor purposely varied, vaccinated or not, was biased by the lurking variable of doctors' knowledge of the treatment.

When conducting physical experiments, the response will normally vary over replicate runs due solely to the fact that the experimental units are different. This is what we defined to be experimental error in the last section. One of the main purposes for experimental designs is to minimize the effect of experimental error. Aspects of designs that do this, such as randomization, replication and blocking, are called methods of *error control*. Statistical methods are used to judge the average effect of varying experimental factors against the possibility that they may be due totally to experimental error. Another purpose for experimental designs is to accentuate the factor effects

(or signal). Aspects of designs that do this, such as choice of the number and spacing of factor levels and factorial plans, are called methods of *treatment design*. How this is done will be explained in the following chapters.

1.5 Types of Experimental Designs

There are many types of experimental designs. The appropriate one to use depends upon the objectives of the experimentation. We can classify objectives into two main categories. The first category is to study the sources of variability, and the second is to establish cause and effect relationships. When variability is observed in a measured variable, one objective of experimentation might be to determine the cause of that variation. But before cause and effect relationships can be studied, a list of independent variables must be determined. By understanding the source of variability, researchers are often led to hypothesize what independent variables or factors to study. Thus experiments to study the source of variability are often a starting point for many research programs. The type of experimental design used to classify sources of variation will depend on the number of sources under study. These alternatives will be presented in Chapter 5.

The appropriate experimental design that should be used to study cause and effect relationships will depend on a number of things. Throughout the book the various designs are described in relation to the purpose for experimentation, the type and number of treatment factors, the degree of homogeneity of experimental units, the ease of randomization, and the ability to block experimental units into more homogeneous groups. After all the designs are presented, Chapter 13 describes how they can be used in sequential experimentation strategies where knowledge is increased through different stages of experimentation. Initial stages involve discovering what the important treatment factors are. Later, the effects of changing treatment factors are quantified, and in final stages, optimal operating conditions can be determined. Different types of experimental designs are appropriate for each of these phases.

Screening experiments are used when the researcher has little knowledge of the cause and effect relationships, and many potential independent variables are under study. This type of experimentation is usually conducted early in a research program to identify the important factors. This is a critical step, and if it is skipped, the later stages of many research programs run amuck because the important variables are not being controlled or recorded.

After identifying the most important factors in a screening stage, the researcher's next objective would be to choose between constrained optimization or unconstrained optimization (see Lawson, 2003). In constrained optimization there are usually six or fewer factors under study and the purpose is to quantify the effects of the factors, interaction or joint effects of factors, and to identify optimum conditions among the factor combinations actually tested.

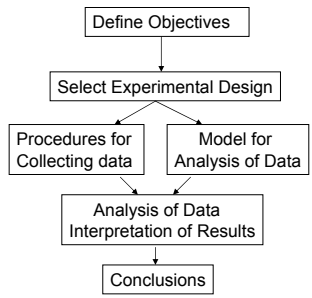
When only a few quantitative factors are under study and curvilinear relationships with the response are possible, it may be possible to identify

improved operating conditions by interpolating within the factor levels actually tested. If this is the goal, the objective of experimentation is called unconstrained optimization. With an unconstrained optimization objective, the researcher is normally trying to map the relationship between one or more responses and five or fewer quantitative factors.

Specific experimental design plans for each of the stages of experimentation will be presented as we progress through the book.

Figure 1.1 shows the relationship between the objectives of experimentation, the design of the experiment, and the conclusions that can be drawn. The objective of a research program dictates which type of experimental design should be utilized. The experimental design plan in turn specifies how the data should be collected and what mathematical model should be fit in order to analyze and interpret the data. Finally the type of data and the mathematical model will determine what possible conclusions can be drawn from the experiment. These steps are inseparable and dependent upon each other. Many mistakes are made in research by trying to dissever these steps. An appropriate analysis of data cannot be completed without knowledge of what experimental design was used and how the data was collected, and conclusions are not reliable if they are not justified by the proper modeling and analysis of the data.

Figure 1.1 *Objectives, Design and Conclusions from Experimentation*



1.6 Planning Experiments

An effective experimental design plan should include the following items: (1) a clear description of the objectives, (2)an appropriate design plan that guarantees unconfounded factor effects and factor effects that are free of bias, (3) a provision for collecting data that will allow estimation of the variance of the experimental error, and (4) a stipulation to collect enough data to satisfy the objectives. Bisgaard (1999) recommends a formal proposal to ensure that a plan includes all of these elements. The proposal should include a checklist for planning the experiments. Below is a checklist that is similar to Bisgaard's.

Examples of some of the steps from this checklist will be illustrated in discussing a simple experiment in the next section.

1. *Define Objectives.* Define the objectives of the study. First, this statement should answer the question of why is the experiment to be performed. Second, determine if the experiment is conducted to classify sources of variability or if its purpose is to study cause and effect relationships. If it is the latter, determine if it is a screening or optimization experiment. For studies of cause and effect relationships, decide how large an effect should be in order to be meaningful to detect.
2. *Identify Experimental Units.* Declare the item upon which something will be changed. Is it an animal or human subject, raw material for some processing operation, or simply the conditions that exist at a point in time or *trial*? Identifying the experimental units will help in understanding the experimental error and variance of experimental error.
3. *Define a Meaningful and Measurable Response or Dependent Variable.* Define what characteristic of the experimental units can be measured and recorded after each run. This characteristic should best represent the expected differences to be caused by changes in the factors.
4. *List the Independent and Lurking Variables.* Declare which independent variables you wish to study. Ishikawa Cause and Effect Diagrams (see SAS Institute, 2004b) are often useful at this step to help organize variables thought to affect the experimental outcome. Be sure that the independent variables chosen to study can be controlled during a single run, and varied from run to run. If there is interest in a variable, but it cannot be controlled or varied, it cannot be included as a factor. Variables that are hypothesized to affect the response, but cannot be controlled, are lurking variables. The proper experimental design plan should prevent uncontrollable changes in these variables from biasing factor effects under study.
5. *Run Pilot Tests.* Make some pilot tests to be sure you can control and vary the factors that have been selected, that the response can be measured, and that the replicate measurements of the same or similar experimental units are consistent. Inability to measure the response accurately or to control the factor levels are the main reasons that experiments fail to produce desired results. If the pilot tests fail, go back to steps 2, 3 and 4. If these tests are successful, measurements of the response for a few replicate tests with the same levels of the factors under study will produce data that can be used to get a preliminary estimate of the variance of experimental error.
6. *Make a Flow Diagram of the Experimental Procedure for Each Run* This will make sure the procedure to be followed is understood and will be standardized for all runs in the design.
7. *Choose the Experimental Design.* Choose an experimental design that is suited for the objectives of your particular experiment. This will include a description of what factor levels will be studied and will determine how the

experimental units are to be assigned to the factor levels or combination of factor levels if there are more than one factor. One of the plans described in this book will almost always be appropriate. The choice of the experimental design will also determine what model should be used for analysis of the data.

8. *Determine the Number of Replicates Required* Based on the expected variance of the experimental error and the size of a practical difference, the number of replicate runs that will give the researcher a high probability of detecting an effect of practical importance.
9. *Randomize the Experimental Conditions to Experimental Units.* According to the particular experimental design being used, there is a proscribed method of randomly assigning experimental conditions to experimental units. The way this is done affects the way the data should be analyzed, and it is important to describe and record what is done. The best way to do this is to provide a data collection worksheet arranged in the random order in which the experiments are to be collected. For more complicated experimental designs Bisgaard (1999) recommends one sheet of paper describing the conditions of each run with blanks for entering the response data and recording observations about the run. All these sheets should then be stapled together in booklet form in the order they are to be performed.
10. *Describe a Method for Data Analysis.* This should be an outline of the steps of the analysis. An actual analysis of simulated data is often useful to verify that the proposed outline will work.
11. *Timetable and Budget for Resources Needed to Complete the Experiments.* Experimentation takes time and having a schedule to adhere to will improve the chances of completing the research on time. Bisgaard (1999) recommends a Gantt Chart (see SAS Institute, 2004a) which is a simple graphical display showing the steps of the process as well as calendar times. A budget should be outlined for expenses and resources that will be required.

1.7 Performing the Experiments

In experimentation, careful planning and execution of the plan are the most important steps. As we know from Murphy's Law, if anything can go wrong it will, and analysis of data can never compensate for botched experiments. To illustrate the potential problems that can occur, consider a simple experimenter conducted by an amateur gardener described by Box *et al.* (1978). The purpose was to determine whether a change in the fertilizer mixture would result in a change in the yield of his tomato plants. Eleven tomato plants were planted in a single row, and the fertilizer type (A or B) was varied. The experimental unit in this experiment is the tomato plant plus the soil it is planted in, and the treatment factor is the type of fertilizer applied. Easterling (2004) discusses some of the nuances that should be considered when planning and

carrying out such a simple experiment. It is instructive to think about these in context with the checklist presented in the last section.

When defining the objectives for this experiment, the experimenter needs to think ahead to the possible implications of conclusions that he can draw. In this case, the possible conclusions are (1) deciding that the fertilizer has no effect on the yield of tomatoes, or (2) concluding that one fertilizer produces a greater yield. If the home gardener finds no difference in yield, he can choose to use the less expensive fertilizer. If he finds a difference, he will have to decide if the increase in yield offsets any increase in cost of the better fertilizer. This can help him determine how large a difference in yield he should look for and the number of tomato plants he should include in his study. The answer to this question, which is crucial in planning the experiment, would probably be much different for a commercial grower than for a backyard enthusiast.

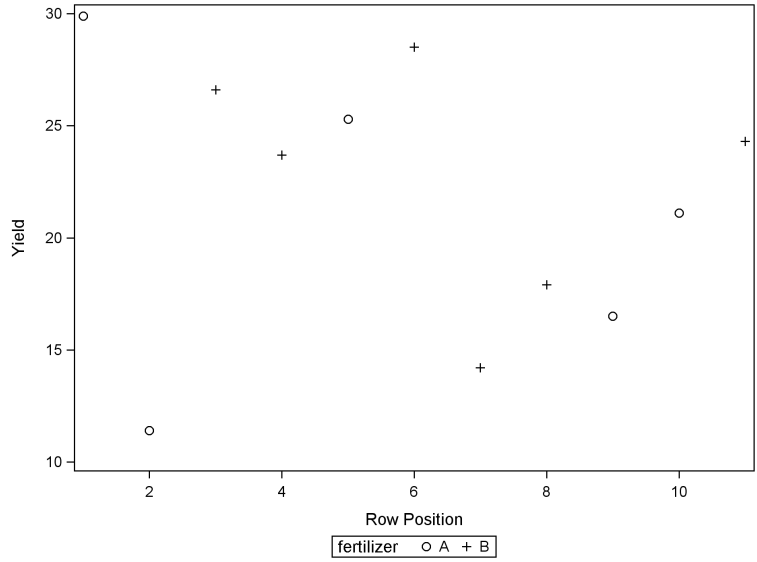
The experimental units for this experiment were defined in the paragraph above, but in identifying them, the experimenter should consider the similarity or homogeneity of plants and how far apart he is going to place the tomato plants in the ground. Will it be far enough that the fertilizer applied to one plant does not bleed over and affect its neighbors?

Defining a meaningful response that can be measured may be tricky in this experiment. Not all the tomatoes on a single plant ripen at the same time. Thus, to measure the yield in terms of weight of tomatoes, the checklist and flow diagram describing how an experiment is conducted must be very precise. Is it the weight of all tomatoes on the plant at a certain date, or the cumulative weight of tomatoes picked over time as they ripen? Precision in the definition of the response and consistency in adherence to the definition when making the measurements are crucial.

There are many possible lurking variables to consider in this experiment. Any differences in watering, weeding, insect treatment, the method and timing of fertilizer application, and the amount of fertilizer applied may certainly affect the yield; hence the experimenter must pay careful attention to these variables to prevent bias. Easterling (2004) also pointed out that the row position seems to have affected the yield as well (as can be seen in Figure 1.2). The randomization of fertilizers to plants and row positions should equalize these differences for the two fertilizers. This was one of the things that Box *et al.* (1978) illustrated with this example. If a convenient method of applying the fertilizers (such as A at the beginning of the row followed by B) had been used in place of random assignment, the row position effect could have been mistaken for a treatment effect. Had this row position effect been known before the experiment was planned, the adjacent pairs of plots could have been grouped together in pairs, and one fertilizer assigned at random to one plot-plant in each pair to prevent bias from the row position effect. This technique is called blocking and will be discussed in detail in Chapter 4.

Easterling (2004) also raised the question: why were only eleven plants used in the study (five fertilized with fertilizer A and six with fertilizer B)? Normally flats of tomato plants purchased from a nursery come in flats of

Figure 1.2 *Plot of Yield by Row Position - Tomato Experiment*



twelve. Was one plant removed from the study because it appeared unhealthy or got damaged in handling? The yield for the plant in the second row position (see Figure 1.2) of the eleven plants used was considerably lower than the others planted in neighboring row positions with the same fertilizer. Was this plant unhealthy or damaged as well?

Any problems that arise during the conduct of experiments should be carefully observed, noted, and recorded as comments on the data collection form described in step 9 of the checklist. Perhaps if this had been done for the tomato experiment, the low yield at row position two could be explained.

This discussion of a very simple experiment helps to emphasize the importance of carefully considering each step of the checklist presented in Section 11.6, and the importance of strict adherence to a flowchart for conducting the experiments, described in step 6 of that checklist. Failing to consider each point of the checklist, and inconsistency in conducting experiments and recording results may lead to the demise of an otherwise useful research project.

1.8 Use of SAS Software

Fisher's original book on Experimental Designs clearly laid the logical principles for experimentation, but users of experimental designs needed to have more detailed descriptions of the most useful designs along with accompanying plans. Consulting statisticians needed to have a systematic explanation of the relation between experimental designs and the statistical theory of least squares and linear hypotheses, and to have an enumeration of designs and

descriptions of experimental conditions where each design was most appropriate.

These needs were satisfied by Cochran and Cox (1950)'s and Kempthorne (1952)'s books. However, Cochran and Cox and Kempthorne's books were published before the age of computers and they both emphasize extensive tables of designs, abundant formulas and numerical examples describing methods of manual analysis of experimental data and mathematical techniques for constructing certain types of designs. Since the publication of these books, use of experimental designs has gone far beyond agricultural research where it was initially employed, and a plethora of new books have been written on the subject. Even though computers and software (to both design and analyze data from experiments) are widely available, a high proportion of the more recent books on experimental design still follow the traditional pattern established by Cochran and Cox and Kempthorne by presenting extensive tables of designs and formulas for hand calculations and methods for constructing designs.

One of the objectives of this book is to break from the tradition and present computer code and output in place of voluminous formulas and tables. This will leave more room in the text to discuss the appropriateness of various design plans and ways to interpret and present results from experiments. The particular computer software illustrated in this book is SAS, which was originally developed in 1972. Its syntax is relatively stable and is widely used. SAS has probably the widest variety of general procedures for design of experiments and analysis of experimental data, including `proc plan`, `proc factex`, `proc optex`, and the menu driven SAS ADX for the design of experiments, and `proc glm`, `proc varcomp`, `proc mixed` and many other procedures for the analysis of experimental data. These procedures are available in the SAS/Stat and SAS/QC software.

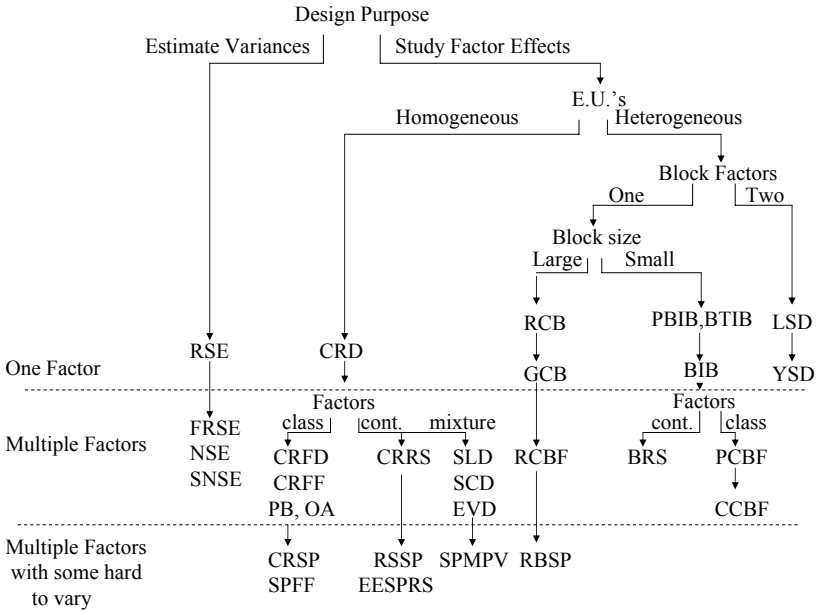
1.9 Review of Important Concepts

This chapter describes the purpose for experimental designs. In order to determine if cause and effect relationships exist, an experimental design must be conducted. In an experimental design, the factors under study are purposely varied and the result is observed. This is different from observational studies or sampling surveys where data is collected with no attempt to control the environment. In order to predict what will happen in the future, when the environment is controlled, you must rely on cause and effect relationships. Relationships obtained from observational studies or sampling surveys are not reliable for predicting future results when the environment is to be controlled.

Experimental designs were first developed in agricultural research, but are now used in all situations where the scientific method is applied. The basic definitions and terminology used in experimental design are given in this chapter along with a checklist for planning experiments. In practice there are many different types of experimental designs that can be used. Which design

is used in a particular situation depends upon the research objectives and the experimental units. Figure 1.3 is a diagram that illustrates when the different experimental designs described in this book should be used. As different experimental designs are presented in chapters to follow, reference will be made back to this figure to describe when the designs should be used.

Figure 1.3 *Design Selection Roadmap*



1.10 Exercises

1. A series of runs were performed to determine how the wash water temperature and the detergent concentration affect the bacterial count on the palms of subjects in a hand washing experiment.

   (a) Identify the experimental unit.

   (b) Identify the factors.

   (c) Identify the response.

2. Explain the difference between an experimental unit and a sub-sample or sub-unit in relation to the experiments described in 1.

3. Explain the difference between a sub-sample and a duplicate in relation to the experiment described in 1.

4. Describe a situation within your realm of experience (your work, your hobby, or school) where you might like to predict the result of some future action. Explain how an experimental design, rather than an observational study might enhance your ability to make this prediction.

5. Kerry and Bland (1998) describe the analysis of cluster randomized studies where a group of subjects are randomized to the same treatment. For example, when women in some randomly selected districts are offered breast cancer screening while women in other districts are not offered the screening, or when some general practitioners are randomly assigned to receive one or more courses of special training and the others are not offered the training. The response (some characteristic of the patients) in the cluster trials must be measured on each patient rather than the group as a whole. What is the experimental unit in this type of study? How would you describe the individual measurements on patients?

CHAPTER 2

Completely Randomized Designs with One Factor

2.1 Introduction

In a completely randomized design, abbreviated as CRD, with one treatment factor, n experimental units are divided randomly into t groups. Each group is then subject to one of the unique levels or values of the treatment factor. If $n = tr$ is a multiple of t, then each level of the factor will be applied to r unique experimental units, and there will be r replicates of each run with the same level of the treatment factor. If n is not a multiple of t, then there will be an unequal number of replicates of each factor level. All other known independent variables are held constant so that they will not bias the effects. This design should be used when there is only one factor under study and the experimental units are homogeneous.

For example, in an experiment to determine the effect of time to rise on the height of bread dough, one homogeneous batch of bread dough would be divided into n loaf pans with an equal amount of dough in each. The pans of dough would then be divided randomly into t groups. Each group would be allowed to rise for a unique time, and the height of the risen dough would be measured and recorded for each loaf. The treatment factor would be the rise time, the experimental unit would be an individual loaf of bread, and the response would be the measured height. Although other factors, such as temperature, are known to affect the height of the risen bread dough, they would be held constant and each loaf would be allowed to rise under the same conditions except for the differing rise times.

2.2 Replication and Randomization

Replication and randomization were popularized by Fisher. These are the first techniques that fall in the category of *error control* that was explained in Section 1.4.

The technique of replication dictates that r bread loaves are tested at each of the t rise times rather than a single loaf at each rise time. By having replicate experimental units in each level of the treatment factor, the variance of the experimental error can be calculated from the data, and this variance will be compared to the treatment effects. If the variability among the treatment means is not larger than the experimental error variance, the treatment differences are probably due to differences of the experimental units assigned to

each treatment. Without replication it is impossible to tell if treatment differences are real or just a random manifestation of the particular experimental units used in the study. Sub-samples or duplicate measurements, described in Chapter 1, cannot substitute for replicates.

The random division of experimental units into groups is called randomization, and it is the procedure by which the validity of the experiment is guaranteed against biases caused by other lurking variables. In the bread rise experiment randomization would prevent lurking variables, such as variability in the yeast from loaf to loaf and trends in the measurement technique over time, from biasing the effect of the rise time.

When experimental units are randomized to treatment factor levels, an exact test of the hypothesis that the treatment effect is zero can be accomplished using a randomization test, and a test of parameters in the general linear model, normally used in the analysis of experimental data, is a good approximation to the randomization test.

A simple way of constructing a randomized data collection form, dividing n experimental units into t treatment groups, can be accomplished using the SAS data step commands. For example, in the bread rise experiment, if the experimenter wants to examine three different rise times (35 minutes, 40 minutes and 45 minutes) and test four replicate loaves of bread at each rise time, the following SAS commands will create the list.

```
/* The unrandomized design */
data unrand;
  do Loaf=1 to 12;
    if (Loaf <= 4) then time=35;
    if (5 <=Loaf <= 8) then time=40;
    if (Loaf >=9 ) then time=45;
    dough_height='_____________';
    u=ranuni(0);
    output;
  end;
run;
/* Sort by random numbers to randomize */
proc sort data=unrand out=crd; by u;
/* Put Experimental units back in order */
data list; set crd;
  Loaf =_n_;
/* Print the randomized data collection form */
proc print double;  var time dough_height; id Loaf;
run;
```

These commands create the levels of the factor (rise time) in order, create an additional column of random numbers using the `ranuni()` function, create the variable `dough_height` that contains a space to record data in the output list, and then sorts the file by the random numbers to create a randomly ordered list. In the second data step the file `crd` is copied into a file called `list`, and the experimental units (loaves) are put back in sequential order. The `proc print` statement prints the SAS data file `list` using double spacing to allow room to write the response data after the experiments are run, and labels each

line with the variable `Loaf`. After running these commands the SAS output window contains a data collection form like the one shown below.

Loaf	time	dough_height
1	45	_____
2	45	_____
3	35	_____
4	45	_____
5	45	_____
6	35	_____
7	35	_____
8	40	_____
9	40	_____
10	35	_____
11	40	_____
12	40	_____

This list shows us that the first and second loaves, or experimental units, should be allowed to rise 45 minutes. The third loaf, or experimental unit, should be allowed to rise 35 minutes, etc. If you run the same commands, you may get a different random order due to the specific random numbers you obtain.

In addition to the data step, two SAS procedures (`proc plan`, and `proc factex`) can be conveniently used to create randomized data collection forms. These procedures will be used to create more complicated designs in forthcoming chapters, so just a simple introduction is presented here. For a completely randomized design SAS `proc plan` can be used to randomize a file containing the treatment indicators. The code listing below randomizes the list in the file `unrand` created in the first list above.

SAS `proc plan` automatically randomizes the order of the experimental

```
/* Randomize the design with proc plan*/
proc plan seed=27371;
  factors Loaf=12;
  output data=unrand out=crd;
run;
/* Put Experimental units back in order */
proc sort data=crd;
  by Loaf;
/* Print the randomized data collection form */
proc print double;  var time dough_height; id Loaf;
run;
```

units in the file `crd`, so no sorting by the random numbers is required. The second data step puts the experimental units back in sequential order so that the treatment levels appear randomized in the printed data collection form.

The other alternative is to use `proc factex`. The lines below call `proc factex` to create the file `crd`.

```
/* Creates CRD in random order using proc factex  */
proc factex;
  factors time/nlev=3;
  output out=crd time nvals=(35 40 45) designrep=4 randomize;
/* Add experimental unit id and space for response*/
data list; set crd;
  Loaf =_n_; dough_height='_____________';
/* Print the data Collection form                 */
proc print double;  var time dough_height; id Loaf;
run;
```

This procedure allows one to specify the level values for the factor and randomize the order of the list. The `designrep=4` specifies that four replicate of each factor level be included in the output file. Alternatively, a randomized list could be constructed using the `=rand()` function and the sort menu in an Excel spreadsheet. This would result in an electronic worksheet that could additionally be used to record the response data during execution of the experiments and later read into SAS for analysis.

2.3 A Historical Example

To illustrate the checklist for planning an experiment, consider a historical example taken from the 1937 Rothamstead Experimental Station Report, unknown (1937). This illustrates some of the early work done by Fisher in developing the ideas of experimental design and analysis of variance for use on agricultural experiments at the research station.

Objectives The objective of the study was to compare the times of planting, and methods of applying mixed artificial fertilizers (NPK) prior to planting, on the yield of sugar beets. Normally fertilizer is applied and seeds planted as early as the soil can be worked.

Experimental Units The experimental units were the plots of ground in combination with specific seeds to be planted in each plot of ground.

Response or Dependent Variable The dependent variable would be the yield of sugar beets measured in cwt per acre.

Independent Variables and Lurking Variables The independent variables of interest were the time and method of applying mixed artificial fertilizers. Four levels of the treatment factor were chosen as listed below:

1. (A) no artificial fertilizers applied
2. (B) artificials applied in January (plowed)
3. (C) artificials applied in January (broadcast)
4. (D) artificials applied in April (broadcast)

Lurking variables that could cause differences in the sugar beet yields between plots were differences in the fertility of the plots themselves, differences in the beet seeds used in each plot, differences among plots in the level of weed infestation, differences in cultivation practices of thinning the beets, and hand harvesting the beets.

Pilot Tests Sugar beets had been grown routinely at Rothamstead, and artificial fertilizers had been used by both plowing and broadcast for many crop plants; therefore, it was known that the independent variable could be controlled and that the response was measurable.

Choose Experimental Design The completely randomized design (CRD) was chosen so that differences in lurking variables between plots would be unlikely to correspond to changes in the factor levels listed above.

Determine the Number of Replicates A difference in yield of 6 cwt per acre was considered to be of practical importance, and based on historical estimates of variability in sugar beet yields at Rothamstead, four or five replicates were determined to be sufficient.

Randomize Experimental Units to Treatment Levels Eighteen plots were chosen for the experiment, and a randomized list was constructed assigning four or five plots to each factor level.

2.4 Linear Model for CRD

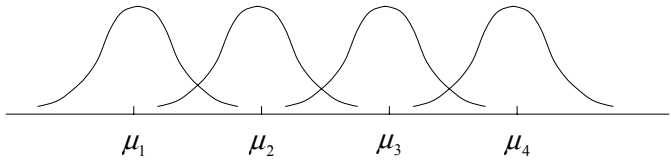
The mathematical model for the data from a CRD, or completely randomized design, with an unequal number of replicates for each factor level can be written as:

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{2.1}$$

where Y_{ij} is the response for the jth experimental unit subject to the ith level of the treatment factor, $i = 1, \ldots, t$, $j = 1, \ldots, r_i$ and r_i is the number of experimental units or replications in ith level of the treatment factor.

This is sometimes called the cell means model with a different mean, μ_i, for each level of the treatment factor. The distribution of the experimental errors, ϵ_{ij}, are mutually independent due to the randomization and assumed to be normally distributed. This model is graphically represented in Figure 2.1.

Figure 2.1 *Cell Means Model*



An alternate way of writing a model for the data is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}. \tag{2.2}$$

This is called the effects model and the τ_is are called the effects. τ_i represents the difference between the long-run average of all possible experiments at the ith level of the treatment factor and the overall average. With the normality assumption $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$ or $\epsilon_{ij} \sim N(0, \sigma^2)$. For equal number of replicates, the sample means of the data in the ith level of the treatment factor is represented by

$$\overline{y}_{i\cdot} = \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij} \tag{2.3}$$

and the grand mean is given by

$$\overline{y}_{\cdot\cdot} = \frac{1}{t} \sum_{i=1}^{t} \overline{y}_{i\cdot} = \frac{1}{n} \sum_{i=1}^{t} \sum_{j=1}^{r_i} y_{ij} \tag{2.4}$$

where $n = \sum r_i$. Using the method of least squares, the estimates of the cell means are found by choosing them to minimize the error sum of squares

$$ssE = \sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \mu_i)^2. \tag{2.5}$$

This is done by taking partial derivatives of ssE with respect to each cell mean, setting the results equal to zero, and solving each equation

$$\frac{\partial ssE}{\partial \mu_i} = -2 \sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \mu_i) = 0.$$

This results in the estimates:

$$\hat{\mu}_i = \overline{y}_{i\cdot}.$$

2.4.1 Matrix Representation

Consider a CRD with $t = 3$ factor levels and $r_i = 4$ replicates for $i = 1, \ldots, t$. We can write the effects model concisely using matrix notation as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.6}$$

.

Where

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{34} \end{pmatrix},$$

and $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

The least squares estimators for $\boldsymbol{\beta}$ are the solution to the normal equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$. The problem with the normal equations is that $\boldsymbol{X}'\boldsymbol{X}$ is singular and cannot be inverted. The solution to this problem using the SAS `proc glm` is to partition the $\boldsymbol{X}$ matrix as:

$$\boldsymbol{X} = \left(\begin{array}{ccc|c} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{array}\right) = [\boldsymbol{X}_1 | \boldsymbol{X}_2]$$

Then $\boldsymbol{X}'\boldsymbol{X}$ can be written in partitioned form as:
$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}$, and $\boldsymbol{X}_1'\boldsymbol{X}_1$ is non-singular. Therefore $\boldsymbol{X}'\boldsymbol{X}^- = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ is a generalized inverse for $\boldsymbol{X}'\boldsymbol{X}$, and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}$ is a solution that is equivalent to the estimates that would be obtained with the restriction that $\beta_t = 0$ (see Freund and Littell, 1981). For the example with $t = 3$ factor levels and $r_i = 4$ replicates for $i = 1, \ldots, t$,

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu} + \hat{\tau_3} \\ \hat{\tau_1} - \hat{\tau_3} \\ \hat{\tau_2} - \hat{\tau_3} \\ 0 \end{pmatrix}$$

2.4.2 L.S. Calculations with SAS proc glm

Table 2.1 shows the data from a CRD design for the bread rise experiment described earlier in this chapter.

Table 2.1 *Data from Bread Rise Experiment*

Rise Time	*Loaf Heights*
35 minutes	4.5, 5.0, 5.5, 6.75
40 minutes	6.5, 6.5, 10.5, 9.5
45 minutes	9.75, 8.75, 6.5, 8.25

Using these data we have

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 12 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 4 & 0 & 0 & 4 \end{pmatrix}, \quad \boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 88.0 \\ 21.75 \\ 33.0 \\ 33.25 \end{pmatrix},$$

and

$$\boldsymbol{X}'\boldsymbol{X}^{-} = \begin{pmatrix} 0.25 & -0.25 & -0.25 & 0 \\ -0.25 & 0.50 & 0.25 & 0 \\ -0.25 & 0.25 & 0.50 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 8.3125 \\ -2.8750 \\ -0.0625 \\ 0.0000 \end{pmatrix}.$$

The SAS commands to read in this data and compute these estimates are:

```
/* Reads the data from compact list   */
data bread;
  input time h1-h4;
  height=h1; output;
  height=h2; output;
  height=h3; output;
  height=h4; output;
  keep time height;
datalines;
35 4.5 5.0 5.5 6.75
40 6.5 6.5 10.5 9.5
45 9.75 8.75 6.5 8.25
run;
/* Fits model with proc glm           */
proc glm;
  class time;
  model height=time/solution;
run;
```

These commands read the data in a compact format, like Table 2.1, then use the output statement to create four lines in the file, `bread`, for each line in the input records. The result of the model `/solution;` option is:

```
Parameter              Estimate

Intercept           8.312500000 B
time       35      -2.875000000 B
time       40      -0.062500000 B
time       45       0.000000000 B
```

We can see the estimates are the same as those shown above. These parameter estimates are not unique because they depend on the ordering of the class variable `time`. `proc glm` recognizes this and designates the estimates as biased by following each value with a `B`.

2.4.3 Estimation of σ^2 and distribution of quadratic forms

The estimate of the variance of the experimental error, σ^2, is $ssE/(n-t)$. It is only possible to estimate this variance when there are replicate experiments at each level of the treatment factor. When measurements on sub-samples or duplicate measurements on the same experimental unit are treated as replicates, this estimate can be seriously biased.

In matrix form, ssE can be written as

$$ssE = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'y} = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-}\boldsymbol{X'})\boldsymbol{y},$$

and from the theory of linear models it can be shown that the ratio of ssE to the variance of the experimental error, σ^2, follows a chi-square distribution with $n-t$ degrees of freedom, i.e., $ssE/\sigma^2 \sim \chi^2_{n-t}$.

2.4.4 Estimable Functions

A linear combination of the cell means is called an estimable function if it can be expressed as the expected value of a linear combination of the responses, i.e.,

$$\sum_{i=1}^{t} b_i(\mu + \tau_i) = E\left[\sum_{i=1}^{t}\sum_{j=1}^{r_i} a_{ij}Y_{ij}\right] \tag{2.7}$$

From this definition it can be seen that effects, τ_i, are not estimable, but a cell mean, $\mu+\tau_i$, or a contrast of effects, $\sum c_i\tau_i$, where $\sum c_i = 0$, is estimable.

In matrix notation $\boldsymbol{L\beta}$ is a set of estimable functions if each row of $\boldsymbol{L}$ is a linear combination of the rows of $\boldsymbol{X}$, and $\boldsymbol{L}\hat{\boldsymbol{\beta}}$ is its unbiased estimator. $\boldsymbol{L}\hat{\boldsymbol{\beta}}$ follows the multivariate normal distribution with covariance matrix $\sigma^2\boldsymbol{L'}(\boldsymbol{X'X})^{-}\boldsymbol{L}$, and the estimator of the covariance matrix is $\hat{\sigma}^2\boldsymbol{L'}(\boldsymbol{X'X})^{-}\boldsymbol{L}$. For example using the data from the bread rise experiment above,

$$L = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \tag{2.8}$$

$\boldsymbol{L\beta} = \begin{pmatrix} \tau_1 - \tau_2 \\ \tau_1 - \tau_3 \end{pmatrix}$, and $\boldsymbol{L}\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\tau_1} - \hat{\tau_2} \\ \hat{\tau_1} - \hat{\tau_3} \end{pmatrix} = \begin{pmatrix} -2.8025 \\ -2.8750 \end{pmatrix}$ is a vector

of contrasts of the effects. The number of degrees of freedom, or number of linearly independent contrasts of effects in a CRD, is always the number of levels of the treatment factor minus one, i.e., $t-1$. Whenever there is a set of $t-1$ linearly independent contrasts of the effects, they are called a *saturated set of estimable contrasts.*

It can be shown that $(\boldsymbol{L}\hat{\boldsymbol{\beta}})'(\boldsymbol{L}(\boldsymbol{X'X})^{-}\boldsymbol{L'})^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}})$ follows the noncentral chi-square distribution, $\chi^2(p,\lambda)$ where the noncentrality parameter

$$\lambda = (\sigma^2)^{-1}(\boldsymbol{L}\boldsymbol{\beta})'(\boldsymbol{L}(\boldsymbol{X'X})^{-}\boldsymbol{L'})^{-1}(\boldsymbol{L}\boldsymbol{\beta}),$$

and $\boldsymbol{L}$ is the coefficient matrix for an estimable contrast like (2.8), and the degrees of freedom p is equal to the rank of $\boldsymbol{L}$.

Estimable contrasts can be obtained from SAS `proc glm` using the estimate command. For example to estimate the average difference in the cell means for the first and second levels of the treatment factor, $(\mu+\tau_1)-(\mu+\tau_2)=\tau_1-\tau_2$, use the command:

```
estimate '1 - 2' time 1 -1;
```

where the string in quotes is the label to be used in the output, `temp` is the name of the treatment factor, and the numbers `1 -1` are the contrast coefficients, c_i. The command:

```
estimate '1 -2 -3' time 1 -1 -1;
```

would produce no output and in the log file the message "`1 -2 -3 is not estimable`" would be printed, since $\sum c_i \neq 0$.

2.4.5 Hypothesis Test of No Treatment Effects

In the model for the CRD, the statistical hypothesis of interest is $H_0 : \mu_1 = \mu_2 = \ldots \mu_t$ or $\tau_1 = \tau_2 = \ldots = \tau_t$ versus the alternative H_a : at least two of the τs differ. If the null hypothesis is true, the model $y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$ simplifies to $y_{ij} = \mu + \epsilon_{ij}$ which can be represented as a single normal distribution with mean μ and variance σ^2 rather than multiple normal distributions like those shown in Figure 2.1.

The sums of squares about the mean is $ssTotal = \sum_{i=1}^{t}\sum_{j=1}^{r_i}(y_{ij}-\overline{y}_{..})^2 = \boldsymbol{y'y} - (\boldsymbol{1'y})^2/(\boldsymbol{1'1})$, where $\overline{y}_{..}$ is the grand mean and $\boldsymbol{1}$ is a column vector of 1s. This sum of squares can be partitioned as:

$$ssTotal = ssT + ssE \tag{2.9}$$

where $ssT = \hat{\boldsymbol{\beta}}'\boldsymbol{X'y} - (\boldsymbol{1'y})^2/(\boldsymbol{1'1}) = (\boldsymbol{L}\hat{\boldsymbol{\beta}})'(\boldsymbol{L}(\boldsymbol{X'X})^{-}\boldsymbol{L'})^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}})$, and $\boldsymbol{L}$ is the coefficient matrix for a saturated set of estimable contrasts. This quantity is called the treatment sums of squares. Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots \mu_t$, both ssT and ssE follow the chi-squared distribution. These sums of squares and their corresponding mean squares, which are formed by dividing each sum of squares by its degrees of freedom, are usually presented

in an Analysis of Variance or ANOVA table like that shown symbolically in Table 2.2.

Table 2.2 *Analysis of Variance Table*

Source	df	Sum of Squares	Mean Squares	F-ratio
Treatment	$t-1$	ssT	msT	$F = msT/msE$
Error	$n-t$	ssE	msE	
Total	$n-1$	$ssTotal$	$msTotal$	

Under the null hypothesis, the F-ratio msT/msE follows the F-distribution with $t-1$ and $n-t$ degrees of freedom, and under the alternative it follows the non-central F distribution with noncentrality parameter

$$\lambda = (\sigma^2)^{-1}(\boldsymbol{L\beta})'(\boldsymbol{L}(\boldsymbol{X'X})^{-}\boldsymbol{L'})^{-1}(\boldsymbol{L\beta}) = \frac{r}{\sigma^2}\sum_{i=1}^{t}(\mu_i - \bar{\mu}_\cdot)^2.$$

It is the generalized likelihood ratio test statistic for H_0, and is the formal method of comparing the treatment effects to the experimental error variance described in Section 2.2.

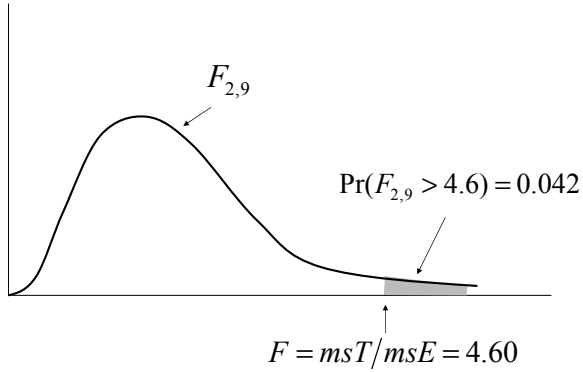
The sums of squares, mean squares, degrees of freedom in the ANOVA table, and associated F-test statistic are automatically calculated by SAS `proc glm`. For example, the result of the `proc glm; class temp; model height=temp;` command for the bread experiment shown earlier is:

```
                The GLM Procedure

Dependent Variable: height

                              Sum of
Source                 DF    Squares   Mean Square F Value Pr > F

Model                   2  21.57291667  10.78645833    4.60    0.0420
Error                   9  21.09375000   2.34375000

Corrected Total        11  42.66666667
```

In this table the ssT and msT and the associated degrees of freedom are on the line labeled `Model`, the ssE is on the line labeled `Error` and the $ssTotal$ is on the line labeled `Corrected Total`. The F-value is the ratio msT/msE and the last column labeled `Pr > F` is the probability of exceeding the calculated F-value if the null hypothesis is true. This is called the P-value and is illustrated graphically in Figure 2.2. If the experimenter chooses the significance level, α, for his hypothesis test, he would reject the hypothesis if the `Pr > F` value on the `proc glm` output is less than the chosen value of α.

For the bread rise experiment there are significant differences among the mean risen dough heights for each rise time at the significance level $\alpha = 0.05$, since $0.042 < 0.05$.

Figure 2.2 $Pr > F$



2.4.6 A Word of Caution

When a completely randomized design in one factor is conducted, the model for analysis is Equation (2.1) or (2.2) and the correct analysis is through the analysis of variance as shown symbolically in Table 2.2. The use of computer software like SAS makes it easy to analyze data and draw conclusions; however, if the experiment was not properly conducted even a sophisticated analysis of the data could be useless. The ϵ_{ij} term in the model (2.1) or (2.2), and its associated sums of squares, ssE, represents replicate experimental units. In many cases experimenters do not have replicate experimental units in each level of the treatment factor and substitute sub-samples or duplicates for them in the analysis. In other cases the experimental units are not properly randomized to treatment factor levels. When this is the situation, performing the analysis as if the design had been properly conducted may be completely wrong and misleading. Wrong conclusions can be drawn that do not hold up to later scrutiny, and a bad reputation is unfairly ascribed to statistically designed experiments and statistical analyses of data.

For example, consider an experiment where a professor would like to determine the effect of teaching methods on student test scores. If he uses one teaching method for the morning class, another for his evening class, and treats test scores for individual students as replicates, the results of his analysis may be totally wrong. This situation is similar to the cluster randomized studies described in exercise 5 of Chapter 1. The experimental unit is the class, since he applied the teaching method to a whole class simultaneously, and the individual students are sub-samples or observational units (since he must test individual students, not the class as a whole). The treatment effect should be judged against the variability in experimental units or classes. The variability among students in a class may be much different than variability from class

average to class average. Sub-sample observations should be averaged before analysis, as explained in Section 1.3. If this were done, he would only have one observation per class per teaching method and no replicates for use in calculating ssE in Table 2.2. There is no denominator for calculating the F-test statistic for teaching method. If he uses the variability in students within a class to calculate ssE, it may be too large or too small, causing him to reach the wrong conclusion about significance of the treatment effect. Further, if he did not randomize which teaching method was used in the morning and evening classes, and if he has no replicate classes that were taught with the same teaching method, his analysis is wide open to biases. Students in the morning classes may be fundamentally different than students in the evening classes, and any difference in average scores between the two teaching methods may be entirely due to differences among the two groups of students. In fact, if the professor knows there are differences in morning and evening students, he may purposely use the teaching method he wants to promote on the better class, thus ruining the objectivity of his research.

2.5 Verifying Assumptions of the Linear Model

The most critical assumption justifying the analysis based on the linear model presented in the last section is independence of the experimental error terms ϵ_{ij}. This assumption is justified if proper randomization of the experimental units to treatment factor levels has been performed. No further steps are needed to justify this assumption.

The other two assumptions for the linear model are constancy of the variance of the experimental error, σ^2, across all levels of the treatment factor, and normality of the experimental errors. Both of these assumptions can be visualized in Figure 2.1.

To verify the equal variance assumption, a simple scatter plot of the response data versus the factor levels can be constructed. Figure 2.3 shows an example of this type plot for the bread rise experiment. In this plot, similarity in the range of variability in dough heights between the different levels of rise time would justify the assumption of equal variance.

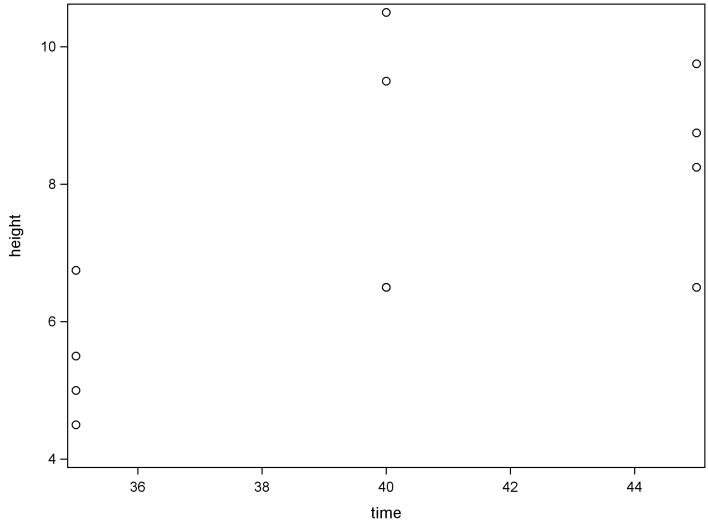
This figure was created in SAS using the commands:

```
proc sgplot data=bread;
  scatter y=Height x=Time;
run;
```

A similar plot that is used to check the equality of variance assumption is a plot of the residuals versus the cell means, or predicted values. The residuals are the differences of the response data and their respective cell means. The mean residual, for each level of the treatment factor, is zero. Therefore, in a plot of residuals it is easier to visualize whether the scatter or variability is equal. A common departure from equal variance occurs when the variance of the experimental errors increases as the mean response increases. By plotting

the residuals versus the cell means, this departure can be detected if the vertical scatter in the residuals increases from left to right on the plot.

Figure 2.3 *Plot of Response versus Factor Level for Bread Rise Experiment*



For the bread rise experiment, the SAS commands:

```
/* Fit model and Examine Residual Plots */
ods graphics on;
proc glm data=bread plots=diagnostics(unpack);
  class time;
  model height=time/solution;
run;
ods graphics off;
```

will fit the model for the CRD design, and when the ods graphics are turned on as shown in this example, the option `plots=diagnostics(unpack)` creates a panel of diagnostic plots to determine whether the assumptions of the model are satisfied. The `(unpack)` feature separates the plots so that they can be individually displayed. One of the plots created in this panel is labeled Residuals by Predicted and is shown in Figure 2.4. For the CRD design with one factor, the predicted values from the model, $\hat{y}_{ij} = \hat{\mu}_i$, are the cell means. As can be seen in Figure 2.4, there is a wider spread in the residuals for larger cell means or predicted values. This indicates that the equality of variance assumption may be violated. This fact is easier to see in Figure 2.4 than it was in Figure 2.3.

To verify the assumption of normality of the experimental error, a normal probability plot of the residuals can be examined. In the panel of diagnostic plots created by `proc glm` the normal probability plot is labeled Q-Q Plot of Residuals, and it is shown in Figure 2.5.

Figure 2.4 *Plot Residuals versus Cell Means for Bread Rise Experiment*

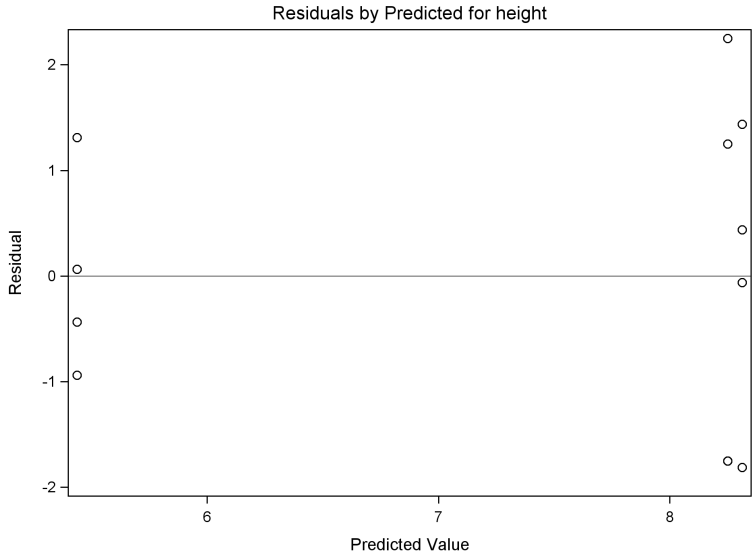
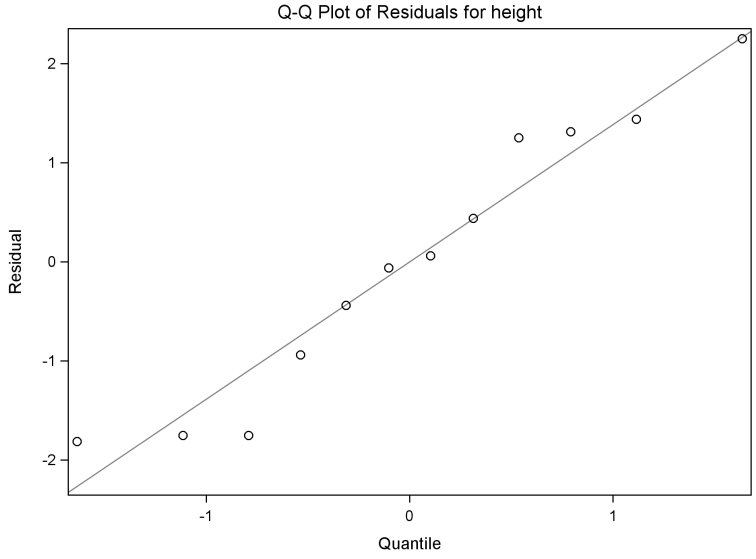


Figure 2.5 *Normal Plot of Residuals for Bread Rise Experiment*



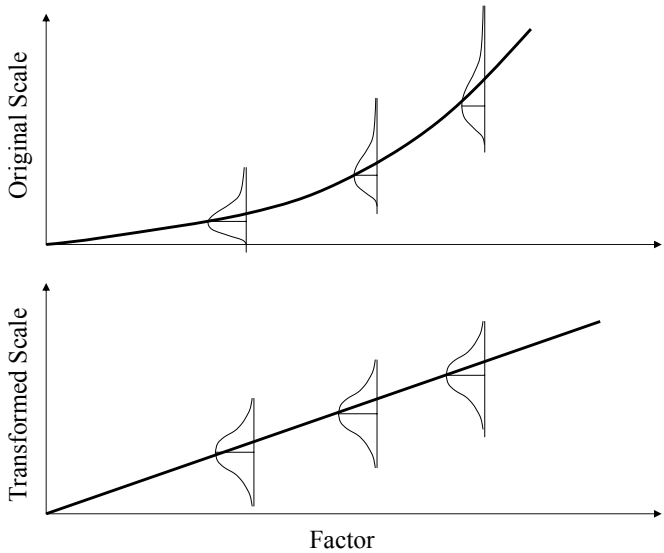
In a normal probability plot, if the points fall along a straight line, the normal distribution would be justified. The more data, and more points on the plot, the closer the points must lie to a straight line to justify the normality

assumption. In Figure 2.5, the points lie reasonably close to a straight line, and the normality assumption would appear justified. The equal variance assumption is more critical than the normality assumption, but they sometimes go hand in hand. When the equal variance assumption is violated, the normality assumption is often violated as well, and the corrective measures used for modifying the analysis when there is heterogeneity of variance will often correct both problems.

2.6 Analysis Strategies When Assumptions Are Violated

One common cause of heterogeneity of variances between levels of the treatment factor is a non-linear relationship between the response and stimulus or treatment. For example, in the upper half of Figure 2.6, it can be seen that the response increases non-linearly in response to the treatment. The density functions, drawn on their sides at three treatment levels, represent how non-linearity often affects the distribution of the response. As the mean or center of the distribution increases, the variance or spread in the distribution also increases, and the distributions have long tails on the right. One way of correcting this situation is to transform the response data prior to analysis.

Figure 2.6 *Representation of Effect of Non-linearities on Distribution of Response*



The bottom half of Figure 2.6 shows the potential result of a variance stabilizing transformation. On the transformed scale, the variance appears constant at different factor levels and the distribution appears more normal.

2.6.1 Box-Cox Power Transformations

One way to recognize the need for a variance stabilizing transformation is to examine the plot of residuals versus cell means described in the last section. If the spread in the residuals tends to increase proportionally as a function of the cell means, as illustrated in Figure 2.4, a transformation, $Y = f(y)$ can usually be found that will result in a more sensitive analysis. Box and Cox (1964) proposed a series of power transformations $Y = y^{\lambda}$ that normally work well. If the variance tends to increase as the mean increases, choose a value of λ less than one, and if the variance tends to decrease as the mean increases, choose a value of λ greater than one. Table 2.3 summarizes some common Box-Cox Power transformations. A common situation where the $\sigma \propto \mu$ is when the response is actually a measure of variability, like the sample variance s^2.

Table 2.3 *Box-Cox Power Transformations*

Relation Between σ and μ	λ	Transformation
$\sigma \propto \mu^2$	-1	Reciprocal
$\sigma \propto \mu^{3/2}$	$-\frac{1}{2}$	Square root of Reciprocal
$\sigma \propto \mu$	0	Log
$\sigma \propto \mu^{1/2}$	$\frac{1}{2}$	Square Root

In a CRD design with replicate experiments in each level of the treatment factor, one way to determine the most appropriate value of λ to use in the Box-Cox transformation is to fit a least squares regression line

$$\log(s_i) = a + b\log(\bar{y}_{i\cdot})$$

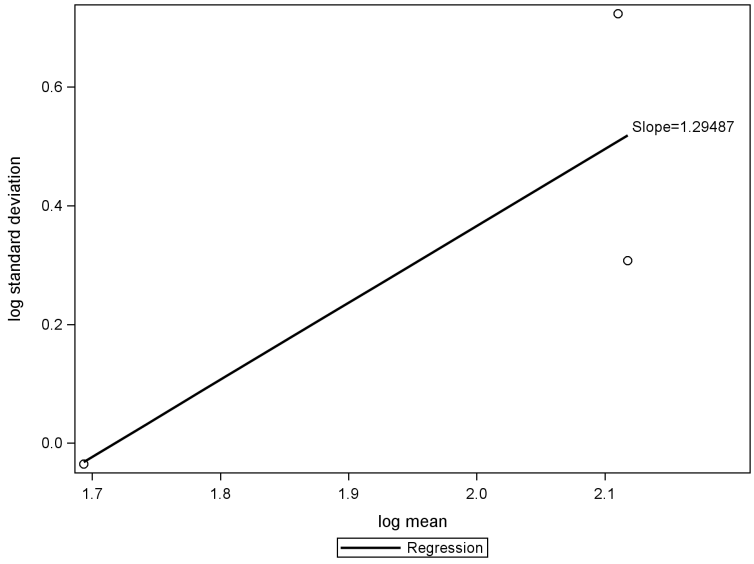
to the summary statistics. If the straight line provides a reasonable fit to the data, use $\lambda = 1 - b$. For example, for the bread rise experiment the SAS commands below use `proc sort` and `proc means` to calculate and store the cell means and standard deviations in the SAS file, `means`.

```
proc sort data=bread; by time;
proc means data=bread; var height; by time;
  output out=means mean=meanht stddev=s;
data logs; set means;
  logmean=log(meanht);
  logs=log(s);
```

Logarithms of both the cell means and standard deviations are computed in the data step and stored in the file `logs`. In the commands shown at the top of the next page `proc reg` is used to calculate the slope and intercept of the regression line and `proc sgplot` is used to plot the log standard deviations by log means with the fitted regression line as shown in Figure 2.7.

```
proc reg data=logs;
model logs=logmean;
run;
proc sgplot data=logs;
  reg y=logs x=logmean/curvelabel="Slope=1.29487";
  xaxis label="log mean";
  yaxis label="log standard deviation";
run;
```

Figure 2.7 *Plot of log(s_i) versus log($\bar{y}_{i\cdot}$) for Bread Rise Experiment*

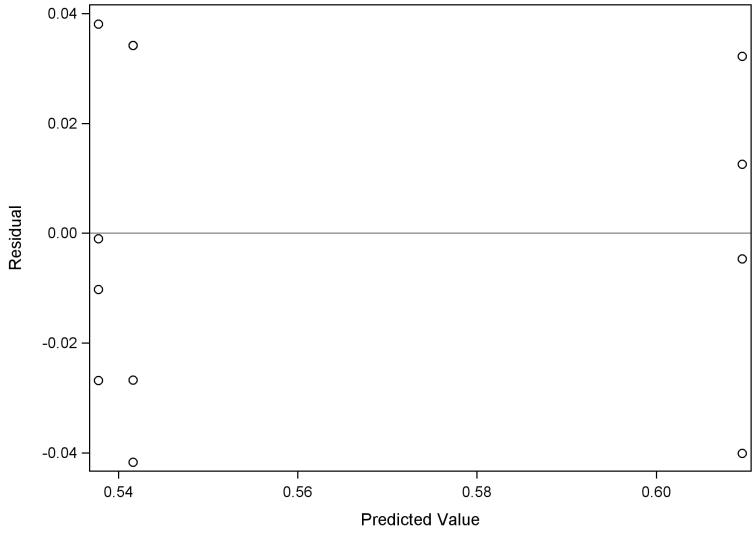


In this plot it can be seen that $\log(s_i)$ increases as $\log(\bar{y}_{i\cdot})$ increases, and that the linear fit provides a good representation of that relationship. Therefore, the appropriate λ to use for transforming the dough heights is $\lambda = 1 - 1.294869$. The SAS data step commands shown earlier to read in the data from the bread rise experiment can be modified to incorporate the transformation as follows:

```
data bread2;
  input time h1-h4;
  height=h1**(1-1.294869); output;
  height=h2**(1-1.294869); output;
  height=h3**(1-1.294869); output;
  height=h4**(1-1.294869); output;
  keep time height;
datalines;
```

After incorporating this transformation, the plot of residuals versus cell means is shown in Figure 2.8. It can be seen in this figure that the spread or variability of the residuals is nearly the same for each value of the predicted value or log cell mean.

Figure 2.8 *Plot of Residuals versus Cell means after* y^{λ} *Transformation for Bread Rise Experiment*



This transformation also makes the F-test more sensitive. The ANOVA table for the transformed response is shown below, where it can be seen that the P-value has decreased from 0.042 (before transformation) to 0.0227 (after transformation).

```
                The GLM Procedure

Dependent Variable: height

                              Sum of
Source                  DF      Squares    Mean Square  F Value  Pr > F

Model                    2   0.01305604   0.00652802     5.94    0.0227
Error                    9   0.00989833   0.00109981

Corrected Total         11   0.02295437
```

Even when transformations are used to justify the assumptions of the analysis and make the F-test more sensitive, the results are normally reported on the original scale. For example, it would be confusing to talk about the (dough height)$^{-.294869}$ in a report.

2.6.2 Distribution-Based Transformations

The distribution assumption for the effects model for the CRD described in Section 2.3, was $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$. However, if it is known that the data follow some distribution other than the normal distribution, such as the Binomial, Poisson or Lognormal, then it would also be known that the standard deviation would not be constant. For example, if the response, Y, was a binomial count of the number of successes in n trials, then due to the central limit theorem, Y would be approximately normal, but $\mu_Y = np$ and $\sigma_Y = \sqrt{np(1-p)}$, where p is the probability of success. In situations like this where the distribution of the response is known to follow some specific form, then an appropriate transformation can be found to stabilize the variance. Table 2.4 shows the transformation for common situations often encountered.

Table 2.4 *Response Distribution-Based Transformations*

Response Distribution	Variance in Terms of Mean μ	Transformation $f(y)$
Binomial	$\frac{\mu(1-\mu)}{n}$	$\sin^{-1}\sqrt{y/n}$ (radians)
Poisson	μ	$\sqrt{y}$ or $\sqrt{y+\frac{1}{2}}$
Lognormal	$c\mu^2$	$\log(y)$

2.6.3 Alternatives to Least Squares Analysis

When the variance of the experimental error is not constant for all levels of the treatment factor, but it is not related to the cell means, a transformation will not be an appropriate way of equalizing or stabilizing the variances. A more general solution to the problem is to use weighted least squares. Using weighted least squares, $\hat{\boldsymbol{\beta}}$ is the solution to the normal equations $\boldsymbol{X'WX\beta} = \boldsymbol{X'Wy}$, where $\boldsymbol{W}$ is a diagonal matrix whose diagonal elements are the reciprocals of the standard deviation within each treatment level. As an illustration of this method, consider the SAS commands, at the top of the next page, for analyzing the data from the bread rise experiment.

As in the previous examples, the data is read in the same way. The sort and means procedures calculate the cell standard deviations and the next data step merges these with the original data and computes the weights, `w`, that are the reciprocals of the standard deviations. The `weight` statement, in the `proc glm` call, solves the weighted least squares normal equations. The results appear below the commands.

```
/* Calculates mean and standard deviations  */
proc sort data=bread; by time;
proc means data=bread; var height; by time;
  output out=means stddev=s;
/* Creates weights                          */
data bread3; merge bread means; by time;
  w=1/s;
proc glm data=bread3;
  class time;
  model height=time;
  weight w;
run;
```

```
                The GLM Procedure

Dependent Variable: height

Weight: w

                              Sum of
Source                DF       Squares    Mean Square F Value Pr > F

Model                  2   18.20937757   9.10468879    6.23    0.0201
Error                  9   13.16060821   1.46228980

Corrected Total       11   31.36998578
```

With these results, it can be seen that F-test from the weighted least squares is more sensitive than the unweighted least squares, and the P-value is similar to what was obtained with the Box-Cox transformation shown in Section 2.5.1.

When the error distribution is not normal, an alternative to analyzing a transformation of the response is to use a generalized linear model (see McCullagh and Nelder, 1989). In fitting a generalized linear model, the user must specify the error distribution and a link function in addition to the model. The method of maximum likelihood is used to estimate the model parameters and the generalized likelihood ratio tests are used to test the hypotheses. When the link function is the identity and the distribution is normal, the generalized linear model analysis will result in the method of least squares and the ANOVA F-test. The SAS procedure `genmod` will fit the generalized linear model and compute appropriate likelihood ratio test statistics. This procedure allows the user to choose from the distributions Binomial, Poisson, Negative Binomial, Multinomial, Gamma, Inverse Gaussian, as well as Normal. For each distribution, `genmod` has a default link function.

To illustrate the use of `proc genmod` to analyze experimental data, consider the following example. A professor wanted to compare three different teaching methods to determine how the students would perceive the course. The treatment factor was the teaching method, the experimental unit was a class of students, and the response was the summary of student ratings for the

course. The professor taught two sections of the course for three consecutive semesters resulting in a total of six experimental units or classes. He constructed a randomized list so that two classes were assigned to each teaching method. This would reduce the chance that other differences in the classes, or differences in his execution of the teaching methods, would bias the results. At the end of each semester, the students were asked to rate the course on a five-point scale, with 1 being the worst and 5 being the best. Therefore, the response from each class was not a single, normally distributed response, y, but a vector $(y_1, \ldots, y_5)$ response that followed the multinomial distribution. The summary data from the experiment is shown in Table 2.5.

Table 2.5 *Counts of Student Rating Scores*

Class	*Method*	1	2	3	4	5
1	1	2	14	12	8	6
2	3	1	11	15	15	10
3	2	3	8	18	14	10
4	3	1	9	17	15	12
5	2	4	12	19	9	7
6	1	3	16	13	10	4

The following SAS commands read in the data and call `proc genmod` to make the generalized linear model analysis.

```
data teaching;
  input count score $ class method;
datalines;
2 1 1 1
14 2 1 1
   ...
proc genmod;
  freq count;
  class method;
  model score=method/dist=multinomial aggregate=method type1;
run;
```

The `$` following `score` in the input statement causes it to be read as a character variable. The `dist=multinomial` option on the model statement specifies the distribution. `proc genmod` then assumes the default cumulative logit link function. The `aggregate=method` option requests that a likelihood ratio test of no method effect be printed. The results appear as shown on the next page.

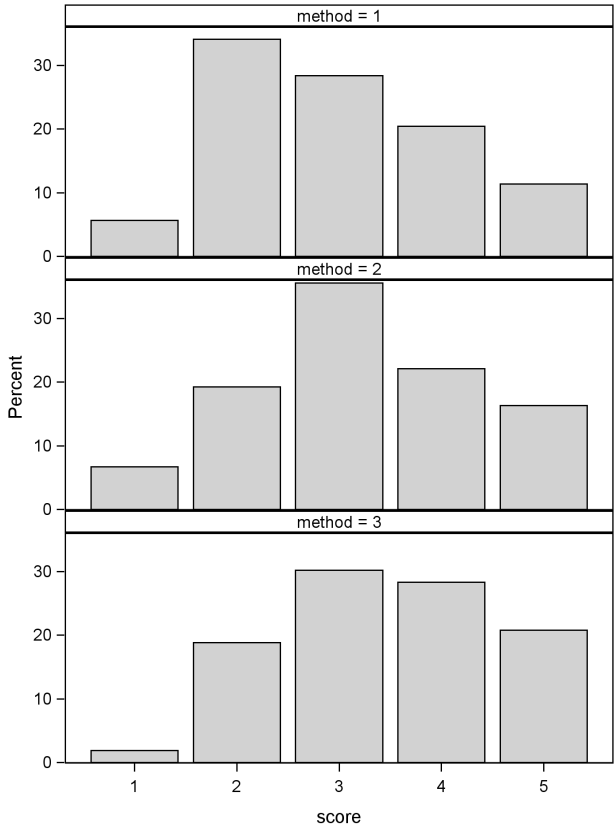
The P-value for the likelihood ratio chi-square statistic is small indicating there is a significant difference between the teaching methods. Teaching method 1 had an average score of 2.98, teaching method 2 had an average score of 3.22, and teaching method 3 appeared to be the best with an average

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercepts	13.7965			
method	4.1486	2	9.65	0.0080

score of 3.47. This can also be visualized in the bar charts in Figure 2.9 which shows that the percentage of high scores given increases for teaching method 2 and 3.

Figure 2.9 *Percentage of Student Rating Scores by Teaching Method*

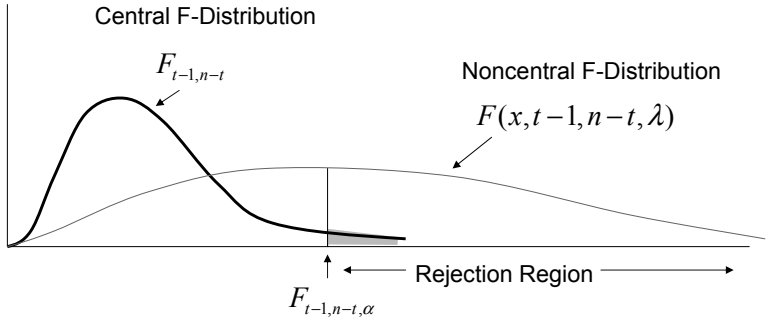


2.7 Determining the Number of Replicates

The significance level, α, of the ANOVA F-test of no treatment effect is the probability of rejecting the null hypothesis $H_0 : \mu_1 = \mu_2, \ldots, = \mu_t$, when it is

true. The power of the test is the probability of rejecting the null hypothesis when it is false. The test statistic msT/msE follows the F-distribution when the null hypothesis is true, but when the null hypothesis is false it follows the noncentral F-distribution. The noncentral F-distribution has a wider spread than the central F-distribution, as shown in Figure 2.10.

Figure 2.10 *Central and Noncentral F-Distribution*



The spread in the noncentral F-distribution and probability exceeding the critical limit from the central F-distribution is an increasing function of the non-centrality parameter, λ. When the distribution is the non-central F, the probability of exceeding the critical limit from the central F-distribution is called the power. The power is greater than the significance level, α, when the null hypothesis is false making the non-centrality parameter greater than zero. The power can be computed for any scenario of differing means, if the values of the cell means, the variance of the experimental error, and the number of replicates per factor level are specified. For a constant difference among cell means, represented by $\sum_{i=1}^{t}(\mu_i - \bar{\mu}_\cdot)^2$, the non-centrality parameter and the power increase as the number of replicates increase.

When the differences among cell means is large enough to have practical importance, the experimenter would like to have high power, or probability of rejecting the hypothesis of no treatment effects. When the difference among the means has practical importance to the researcher we call it *practical significance*. Practical significance does not always correspond to statistical significance as determined by the F-test from the ANOVA. Sometimes the number of replicates in the experiment is too few and the probability or power of detecting a difference of practical significance too low. Statistical significance can be made to coincide with practical significance by determining the appropriate number of replicates that result in the desired power. Doing this is the second technique that falls in the category of *error control* discussed in Chapter 1. The idea that increasing the number of replicates increases the sensitivity of the experiment is also due to Fisher (1935).

For example, if there is a difference among the cell means so that the cor-

rected sum of squares ($css = \sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2$) is greater than zero, then the power or probability of rejecting $H_0 : \mu_1 = \mu_2, \ldots, = \mu_t$ is given by

$$\pi(\lambda) = \int_{F_{t-1,t(r-1),\alpha}}^{\infty} F(x, t-1, t(r-1), \lambda)dx \tag{2.10}$$

where $F_{t-1,t(r-1),\alpha}$ is the αth percentile of the central F distribution with $t-1$ and $t(r-1)$ degrees of freedom, $F(x, t-1, t(r-1), \lambda)$ is the noncentral F-distribution with non-centrality parameter $\lambda = \frac{r}{\sigma^2}\sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2$. For a fixed value of $\frac{1}{\sigma^2}\sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2$, the non-centrality parameter, and the power increase as a function of the number of replicates, r. This probability can be calculated for various values of r until a value is found with adequate power. In this way the appropriate number of replicates can be determined. The `fprob` function in SAS facilitates these computations.

In the bread rise experiment, suppose less than a three-inch difference in risen dough heights is of no consequence. However, if changing the rise time from 35 minutes to 45 minutes causes a difference of more than three inches in risen dough height, the experimenter would like to know about it, because he will need to monitor rise time closely in the future to produce loaves of consistent height. In this case we can regard $\Delta = 3.0$ as a practical difference in cell means. The smallest $css = \sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2$ could be, with at least two cell means differing by Δ, would the case when one cell mean was $\Delta/2$ higher than the grand mean, a second was $\Delta/2$ less than the grand mean, and a third was equal to the grand mean. This would result in

$$css = \sum_{i=1}^{t}(\mu_i - \bar{\mu}_.)^2 = \left(\frac{\Delta}{2}\right)^2 + 0^2 + \left(-\frac{\Delta}{2}\right)^2 = \left(\frac{\Delta^2}{2}\right) = \left(\frac{3^2}{2}\right) = 4.5$$

Assuming the variance of the experimental error $\hat{\sigma}^2 = 2.1$ was estimated from the sample variance in risen dough heights in a pilot experiment where several loaves were allowed to rise for the same length of time, then the noncentrality factor can be calculated as $\lambda = \frac{r}{2.1} \times (4.5)$. The power is calculated for $r = 2, \ldots, 6$ using the SAS data step commands shown at the top of the next page.

```
*Example power computation in SAS data step using Bread Example;
data Power;
do r=2 to 6;
   nu1=3-1; * df for numerator;
   nu2=3*(r-1); * df for denomonator;
   alpha=.05;
   Fcrit=finv(1-alpha,nu1,nu2); *F critical value;
   sigma2=2.1;
   css=4.5;
   nc=r*(css)/sigma2;*noncentrality parameter for noncentral F;
   power=1-probf(Fcrit,nu1,nu2,nc);
   output;
end;
keep r nu1 nu2 nc power;
title Power Calculation in Data Step;
proc print; run;
```

Below are the results.

```
          Power Calculation in Data Step

  Obs     r     nu1     nu2        nc        power
   1      2      2        3      4.2857     0.19480
   2      3      2        6      6.4286     0.40419
   3      4      2        9      8.5714     0.59034
   4      5      2       12     10.7143     0.73289
   5      6      2       15     12.8571     0.83299
```

From this we can see that with $r = 5$ replicates there would be a 73% chance of detecting a difference in cell means as large as 3.0, and with $r = 6$ there is a 83% chance. With fewer than five replicates there is at least a 40% chance this difference will be missed. As a rule of thumb, the number of replicates that result in power between 0.80 and 0.90 is usually sufficient for most experimental studies.

A similar power calculation can be accomplished with the SAS Analyst Sample Size tool for the One Way ANOVA, with `proc power`, with the `proc glmpower`, or with the SAS Power and Sample Size Application. The interactive input for the SAS Analyst Sample Size tool is similar to the data step commands above, and the $css = \sum_{i=1}^{t}(\mu_i - \bar{\mu}_\cdot)^2$, must be specified. SAS Analyst will no longer be available after SAS version 9.2.

The SAS `proc power`, `proc glmpower` procedure, and the SAS Power and Sample Size Application require the user to enter the values of the cell means, μ_i, that he believes exhibit a practical significance rather than the $css = \sum_{i=1}^{t}(\mu_i - \bar{\mu}_\cdot)^2$, and σ rather than σ^2. The following commands will produce an equivalent result to the data step commands above for the bread rise experiment using `proc power`.

```
* Eample Power Calculation Using proc power;
proc power;
   OneWayANOVA
      Alpha = 0.05
      GroupMeans = (-1.5 0 1.5)
      StdDev = 1.449
      Power = .
      NPerGroup = 2 to 6 by 1;
run;
```

`proc glmpower` allows for more complicated models and will be illustrated in Chapter 3. The SAS Power and Sample Size Application is accessed by Start ► Programs ► SAS ► SAS Power and Sample Size ► SAS Power and Sample Size 3.1(or the latest release). It provides a GUI interface for `proc power` and `proc glmpower`.

2.8 Comparison of Treatments after the F-Test

When the F-test for the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots \mu_t$ is rejected, it tells us that there are significant differences between at least two of the cell means, but if there are several levels of the treatment factor, it does not necessarily mean that all cell means are significantly different from each other. When the null hypothesis is rejected, further investigation should be conducted to find out exactly which cell means differ. In some cases the investigator will have preplanned comparisons he would like to make; in other situations he may have no idea what differences to look for.

2.8.1 Preplanned Comparisons

Considering the treatment factor levels in the sugar beet yield experiment conducted at Rothamstead in 1937 and described in Section 2.3, some preplanned comparisons that might have been of interest are:

1. $H_0 : \mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$
2. $H_0 : \mu_2 = \mu_3$
3. $H_0 : \mu_3 = \mu_4$

The first comparison asks the question: Does a mix of artificial fertilizers change yield? The second comparison asks the question: Is there a difference in yields between plowed and broadcast application of artificial fertilizer? The third comparison asks the question: Does timing of the application change the yield?

These hypotheses can all be expressed in the general form $H_0 : \sum_{i=1}^{t} c_i\mu_i = 0$, where $\sum_{i=1}^{t} c_i = 0$. Since $\sum_{i=1}^{t} c_i\mu_i = 0$ are estimable functions, each of these hypotheses can be tested by computing the single estimable function $\boldsymbol{L}\hat{\boldsymbol{\beta}}$ and its standard error $s_{\boldsymbol{L}\hat{\boldsymbol{\beta}}} = \sqrt{\hat{\sigma}^2\boldsymbol{L}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{L}}$. The ratio of the estimable function to its standard error follows the t-distribution. The `estimate` command in SAS `proc glm` performs this test. For the sugar beet experiment the commands on the next page produce the output that follows the commands.

```
proc glm;
  class treat;
  model yield=treat;
  estimate 'fertilizer effect' treat 1 -.33333 -.33333 -.33333;
  estimate 'plowed vs broadcast' treat 0 1 -1 0;
  estimate 'January vs April' treat 0 0 1 -1;
  means treat/dunnettl('A');
run;
```

```
                                 Standard
Parameter              Estimate     Error   t Value Pr > |t|

fertilizer effect    -8.80001200  0.82520231  -10.66    <.0001
plowed vs broadcast  -3.80000000  0.97518951   -3.90    0.0016
January vs April      0.10000000  0.91941749    0.11    0.9149
```

The P-values in the column labeled `Pr >|t|`, in the above output, can be interpreted the same way the P-values for the F-statistic were interpreted, and we can see that: (1) artificial fertilizers enhance yield, (2) broadcast application results in higher yields than plowed application, and (3) there is no significant difference in yield between April and January application time.

When factor levels are quantitative, such as the rise time in the bread dough rise experiment, preplanned comparisons often involve looking for the significance of linear or higher order polynomial trends in the response. Contrast coefficients, c_i for testing orthogonal polynomial trends, can be obtained from SAS `proc iml orpol` function. The required inputs for this function are a vector containing the levels of a quantitative factor and the degree of polynomial desired. The result is an orthogonal matrix with the contrast coefficients desired beginning in the second column. For example, for the bread dough rise experiment the commands

```
proc iml;
  t={35 40 45};
  C=orpol(t);
  print C;
quit;
```

produce the result

```
                C

0.5773503 -0.707107 0.4082483
0.5773503         0 -0.816497
0.5773503 0.7071068 0.4082483
```

The estimate commands in the following call to `proc glm` take the second and third columns as the coefficients for the linear and quadratic polynomials.

```
proc glm data=bread;
  class time;
  model height=time;
  estimate 'Linear Trend' time -.707107 0 .707107;
  estimate 'Quadratic Trend' time .4082483 -.816497 .4082483;
run;
```

The result of the contrast statement is shown below, where we can see that there is a significant (at the $\alpha = 0.05$ level) linear trend, but no significant quadratic trend.

```
                                 Standard
Parameter           Estimate        Error  t Value  Pr > |t|

Linear Trend       2.03293262  0.76546578     2.66    0.0262
Quadratic Trend   -1.12268280  0.76546594    -1.47    0.1765
```

2.8.2 Unplanned Comparisons

When a set of preplanned comparisons can be expressed as a saturated set of orthogonal contrasts, like the examples shown in the last section, these comparisons are independent and equivalent to partitioning the overall F-test of $H_0 : \mu_1 = \ldots = \mu_t$. However, if the comparisons are not planned in advance of running the experiment, the analyst might be tempted to choose the comparisons he or she would like to make based on the means of the data. This implicitly means that all possible comparisons have been made. When testing all possible comparisons, each at the α=0.05 significance level, the overall significance level can be much higher than 0.05, greater than 50% in some cases. This means that even when there is no difference in the cell means $\mu_1, \ldots, \mu_t$ there could be a high probability of finding one or more comparisons significant when each is tested individually. In order to reduce the overall (or experimentwise) chance of a type I error, an adjustment must be made.

For pairwise comparisons of the form $H_0 : \mu_i = \mu_j$ for $i \neq j$ Tukey's HSD (or honestly significant difference) method adjusts the critical region by using the studentized range statistic instead of the student's t-distribution. Using the HSD reject $H_0 : \mu_i = \mu_j$ in favor of the alternative $H_a : \mu_i \neq \mu_j$ if $|\hat{\mu}_i - \hat{\mu}_j| > q_{I,n-t,\alpha/2} s_{\hat{\mu}_i - \hat{\mu}_j}$ where $q_{I,n-t,\alpha/2}$ is the $\alpha/2$ percentile of the studentized range. If $X_1, \ldots, X_I$ are independent random variables following $N(\mu, \sigma^2)$ and $R = \max_i X_i - \min_i X_i$ then $R/\hat{\sigma}$ follows the studentized range distribution (see Tukey, 1949a).

There are two different commands in SAS `proc glm` that will do pairwise comparisons using Tukey's HSD method. Both produce differently formatted output. The first is the means statement with the `tukey` option as shown below.

```
proc glm data=Sugarbeet;
  class treat;
  model yield=treat;
  means treat/tukey;
run;
```

The output illustrates the results from the sugar beet experiment and is shown on the next page.

```
Comparisons significant at the 0.05 level are indicated by ***.


                   Difference
       treat          Between      Simultaneous 95%
     Comparison         Means     Confidence Limits

       C - D           0.1000      -2.5723    2.7723
       C - B           3.8000       0.9655    6.6345  ***
       C - A          10.1000       7.2655   12.9345  ***
       D - C          -0.1000      -2.7723    2.5723
       D - B           3.7000       0.8655    6.5345  ***
       D - A          10.0000       7.1655   12.8345  ***
       B - C          -3.8000      -6.6345   -0.9655  ***
       B - D          -3.7000      -6.5345   -0.8655  ***
       B - A           6.3000       3.3122    9.2878  ***
       A - C         -10.1000     -12.9345   -7.2655  ***
       A - D         -10.0000     -12.8345   -7.1655  ***
       A - B          -6.3000      -9.2878   -3.3122  ***
```

The first column of the output lists the comparison made, the next column lists the difference in cell means, and the next column is a 95% confidence interval on the difference of means of the form $|\mu_i - \mu_j| \pm q_{I,n-t,0.025} s_{\hat{\mu}_{i\cdot} - \hat{\mu}_{j\cdot}}$. The final column is an indicator whether the null hypothesis has been rejected. For example, the confidence interval for the first comparison, $\mu_C - \mu_D$ includes zero and therefore no indicator (`***`) for significance is present, and the sugar beet yield for treatment (C - artificial applied broadcast in January) is not significantly different than the yield for treatment (D - artificial applied broadcast in April). All other pairwise comparisons show a significant difference.

Another way of obtaining the same result is through using the `lsmeans` statement with the `adjust=tukey` option. More on the difference in the `means` statement and the `lsmeans` statement will be given in Chapter 3. The command to do this is:

```
proc glm data=Sugarbeet;
  class treat;
  model yield=treat;
  lsmeans treat/pdiff adjust=tukey;
run;
```

and the output is shown on the next page.

In this output the actual cell means are listed at the top, followed by a table of P-values for the pairwise comparisons using Tukey's HSD method. The P-value in the third column fourth row 0.9995>.0.05 indicating the mean for treatment C is not significantly different than the mean for treatment D. All other P-values are less than 0.05 providing the same information as the `means` statement.

A less conservative method of comparing all possible cell means was developed independently by Newman (1939) and Keuls (1952). This method is also

```
                    The GLM Procedure
                  Least Squares Means
      Adjustment for Multiple Comparisons: Tukey-Kramer

                                          LSMEAN
              treat     yield LSMEAN      Number

              A           38.7000000           1
              B           45.0000000           2
              C           48.8000000           3
              D           48.7000000           4


             Least Squares Means for effect treat
             Pr > |t| for H0: LSMean(i)=LSMean(j)

                  Dependent Variable: yield

 i/j              1             2             3             4

    1                       0.0001        <.0001        <.0001
    2         0.0001                      0.0078        0.0094
    3         <.0001        0.0078                      0.9995
    4         <.0001        0.0094        0.9995
```

based on the studentized range statistic, but is based on the range of the particular pair of means being compared, within the entire set of ordered means, rather than the range of the largest - smallest as Tukey's HSD. The means comparison using the student Newman-Keuls method can be made using the `snk` option of the `means` statement as follows:

```
proc glm data=Sugarbeet;
  class treat;
  model yield=treat;
  means treat/snk;
run;
```

The output is shown below.

```
 Means with the same letter are not significantly different.


    SNK Grouping          Mean      N    treat

                A       48.8000      5    C
                A
                A       48.7000      5    D

                B       45.0000      4    B

                C       38.7000      4    A
```